2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application

Advanced Local Binary Pattern Descriptors for Crowd Estimation

Wenhua Ma[1,2],Lei Huang[1],Changping Liu[1]
1 *Institute of Automation*
Chinese Academy of Sciences, Beijing 100190, China
2 *Graduate University of the Chinese Academy of Sciences*
Beijing 100190, China
mwh-999@hotmail.com, lei.huang@mail.ia.ac.cn, changping.liu@mail.ia.ac.cn

Abstract

Local Binary Pattern (LBP) is a powerful texture descriptor that is gray-scale and rotation invariant. In this paper, an extension of the original LBP is proposed. LBP operator is adopted in multi-layer block domain, instead of pixel domain. Meanwhile, feature dimension is effectively reduced by Dual-Histogram LBP (DH-LBP). Combining merits of the two, we propose the advanced LBP (ALBP) and use that in solving the practical problem of crowd estimation. Experiment results demonstrate the performance and the potential of our method.

1. Introduction

Local Binary Pattern (LBP), originally designed for efficient texture classification, provides a simple and effective way for object/pattern recognition[1]. The original 8-bit LBP encodes 256 simple feature detectors (i.e. edge, curve/line, corner, spot) at different orientations in a single operator, and the LBP histogram contains the occurrence statistics of each feature over a region. LBP based methods have been used successfully in face recognition[2], background modeling[3], texture classification[4] as well as image segmentation [5].

The most important properties of LBP are its tolerance against illumination changes and its computational simplicity. In order to make the LBP more suitable to real-world scenes, we propose a modification to the operator and induce it into solving the practical problem of crowd estimation. Crowd estimation is crucial for crowd monitoring and control. It differs from pedestrian detection or people counting in that no individual pedestrian can be properly segmented in the image, which is often the case in practical. For instance, when a crowd of people move at once no particular pedestrian can be properly segmented. In other cases, a single pedestrian may be too small to be reliably and robustly detected in an image. People counting under these cases is technically challenging and computationally expensive. On the contrary, a macroscopic approach that represents the overall nature of space would suffice for crowd estimation. Texture pattern coarseness is proved to be related to crowd density. Marana et al has shown in [6] that texture descriptors like Grey Level Dependence Matrix, Fourier Spectrum Analysis and Minkowski Fractal dimension are efficient to estimate crowd density in real time with simple background. However, these raw texture descriptors are limited in distinctive power and sensitive to noise. Ref.[7] induces histogram-based feature in people counting, in which edge orientation and blob size histograms are adopted as feature vectors.

We have proposed the advanced Local Binary Pattern (ALBP) feature as texture descriptor in our automatic surveillance system, which can generate a density map as well as a total density estimation of a target area. In addition to introducing the ALBP as a generative texture descriptor, we report on two main contributions to crowd estimation research: 1) Improve the patch-based method used in [8] by fixing the size of image cells according to their "capacity". 2) Propose methods of calculating total density of an image based on image cells in it. The rest of the paper is organized as follow: section 2 introduces the construction of ALBP features, section 3 proposes details of the system for crowd estimation. Experiment results are reported in section 4. Conclusions and future research directions can be found in section 5.

2. LBP-based features

The original LBP operator introduced by Ojala et al[4] is a powerful texture descriptor. LBP is by definition invariant against any monotonic transformation of the gray scale. The discrete occurrence histogram of the LBP computed over an image or a region of image has shown to be a very powerful texture feature.

978-0-7695-3490-9/08 $25.00 © 2008 IEEE
DOI 10.1109/PACIIA.2008.258

2.1. LBP in Block Mean Domain

The LBP operator in this paper is different from the original LBP. The original LBP operates directly on the image pixels and uses the central pixel value as threshold makes it sensitive to noise[3, 4]. Using block coefficients and the mean threshold value is better.

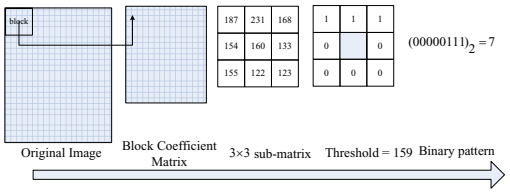

Figure 1. LBP code in block mean domain

As shown in Figure 1, firstly, an original image is divided into a set of non-overlapping uniform-sized blocks. Coefficients of all these blocks are used for further process. For simplicity, only the mean value of the block is used in current work, which is calculated as (1):

$$BM = \frac{\sum I(x, y)}{s^2} \tag{1}$$

BM is short for block mean, and $I(x, y)$ is intensity of pixel (x, y) in the block, s is the size of the block in pixel. s^2 is block capacity for square blocks are used.

These block means are grouped into a matrix, from which 3×3 sub matrices of neighboring elements are formed. The eight elements surrounding the center are thresholded by the average value of the nine elements, forming a binary pattern vector with eight bits, denoted as BFV.

2.2. Multi-scale analysis

Multi-scale information is included by combining BFVs extracted from blocks of different sizes together, shown as (2). α_i is a weighting factor controlling the relative impact of different BFVs and BFV_i denotes BFV of the ith-layer blocks. Multi-scale analysis enable us to deal with local and global information at the same time, which is expected to improve the performance of our system.

$$MBFV = [\alpha_1 \times BFV_1, \alpha_2 \times BFV_2, \ldots \alpha_n \times BFV_n] \tag{2}$$

2.3. Feature reduction

Uniform LBP (ULBP) is an extension to the original LBP, which reduces feature dimensions and increases noise immunity. In our work, a new indication of ULBP called the Dual-Histogram LBP (DH-LBP) is proposed to further reduce feature dimensions. Let black dot represents 0, white dot represents 1, ULBP can be visually interpreted as a circle connected by a black curve and a white curve, the pattern of which can be fixed uniquely by two parameters: white curve length L and white curve start point C. Uniform LBP codes are encoded by the connected histograms of L and C. Meanwhile, because the probability for bright spot (all 0s) and dark spot (all 1s) is quite small, they fall into the same class called Specials with the non-uniform LBP. The DH-LBP feature vector deduced from 8-bit OLBP has 16 parameters. As shown in Figure 2.

Content	Specials	Histogram of white curve length L, ranges from 1 to 7							Histogram of white curve start point C, ranges from 0 to 7							
Digit	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

Figure 2. Construction of DH-LBP feature

2.4. Similarity measures used

Feature vector histograms can be compared using various similarity measures, such as the L_1 norm distance, L_2 norm distance, standard Euclidean measure and so on. It is denoted in [9] that, L_1 norm distance gives overall best performance in feature vector histogram matching, which has low computational complexity and good accuracy also. Based on their work, we have done further research and found that standard L_1 norm distance is more suitable for our work, this can be demonstrated later by the experiment results in section 4. For two histograms H_i and H_j with bins numbered as $b = 1 \ldots d$, standard L_1 norm distance is defined as (3),where $\sigma(b)$ denoting the variance of all histograms at bth bin in terms of L_1 norm distance.

$$SL_1(i, j) = \sum_{b=1}^{d} \frac{|H_i(b) - H_j(b)|}{\sigma(b)} \tag{3}$$

3. Crowd estimation

The major goal of our work is to generate a stable and accurate density map for the input video frames. This is done by dividing the image into a set of cells and calculating local density within the cells, similar to that used in [8]. In our work, image cell is manually labeled with a density level ranging from 1 to 5 during the training phase, corresponding to the very low (VL), low (L), moderate (M), high (H) and very high (VH) crowd densities respectively. ALBP feature vectors are extracted from these image cells and relationship between feature and crowd densities is learned by supervised learning.

3.1. Crowdedness definition standard

Image-cell based analysis is supposed to make estimations more accurate and more robust to environmental

change. Hence, a uniform and explicit standard for crowdedness definition under various surroundings as well as different locations is necessary. This is achieved by ensuring that different image cells have the same "capacity". In other words, image cells are able to cover equal number of people at most. Assuming that pedestrians have similar size in nature, under a certain scene, we first pick a region of interest (ROI) and specify size of the nearest and farthest image cells (assign the size twice of the width of a reference pedestrian). After that, we can approximate the perspective model by linearly interpolating between the two extremes of the scene and generating other cells.

Table 1. Crowdedness definition of image cell

Set	VL	L	M	H	VH
NP	0-0.5	0.5-2	2-4	4-6	> 6
AOP	$\leqslant$ 10%	10%-30%	30%-60%	60%-90%	>90%

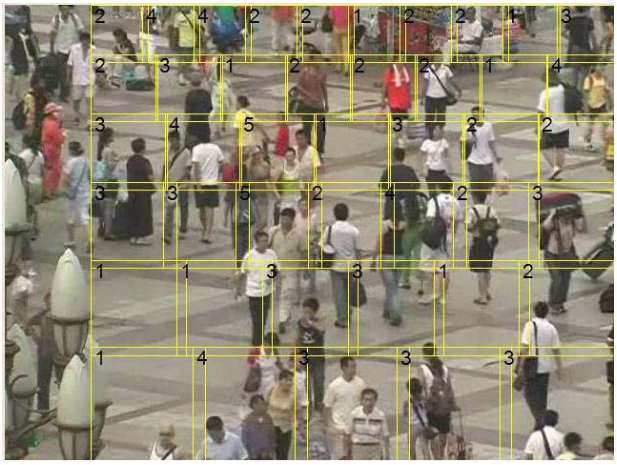

Figure 3. Manually labeled image cells

As shown in Table 1, two factors are taken into consideration to define the crowdedness level of an image cell, namely: number of pedestrians in it (NP), and the area occupied by pedestrians (AOP). They are somehow related to each other and both have direct correlations with people's perception of crowdedness. Example of a manually-labeled image is presented in Figure 3.

3.2. Classification method

As the classification rule, we employ nonparametric clustering under one labeled class and simulate prototype distributions based on K-means clustering. K clusters (K may be different for specific ranges) are generated for each density class. Distance between image cell and cluster is measured in terms of distance specified in formula (3). During testing phase, an image cell is classified using nearest neighbor algorithm, assigned to the same class with that of its nearest cluster.

4. Experiment result

4.1. Database

The crowd estimation system is implemented under C# and DirectShow. Four clips of real videos are used: two for training and the other two for testing. The training set is further divided into two for training and validation. We use the validation set to optimize feature parameters like block sizes and weighting factors. Table 2 shows basic information of training set and testing set. We allow the image cells to have moderate overlapping areas in order to cover as large image area as possible.

4.2. Performance of image cell classification

In order to have uniform representation of the texture measurements, image cells are normalized to the same size before feature extraction. Classification Accuracy is defined as the proportion occupied by correctly classified samples:

$$CA = \frac{\#(correct)}{\#(total)} \times 100\% \tag{4}$$

where CA is short for Classification Accuracy, and $\#(\bullet)$ indicated numbers of the variable.

To start, feature parameters are changed iteratively to establish their values for best classification results on validation set. The final feature vector is constructed by four-layed blocks, $s = 2, 3, 5, 7$ respectively. Then, the whole training set is used for constructing the classifier.

As shown in Table 3, the ALBP feature obtains 83.28% classification accuracy, which is quite promising. For comparison, other texture features are also used in our experiment. Total edge performs the worst (about 40.67% merely), color entropy is a little better than that (about 48.01%). Performance improves significantly as more sophisticated features are used. Histogram-based features, such as Edge orientation histogram and LBP all outperformed simple texture features, because histogram-based representation contains statistical information, making it more expressive, also more powerful in handling noise. On the other hand, compared with Original LBP and ULBP, DH-LBP reduces feature dimension remarkably (6.25% compared with original LBP, 27% compared with uniform LBP) with just a little sacrifice in accuracy (about 1.91% compared with ULBP and 0.89% compared with Original

Table 2. Basic information of training set and testing set

Scene	Usage	ROI size	Max.cell size	Min.cell size	Cell number	Frame number
subway station	train	640×430	180×180	80×80	23	120
train station	train	550×480	120×120	60×60	43	60
square	test	600×430	160×160	120×120	13	100
subway station	test	500×180	120×120	60×60	14	50

Table 3. Classification Accuracy on image cells using different features

Feature	Total Edge	Color Entropy	GLDM	Edge orientation	Original LBP	ULBP	MULBP	DHLBP	ALBP
Dimension	1	1	16	8	256	59	236	16	64
CA (%)	40.67	48.01	70.26	73.02	80.52	81.54	83.32	79.63	83.28

LBP). What is noteworthy is that by applying ULBP and DH-LBP in four scales, CA increased by 1.78% and 3.65% respectively . Proving that multi-scale analysis is indeed beneficial. Summing up the above, ALBP gives a good trade-off between computation complexity and classification accuracy.

4.3. Performance of total density estimation

Alternative methods can be used in total density estimation, such as mean estimation and maximum estimation. Mean estimation defines the total density as mean of the density of all image cells in a ROI:

$$TCD = \frac{\sum_{i=1}^{5} i \times s(i)}{N_{total}} \tag{5}$$

where and TCD is total crowd density of the image. $s(i)$ is the number of image cells classified into density class i, N_{total} is the total number of image cells in ROI. While maximum estimation defines the total density as the density range with highest $s(i)$.

Because crowdedness is subjective and sensuous, estimations should be evaluated by human operators. In our work, 20 subjects were asked to view the images and score total crowd density in 1-5 scale, also corresponding to the very low, low, mediate, high and very high densities. The average scores were calculated for all the subjects and adopted as ground truth. Estimations of the above two methods are compared with the ground truth in Figure 4, in which crowd densities are demonstrated every 10 frames. It can be seen that mean estimation is more approximate to the ground truth for most of the images, therefore, mean estimation is adopted for total density calculation.

Average square error between estimation and ground truth is induced to evaluate total density estimation:

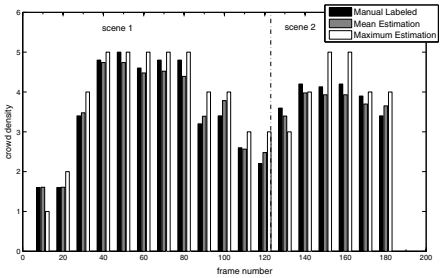

Figure 4. Estimation of Total crowd density

$$ASE = \frac{\sum_{i=1}^{N_{frame}} (TD_i - ETD_i)^2}{N_{frame}} \times 100\% \tag{6}$$

where ASE is short for Average Square Error, and TD_i denotes the total crowd density of the ith frame while ETD_i is its estimation. N_{frame} represents number of the images.

Table 4. ASE of different similarity measures

Measure	L_2	SED	L_1	SL_1
ASE on training set (%)	5.78	3.88	3.94	3.57
ASE on testing set(%)	10.72	6.73	6.08	5.37

Estimations based on different similarity measures are shown in Table 4. L_1 norm distance, L_2 norm distance, Standard Euclidean distance (SED) and Standard L_1 norm distance (SL_1) are tested. Among these, Standard L_1 norm distance gives best estimation. That's why it is adopted as feature similarity measure.

Since the testing set is highly different from training set in many aspects such as pedestrian size, surrounding construction, camera orientation and so on, performance on testing sets can be used to judge the robustness of the system. ASE on training and testing sets are shown in Table 5. Scene 1 and scene 2 are training sets, while scene 3 and scene 4 are testing sets. Compared with training sets, ASE on testing sets increased slightly (about 1.80%). Proving that our system gives very good and robust performance even if in totally unfamiliar environments. Figure 5a) shows the estimation in scene 3 and Figure 5b) shows the estimation in scene 4.

Table 5. ASE on training set and testing set

Scene	1	2	3	4	Training	Testing
ASE(%)	3.47	3.78	5.96	4.20	3.57	5.37

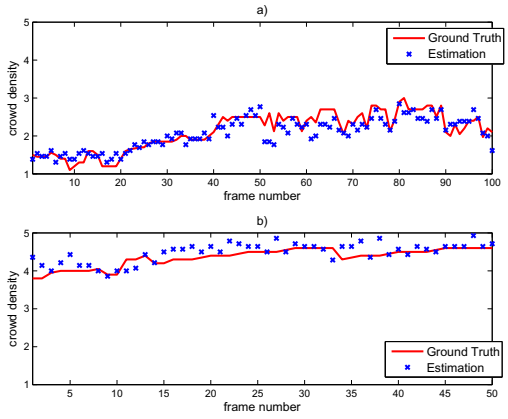

Figure 5. Total crowd density on testing set

5. Conclusion and future work

The ALBP approach, despite its simplicity, provides better distinctive performance compared with previous methods. Experiment results involving different spatial scales show that multi-scale analysis is beneficial.

Uniform and explicit crowdedness criterion based on image cells is also our innovative work in crowd estimation. In virtue of that, our system can perform effectively under different surroundings and give rather authentic crowd density estimations.

Regarding future work, one thing deserving a closer look is selection of block coefficients, many candidates, such as block variance and energy components of different frequencies can be used. Feature selection will also be incorporated in our future work.

References

[1] M. Pietikainen T. Ojala and D. Harwood. A comparative study introduction of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(7):51–59, January 1996.

[2] M. Pietikainen T. Ahonen, A. Hadid. Face description with local binary patterns: application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006.

[3] M. Pietikainen Marko Heikkila. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):657–662, 2006.

[4] T.Maenpaa T. Ojala, M. Pietikainen. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[5] In-So Kweon Dong-Su Kim, Wang-Heon Lee. Automatic edge detection using 3×3 ideal binary pixel patterns and fuzzy based edge thresholding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[6] A. Marana, L. da Costa, R. Lotufo, and S. Velastin. On the efficacy of texture analysis for crowd monitoring. *SIBGRAPHI '98: Proceedings of the International Symposium on Computer Graphics, Image Processing, and Vision*, pages 354–361, 1998.

[7] Hai Tao Dan Kong, Doug Gray. Counting pedestrians in crowds using viewpoint invariant training. *Proceedings British Machine Vision Conference*, 2005.

[8] Xinyu Wu, Guoyuan Liang, Ka Keung Lee, and Yangsheng XU. Crowd density estimation using texture analysis and learning. *IEEE Transactions on System, Man, and Cybernetics*, 31(6):645–654, 2001.

[9] Xin Chen and Charles H. Reynolds. Performance of similarity measures based on histograms of local image feature vectors. *Computer Science*, 42(6):1407–1414, 2002.