HPC Feature for Crowd Outdoor Scenes Estimation

H. H. Lin[#1], and K.T. Win[#2]

[#]*Universityof Computer Studies, Mandalay*
Myanmar

Abstract *— Crowd people estimation in outdoor scenes is a difficult task and an important research area in the computer vision of person monitoring security fields. The effectiveness of feature is very crucial for this estimation. Although the existing works system got the suitable result, but have not worked well due to the weakness of their feature detector. Only the histogram gradient feature, or principal component feature can't perfectly represent for all locations of the crowd people area. So, effective feature detection plays an important factor in the crowd estimation system. To cover these factors, this work proposes the hybrid HPC feature (processed as a cascaded feature) to establish the crowd people estimation system in order to predict the accurate people number. To get the HPC cascaded feature, histogram of gradient (H) feature, principal component analysis (P) feature and convolutional neural network (C) feature is used as the step by step procedure. Although three feature detectors are used as a cascaded, this work gets the less time complexity. Convolutional neural network applies as the classifier to estimate the system result. The experimental performance is evaluated on PET 2009 and UCSD crowd datasets and this work gets the significant and lowest error rate.*

Keywords *— crowd estimation, histogram feature, component feature, convolution feature, cascaded feature.*

I. INTRODUCTION

Outdoor crowd estimation is an active research area in an automatic surveillance field (such as train stations, pedestrianized streets, stadiums, crowd management, parks, and etc.). There have been many crowd estimation approaches for the security purposes such as prevention and management. In the crowd estimation system, detection plays a critical role, since it is the basic important step for crowd counting or estimation.

The major challenge of pedestrian detection and estimation systems is the development of reliable on-board pedestrian detection systems. Due to pose changing, object carrying, background clustering, illumination, noise and shadow, it is really difficult to cope with the demanded robustness of this kind of system. It is very hard to estimate the person number due to the illumination and occlusion.

In the previous works, many researchers have attempted the people estimation or counting problem. Although their works got the suitable results, they have still some weak points. There have been clustered into two groups: direct approach (counting the person individual) and indirect approach (counting the person based on estimation) according to the surveying of the literature. Certainly, some systems are introduced to count the person number by directly or indirectly. This is shown in Fig. 1.

While the direct approach implements to localize each person on the scene with classifiers, the indirect one attempts to apply the learning algorithms or statistical analysis of the whole crowd. In the direct approach, is known as the detection-based approach, everyone is individually detected by applying some segmentation or detection to obtain the people number result.

Sharma et al. [1] proposes a part based edgelets feature training (person silhouette contours, arms, face, shoulders, head, etc.). Every person is localized, traced along the whole frames and counted. Some previous direct approaches focus on blob detection [2], omega shape (head plus shoulders) detection [3] and face detection [4]–[7]. Due to increasing of crowd scene density, it can't possible to sum individually since some single person features fail to detect and give an accurate crowd result [8].

The benefit of the direct approach is the high accuracy when individual person is detected to focus the location of people. But they don't consider the different people densities or partial occlusions. Hence, the weakness of this approach is that incorrect foreground segmentation, and unreliable output, especially in crowded conditions.

In the indirect approach, is known as map-based, the feature measurement is performed without the separation of every person in the scene to show the relationship.

D. Ryan et al. [9] identifies the local feature set from the foreground groups in the input images. These features are based on area, perimeter, perimeter-area ratio, edges and histogram of edge angle. The density map is calculated to load every pixel for perspective compensating. The size of every group is estimated by using a linear least square model that applies the extracted features. Finally, they resulted out the total group sizes by summing all.

Previous indirect methods are corner points training [10] and dynamic texture model [11]. This is more reliable because they are focused on features, but it is difficult to discover an exact correspondence between these features and the person's number, mainly if people can be in diverse distances from the group and camera view with different densities. Albiol [12] developed the corner point algorithm (Harris). They can't handle the perception effects and can't influence the crowd density in corner point'sdetection. Since the Harris' corner detector can be unsteady for objects moving towards the camera or away from it.

In this paper, we propose a cascaded feature extraction idea for effective crowd outdoor scenes estimation system. It attempts to offer an accurate count estimation by handling the challenging factors. The foreground area is localized by the use of mean shape background estimation (ms) and k-mean clustering algorithm for accurate segmentation. These localized segmented regions of the whole image is fed as the input image for the feature extraction step. In feature extractionstep, Histogram of gradients (HOG), principal component analysis (PCA) and convolutional neural network (CNN) are used to develop the cascaded feature extraction (histogram feature, component feature and the convolution feature).

In this paper the effective terminology of outdoor scenes crowd estimation system has been addressed. This paper also shows the distinct progress in crowd estimation that has been achieved by proposed methods. This paper is organized as follows. Section II presents the proposed approach to contribute in crowd estimation field. Section III investigates the performance comparison with state-of-art results. Finally, the concerns and conclusion has been attempted in the last section.

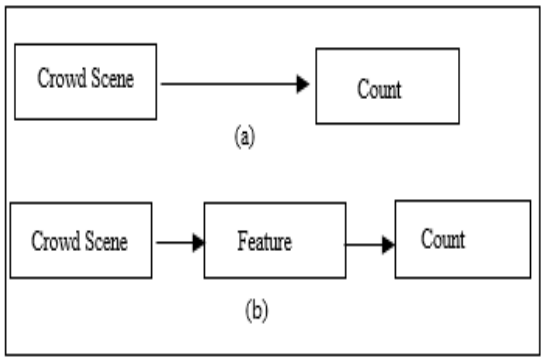

Fig. 1. (a) Crowd counting vs (b) Crowd estimation Method

II. PROPOSED APPROACH

This paper presents the implementation of outdoor scenes crowd estimation system that consists of pre-processing, foreground detection, cascaded feature extraction and classification step. This system is based on the analysis of the crow estimation task. Fig. 2 shows the basic structure of the proposed architecture.

A. Pre-processing and Foreground Detection

The original video sequence is set as the input. Image size and its quality are a serious problem. Hence, it needs to resize and enhance the input image. The mean shape background estimation [13], frame differencing method with a fixed threshold (60) is used to detect the foreground object. After this process, a k-means clustering is followed to cluster and group the foreground pixels.

B. Cascaded Feature Extraction

The aim of the feature extraction is to discover the better distinctive features. This work uses three feature descriptors (cascaded feature) as the feature extraction, feature reducing and feature learning. In the feature extraction step, Histogram of oriented gradients (HOG) is used. The high resulted hog feature set is reduced by the principal component analysis as the feature reducing step. Then, the reduced valuable features are put into the convolutional neural network as the feature training. This convolutional neural network method is not only as the feature extraction/selection algorithm to extract relevant features for people detection, but also used as the classifier for an accurate counting estimation.

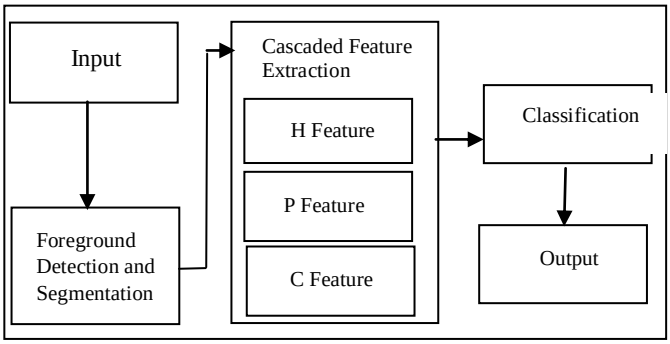

Fig. 2. Proposed System Architecture

a). HistogramFeature (H):

Feature extraction is the transformation of input data into a set of features. Histogram of oriented gradient (HOG) is used to extract the foreground object feature. It tends to count the gradient occurrences orientation in localized foreground portions. The H features are extracted to every frame of the segmented foreground image. This image is divided into 8×8 small connected regions. For each cell, a histogram of gradient directions or edge orientations is calculated. The proposed system aggregates the gradient to the corresponding cells, makes a histogram on each cell, and normalizes the histogram along four directions. 0 to 180 degrees is used to discretise each cell into 9 angular bins according to the gradient orientation. The pixel of each cell gives the weighted gradient to its corresponding angular bin. Block (spatial region) is

the groups of adjacent cells and it is the fundamental fact for histogram normalization and grouping. The block histogram is the normalized group of histograms. This is calculated by the equation 1 and 2.

$$mag = \sqrt[2]{(g\_x)_2 + (g_y)_2} \quad (1)$$

$$ang = \arctan {}^{(g_{-x})}\Big/_{(g_{-y})} \quad (2)$$

where mag is the value of magnitude, ang is the value of the angle, g_x and g_y are the horizontal and vertical gradient pixel values, and arctan is the tangent inverse of the gradient values. Unlike the other histogram-gradient approaches, this paper normalized the sum of four cell feature together, instead of one-fourth reducing of the dimensional feature vector (Fig. 3).

Finally, an H feature of 8100 × number of frame single feature vector is extracted from the detected foreground image since because there are 225 blocks into 36 values 1D feature vector. This feature evolves many feature dimension numbers (most feature based on the unusable background feature) and gets the worse results.

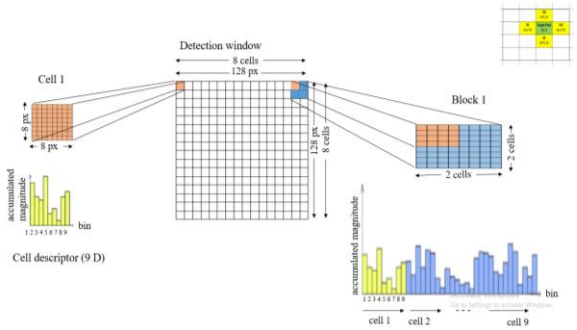

Fig. 3. Illustration of H Feature Calculation

b). Component Feature (P):

To handle the gradient feature over-fitting, principal component analysis (PCA) is used to reduce the feature dimension. It is a linear subspace projection technique, to reduce the feature vector of the proposed system. The main goal is to detect the correlation between features and to find the direction that maximize the separation between different classes.

The proposed system finds the mean vector of the feature variables to centreon the origin. It projects the data onto the line, measure the distance from this point to the origin and finds the best fitting line by maximizing the distances from the projected points to the origin. This process is calculated by the equation 3 and 4.

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} - Mean \quad (3)$$

$$Covariance\ Matrix = C * C' \quad (4)$$

Obtain the Eigenvectors and Eigenvalues from the covariance matrix, Ax = ʎx where A is the transformation matrix, x is the vector and ʎ is a scalar value. The size of H_P feature matrix is 1024 × number of frames.

c). Convolution Feature (C):

To get the effective feature, convolutional neural network (CNN) is used. The input of CNN is the H_P feature input matrix (input size) M × M with a kernel of size N × N. The reduced number 1024 is split into the 32 × 32 feature matrix. The general explaining of the CNN is showed in Fig. 4. If the convolution process will be 28 × 28 feature map. This is computed as the following equation 5.

$$G_i^j = f\left(\sum_{k \in I_i} (G_k^{j-1} \otimes F_{k_j}^J + B_i^j)\right) \quad (5)$$

$$wei = \left(\sum_{i=1}^{n} wei_{ij}\right)_{j=1}^{n} \quad (6)$$

where I is the input image set, G_i^j is the i feature map of the j layer, F is the filter, B is the bias, x is the pixel value, wei is the weight and n is the number of neurons. After that pooling is used to reduce the feature rectangular blocks and subdivides it to yield a single output from that block.

$$h_j^n(x,y) = max_{s \in N(\tilde{x}), y \in N(\tilde{y})} h_j^{n-1}(\tilde{x}, \tilde{y}) \quad (7)$$

where h_j^n is the j feature map of the n layer. Logistic sigmoid non-linearity rectified linear unit (ReLU) is applied as an activation function. It proceeds the real number values. The large negative values turn to 0 and the large positive values turn to 1.

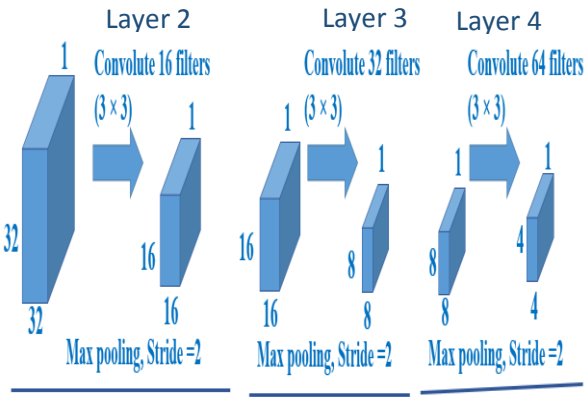

Fig. 4. Illustration of Convolution Feature Map

In this work, five layers (one input layer, three hidden layers in between, and one output fully connected layer) are used. This last layer is used as the classifier. The hidden layers are used to extract the feature. The output feature map of the first hidden layer (CNN layer 2) is 3 × 3× 32 feature map.

The second hidden layer result out the 3 × 3 × 2048 feature map. The last hidden layer (CNN Layer 4) emerges 3 × 3 × 8192 feature map as the result.

The last fully convolutional layer is used not only the feature detector but also the classifier. This is the advantages of the CNN (can automatically extract and classify the feature to study the network by using global information, the location of input space disappears and need not apply). The layer 5 result out the 2048 feature map before classifying the label and computing the loss function. This 2048 feature map is used as the proposed cascaded HPC feature. The loss function is calculated by the following equation 8.

$$L = -\sum_j y^j log_{10}\theta(o)^j \quad (8)$$

where L is loss value, y is true label, o is the last output layer of the network, j is the dimension of the vector and ɵ is the probability estimated value.

III. EXPERIMENTAL RESULTS

In order to calculate the proposed cascaded feature extraction (HPC) methods for outdoor crowd estimation system, experiments are tested on the available challenging dataset (PET 2009, and UCSD). The summary of the datasets with detail information used in the experiments is shown in Table I.

To evaluate the performance, this paper uses MAE (Mean Absolute Error) and MRE (Mean Relative Error). MAE can give clear insight that the predicted value is the same with the ground truth data. Otherwise, there exists the higer miss rate. MRE is like as the relative square error. This is the rate of the comparative the predictor value with the actual value. This is calculated by the equation 9 and 10.

$$MAE = \frac{1}{N}.\sum_{t=1}^{N}|h_a - h_b| \quad (9)$$

$$MRE = \frac{1}{N}.\sum_{t=1}^{N}\frac{|h_a - h_b|}{h_a} \quad (10)$$

where N is the total number of testing frames, h_a and h_b is the ground truth and the estimated count of person in frame t. The execution environment is on Core i7 with 3.1 GHz, memory 4GB, NVidia GeForce and matlab 2018a.

TABLE I. DETAIL FACTS OF CROWD DATASETS

Name	*Year*	*Frame No.*	*Resolution*	*Density*
PETs (S1.L1)	2009	500 (RGB)	768 × 576 (Outdoor)	Medium crowd
PETs (S1.L2)	2009	520 (RGB)	768 × 576 (Outdoor)	High density crowd
UCSD	2008	2000 (Gray)	238 × 158 (Outdoor)	Medium crowd

A Performance Result

To verify the performance results, this paper analyses the different feature extraction methods. Before this work contributes in cascaded nature, the performance result of the individual feature detectors is shown in Table II. In this tes, this made as the cascaded feature, according to this paper contribution point. So, four ways are tested in this experiment. These are shown in TABLE III, IV, V, and VI.

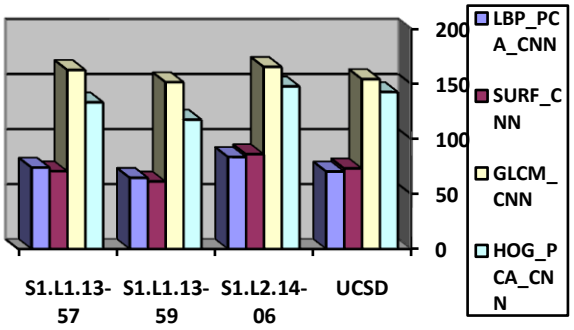

Fig. 5. Time comparison results of various feature methods of datasets

The above Fig. 5 shows the time complexity of various feature extractionmethods tested on four sequences of PET 2009 and UCSD dataset. The performances of the proposed system are compared with the previous method [14] are shown in TABLE. VI. Each frame consumes 1.5 frame per second.

TABLE II . PERFORMANCE COMPARISON OF VARIOUS FEATURE EXTRACTION METHODS ON S1.L1.13-57 SEQUENCE

Methods	S1.L1.13-57 Sequence	
	MAE	*MRE %*
LBP	20.86	29.98
PCA	22.98	32.14
SURF	17.35	20.73
GLCM	14.42	18.66
HOG	**0.56.37**	**3.44**

TABLE III. PERFORMANCE COMPARISON OF VARIOUS CASCADED FEATURE SETS ON S1.L1.13-59 SEQUENCE

Methods	S1.L1.13-59 Sequence	
	MAE	*MRE %*
LBP_PCA_CNN	8.86	10.98
SURF_CNN	3.35	3.13
GLCM_CNN	1.25	9.52
HOG_PCA_CNN	**0.45**	**3.06**

TABLE IV. PERFORMANCE COMPARISON OF VARIOUS CASCADED FEATURE SETS ON S1.L2.14-31 SEQUENCE

Methods	S1.L2.14-31 Sequence	
	MAE	*MRE %*
LBP_PCA_CNN	7.7	19.78
SURF_CNN	3.47	14.71
GLCM_CNN	0.85	11.29
HOG_PCA_CNN	**0.56**	**7.72**

TABLE V. PERFORMANCE COMPARISON OF VARIOUS CASCADED FEATURE SETS ON UCSD

Methods	UCSD	
	MAE	*MRE %*
LBP_PCA_CNN	2.67	23.09
SURF_CNN	1.98	20.22
GLCM_CNN	0.80	15.46
HOG_PCA_CNN	**0.35**	**5.42**

TABLE VI. COMPARISON OF PROPOSED SYSTEM WITH [10, 14]

Methods	PET 2009 (Average)		UCSD	
	MAE	*MRE %*	*MAE*	*MRE %*
Proposed system	**0.46**	**4.13**	**0.35**	**5.42**
Albiol [10]	5.56	16.01	2.82	15.38
I.Topkaya [14]	1.48	9.045	1.3	4.9

In these tables, Local Binary Pattern (LBP) – PCA – CNN cascaded feature got the highest error rate although their easy and less computational complexity. Because this only used the pixel difference (magnitude information is ignored), so the people area are misclassified. SURF_CNN cascaded feature got the worst result due to the unsuitable key point feature. The GLCM_CNN' result is suitable but not the lowest. Among them, our proposed cascaded feature set (Hog_Pca_Cnn) got the lowest error rate to prove the effectiveness of the proposed system.The proposed system tests how the system results change the number of input feature size. This is showed in TABLE. VII and Fig. 5.

TABLE VII. VARIOUS NUMBER OF HPC FEATURE DIMENSION

No. of H feature →	No. of P feature →	No. of C feature
8100	784	28 × 28
8100	1024	32 × 32
8100	4096	64 × 64
8100	16384	128 × 128

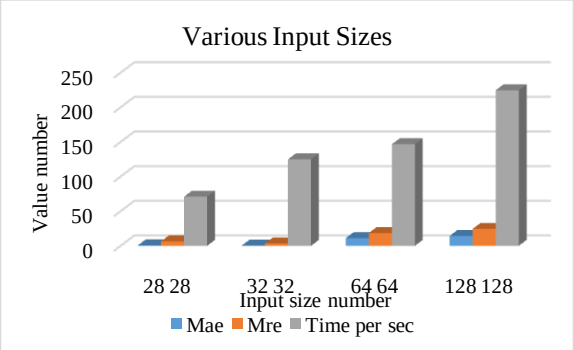

Fig. 6. Various Input Size Testing Result

IV. CONCLUSION

A hybrid HPC significant feature is developed for outdoor crowd people estimation. This system considers the way to avoid the challenging factor of the unreliable crowd estimation system by focusing on various viewpoints. The predicted crowd result is nearly same with the ground truth of the datasets. Hence, this system will suitable to use as the real time system in monitoring security fields. As a discussion, we tested only on the offline outdoor crowd benchmark dataset. In the near future, we will test on the real time outdoor dataset.

REFERENCES

[1] P. K. Sharma, C. Huang, and R. Nevatia, "Evaluation of people tracking, counting and density estimation in crowded environments," in IEEE International Workshop on Performance Evaluation of Tracking andSurveillance, Miami, USA, 2009, pp. 39–46.

[2] J.-W. Kim, K.-S. Choi, B.-D. Choi, and S.-J. Ko, "Real-time vision-based people counting system for the security door," in International Technical Conference On Circuits Systems, Computers and Communications,Phuket, Indonesia, 2002, pp. 1418–1421.

[3] J.-W. Kim, K.-S. Choi, B.-D. Choi, and S.-J. Ko, "Real-time vision-based people counting system for the security door," in International Technical Conference On Circuits Systems, Computers and Communications,Phuket, Indonesia, 2002, pp. 1418–1421.

[4] X. Zhao, E. Dellandra, and L. Chen, "A people counting system based on face detection and tracking in a video," in IEEE International Conference on Advanced Video and Signal Based Surveillance, Washington, USA,2009, pp. 67–72.

[5] A. L. Koerich, L. E. S. Oliveira, and A. Britto Jr., "Face recognition using selected 2DPCA coefficients," in 17th International Conference on Systems, Signals and Image Processing, Rio de Janeiro, Brazil, 2010,pp. 10–16.

[6] L. E. S. Oliveira, M. Mansano, A. L. Koerich, and A. S. Britto Jr., "2d principal component analysis for face and facial-expression recognition,"Computing in Science & Engineering, vol. 13, no. 3, pp. 9–13, 2011.

[7] T. H. H. Zavaschi, A. S. Britto Jr., L. E. S. Oliveira, and A. L. Koerich, "Fusion of feature sets and classifiers for facial expression recognition,"Expert Systems with Applications, vol. 40, no. 2, pp. 646–655, 2013.

[8] C. R. P. Morerio, L. Marcenaro, "People count estimation on small crowds," in 9th IEEE International Conference on Advanced Video andSignal-Based Surveillance, Beijing, China, 2012, pp. 476–480.

[9] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using group tracking and local features," in IEEE International Conference on Advanced Video and Signal Based Surveillance, Washington, USA,2010, pp. 218–224.

[10] A. Albiol, M. Silla, A. Albiol, and J. Mossi, "Video analysis using corner motion statistics," in IEEE International Workshop on PerformanceEvaluation of Tracking and Surveillance, Miami, USA, 2009, pp. 31–37.

[11] A. B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Miami, USA,2009, pp. 101–106.

[12] C. Harris and M. Stephens, "A combined corner and edge detector",in Proceedings of the 4th Alvey Vision Conference,pages 147–151, 1988.

[13] H. H. Lin and K.T. Win, "People Counting with Extended Convolutional NeuralNetwork",in Proceeding of 27th International Conference on Computer Theory and Applications, (ICCTA), pp. 99- 104, Alexandria, Egypt, 2017.

[14] Topkaya, H.Erdogan and Fatih, "Counting People by Clustering Person Detector Outputs", Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2014.