Crowd Counting using Multiple Local Features

David Ryan, Simon Denman, Clinton Fookes, Sridha Sridharan
Image and Video Laboratory
Queensland University of Technology
Brisbane, Australia
d23.ryan@student.qut.edu.au, s.denman@qut.edu.au, c.fookes@qut.edu.au, s.sridharan@qut.edu.au

***Abstract*—In public venues, crowd size is a key indicator of crowd safety and stability. Crowding levels can be detected using holistic image features, however this requires a large amount of training data to capture the wide variations in crowd distribution. If a crowd counting algorithm is to be deployed across a large number of cameras, such a large and burdensome training requirement is far from ideal. In this paper we propose an approach that uses local features to count the number of people in each foreground blob segment, so that the total crowd estimate is the sum of the group sizes. This results in an approach that is scalable to crowd volumes not seen in the training data, and can be trained on a very small data set. As a local approach is used, the proposed algorithm can easily be used to estimate crowd density throughout different regions of the scene and be used in a multi-camera environment. A unique localised approach to ground truth annotation reduces the required training data is also presented, as a localised approach to crowd counting has different training requirements to a holistic one. Testing on a large pedestrian database compares the proposed technique to existing holistic techniques and demonstrates improved accuracy, and superior performance when test conditions are unseen in the training set, or a minimal training set is used.**

***Keywords*-Crowd Counting, Crowd Density, Local Features, Foreground segmentation**

I. Introduction

In large public places, it is often impossible to monitor every person for suspicious behaviour. The threats posed in crowded environments are of a different nature to those posed by an individual, and arise from the crowd's collective properties: "a crowd is something other than the sum of its parts" [6]. These threats include fighting, rioting, violent protest, mass panic and excitement. The most common indicator of such behaviour is crowd size, which may also be an indicator of congestion, delay or other abnormality. As crowd size is a *holistic* description of the scene, the majority of crowd counting techniques have utilised holistic features to estimate crowd size. However, due to the wide variability in crowd behaviours, distribution, density and overall size, holistic systems require a very large training set. In a facility containing numerous cameras, it is not practical to supply hundreds of frames of ground truth for potentially hundreds of cameras.

In this paper we propose a novel approach that uses *local* features, defined here as features which are specific to an individual or small group within an image. While existing techniques have used similar local features such as foreground pixels, they are analysed at a holistic level. Local features are used here to estimate the number of people within each *group*, so that the total crowd estimate is the sum of all group sizes. As local features are used, training data must also be annotated with local information. To provide appropriate training data, a unique method of localised ground truth annotation is proposed which greatly reduces the required training data.

As well as the reduced training requirement, a localised approach also enables the estimation of crowd densities at different locations within the scene (unlike holistic systems which can only provide a density for the whole scene), and allows for a simplistic extension to a multi-camera environment. The ability to determine local crowd densities greatly improves the systems ability to detect abnormalities in a scene. While the overall number of people in a scene may be considered normal, there may be a very high concentration of people in a small area. Holistic systems are unable to detect such an abnormality, however the proposed local approach can easily detect such an occurrence.

The proposed system is tested on a 2000 frame database [4] featuring crowds of size 11-45 people. The proposed technique is compared to two holistic techniques, and is shown to outperform holistic techniques in terms of accuracy, scalability and practicality. The system is shown to be highly scalable, as it is capable of extrapolating to count crowds which are larger or smaller than those encountered during training; and highly practical, as it is able to count crowds when trained on as few as 10 frames of training data.

The remainder of the paper is structured as follows: Section II provides an overview of existing crowd counting techniques, Section III outlines the proposed algorithm, Section IV describes the proposed ground truth annotation method, Section V presents experimental results and Section VI presents conclusions and possible directions for future work.

II. Existing Work

The task of crowd counting has been approached from a number of angles, but the techniques share a common

framework: feature extraction using image processing, followed by crowd counting using classification. The output of the classifier is a measure of crowding, which is a holistic description of a scene. Therefore it is logical to use holistic features which are indicative of larger crowds. Local features, however, provide more detailed information about a scene. As computer power increases, these techniques have become more popular.

Holistic features, such as textural information [12], Minkowski Fractal Dimension [11], and Translation Invariant Orthonormal Chebyshev Moments [15] have been used to measure crowd density. Holistic features such as these are highly sensitive to external changes (such as lighting conditions), and it has been shown that for outdoor environments, the natural fluctuations in lighting between morning and afternoon reduce system performance [15].

More recent crowd counting algorithms have utilised specific features which are indicative of crowding, such as edge and foreground pixels. While these features are local to points of interest in an image, they are considered at a holistic level. Many techniques [6], [14], [9] have used foreground segmentation to determine the crowd count. The relationship between the total number of foreground pixels and the number of people in the scene has been shown to be approximately linear [6]. However, local nonlinearities arise due to the effects of perspective and occlusion.

Paragios [14] proposed the use a geometric factor to weight each pixel according to its location on the ground plane, to overcome the problem of perspective. Occlusions have been addressed using blob size histograms [9], or by using more features [4]. The blob size histogram captures the range of blob sizes present in an image (compared to a foreground pixel count), and enables the classifier to distinguish between groups of people and individuals. By contrast, Chan et al. [4] extract features in a greater quantity, however additional features greatly increase the quantity of training data required.

Local features are specific to an individual or small group of people within an image. For example, head detection has been proposed to estimate crowd sizes [10]. Tracking [13] and blob segmentation [16] have been employed, however these approaches are best suited to situations where crowds are small. Celik [3] assumed linearity between blob size and group size, and Kilambi [8] used an elliptical cylinder model and tracking to estimate group size. While these systems all employ local features, they often rest on specific assumptions, including image quality. When presented with low-quality video and poor segmentation, it is difficult to classify or track the local features unless ground truth is also annotated on a local level.

Local features have been employed to other crowd related problems though, such as crowd detection [2] (detection of human like objects and repeating structures) and analysis of crowd stability [1] (using optical flow over time). However neither of these algorithms is concerned with the overall size of the crowd.

III. Crowd Counting Using Multiple Local Features

A. System Description

A crowd counting system is proposed which uses local rather than holistic features. These features are 'local' with respect to the blob segments in a foreground mask, obtained using a foreground segmentation technique [7]. A crowd estimate is obtained for each blob in an image, so that the total estimate for the scene is the sum of the estimates for each individual blob. In order to train the system, ground truth annotation is performed *after* the first stage of image processing, once the foreground is extracted. The group size is manually counted for each blob in an image, therefore each frame provides several instances of ground truth.

This approach is built on the assumption that it is easier for a system to estimate the number of people in each group than to estimate the entire crowd at once. It is possible for a crowd of 20 people to be distributed as two large groups or as ten pairs (for example). Viewed from a holistic perspective, these various crowd distributions can give rise to vastly different image features. Existing techniques cope by extracting a larger quantity of holistic features (29 features are used in [4]), necessitating more training data and/or intensive classification strategies. We hypothesise that the relationship between image features and group size is more reliable and consistent on a local scale.

B. Perspective Normalisation

To account for perspective, a density map is calculated using the relative sizes of two reference persons. This is calculated in the same manner as [4]. The weight applied at pixel (i, j) to a two-dimensional feature is $W(i, j)$. For one-dimensional features, such as edges, the square root of the weight is applied.

C. Feature Extraction

Several features are extracted from each blob segment in order to estimate the number of people in the group. The features extracted are similar to those used in [9] and [4], taken locally. These features are:

- **Area:** The total pixel count for the blob segment, each pixel weighted by its value in the density map.

  $$B_{size} = \sum W(i, j)$$

  where $(i, j) \in B$, and B_{size} is the calculated area of blob B.
- **Perimeter:** The total pixel count for the blob's perimeter, each weighted by the square root of its value in the density map.
- **Perimeter-Area Ratio:** The ratio of perimeter to area, a measure of shape complexity [4].

- **Edges:** The total pixel count of edges within the blob, extracted from the image using Canny edge detection. Each pixel was weighted by the square root of its value in the density map.
- **Edge Angle Histogram:** The histogram of edge angles, obtained from the edge detection. Six histogram bins are used in the range $0°$ - $180°$ [9]. Each pixel's contribution to a histogram bin is the square root of its value in the density map.

D. Crowd Counting

The features extracted from each blob serve as inputs to a classifier. The output of the classifier is g_i, the group size estimate for the ith blob. A neural network was used to perform classification, as this has proven successful in previous research [12], [9]. In order to test whether local features can be classified using simpler strategies, a basic linear model was also tested:

$$g_i = w_0 + \sum_{n=1}^{N_F} w_n f_n \tag{1}$$

where w_n is the weight assigned to feature f_n, given N_F features. The weights are calculated using least squares regression. The total crowd estimate for a frame containing N_B blobs is then calculated:

$$C = \sum_{i=1}^{N_B} g_i \tag{2}$$

The estimate will vary from frame to frame as pedestrians enter and exit a scene simultaneously. A rapidly fluctuating estimate is not usable or accurate. A median filter provides smoothness and stability to the estimate, as well as making it robust against outlier estimates.

A median filter of length $2n + 1$ will select the median estimate from n frames either side of the frame in question. This is a non-causal filter which, implemented in practice, will introduce a delay of n frames. For this application we use a median filter of length 41 ($n = 20$). At a frame rate of 10 fps, the delay is 2.0 seconds.

IV. Algorithm Training and Ground Truth Annotation

The proposed algorithm is trained and tested on the data used in [4]. This database contains 2000 frames of pedestrian traffic moving in two directions. The video has been downsampled to 238×158 pixels and 10 fps, grayscale. An example frame is shown in Figure 1.

As the proposed algorithm calculates crowd size by determining the number of people in each blob, the ground truth annotation must specify a person count for each blob. As such, ground truth annotation is performed after foreground segmentation. A GUI was written which enables the operator to do this.

(a) Frame 1280.

(b) Foreground mask.

(c) Region of interest.

Figure 1. A frame from the testing database.

(a) Correct extraction of individuals, with additional noise (i.e. small bar near centre).

(b) Person (top, centre) is fragmented into two blobs, one of which is merged with nearby blob(s).

(c) Person (left) is fragmented into two blobs.

(d) Person (top, centre) blends into background leaving few foreground pixels. This person is barely visible to the human eye.

Figure 2. Typical errors in foreground extraction.

Ideally, a single blob will correspond to a whole number of people as shown in Figure 2(a). However, foreground segmentation on a low resolution grayscale image is prone to errors, examples of which are shown in Figure 2. There are three types of segmentation errors that can occur:

1) A single person is split into multiple foreground blobs (Figure 2(c)). In this case, the contribution of the person is split across multiple blobs, in direct proportion to the number of pixels contained in each blob (i.e. for a person fragmented into three blobs representing the upper body and each leg, the blobs may receive weights of 0.6, 0.2 and 0.2 for the upper body and each

leg respectively). The assignment of these weights is made by the computer according to the blob sizes.

2) Part of a person is split in isolation from the group they are with (Figure 2(b)). In this case, the contribution of the person is split across multiple (n) blobs equally ($1/n$ to each). Proportional contributions would not be suitable, because some fragments are merged with neighbouring blobs.
3) The motion detection fails to detect a person (Figure 2(d)). In this case, no assignment is made because the person has blended completely into the background so that very few, if any, foreground pixels are present. If this is a common occurrence, then the problem must be addressed at the segmentation stage (if possible). (In the database used there are only a small number of instances where this occurs, and these only occur in one part of the scene where the background is dark). Assuming it is a rare occurrence, no contribution is assigned to the faded person. The reason for this is that assigning a large weight to a tiny blob may lead to misclassification at other locations in the scene, where tiny blobs are merely products of noise, such as in Figure 2(a).

The correspondences between pedestrians and foreground blobs are entered via the GUI. The above scenarios and the methods for handling them are used throughout the ground truth process to ensure that labelling is performed in a consistent manner.

V. Evaluation and Results

A. Testing Criteria

The performance of the proposed system is assessed using three criteria:

1) Accuracy,
2) Scalability,
3) Practicality.

Accuracy is measured by comparing the detected number of pedestrians with the number annotated in the ground truth. Scalability is evaluated by using training and testing sets such that the types of crowds seen in testing are not present in the training set. Practicality is evaluated through the use of reduced training sets.

1) Accuracy: Although this system is trained on the basis of individual blobs, the testing still takes place on a holistic level. The accuracy of a system can be any measure of how closely the estimate follows the ground truth. The ground truth for the holistic crowd count was taken as the number of (x, y) person coordinates which lay within the region of interest. However, the exact point in time at which a person is deemed to have entered or exited a frame is never clearly defined. It may take several seconds between a pedestrian reaching the border of the region of interest, and being fully inside or outside of it.

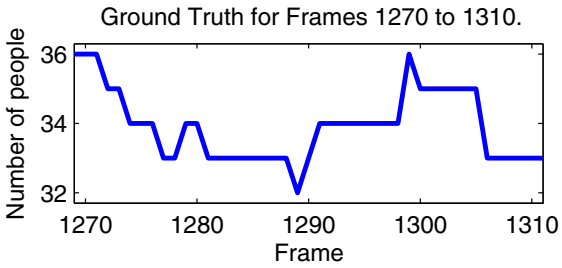

Figure 3. Ground truth for frames 1270 to 1310.

Figure 3 shows the number of people inside the region of interest over 40 frames (4.0 seconds). Based on the number of increments and decrements in this graph, there are at least 13 instances of pedestrians either entering or exiting the scene in this time. An example frame from this sequence is shown in Figure 1(a). The pedestrian at the bottom left in this sequence takes more than 30 frames to fully enter the scene. With groups entering and exiting the scene at this rate, yet taking several frames to do so, it would be difficult even for a human to estimate the exact crowd size, and impossible for them to remain consistent in their definition of what constitutes being 'in' or 'out' of the scene. In a scene such as this, where crowd size varies between 11 and 45 people, it is suggested that an estimate within 3 of the ground truth is acceptable. For testing purposes we consider the following measures of accuracy:

- **Error:** The mean value of the absolute difference between the crowd estimate and the ground truth.
- **MSE:** The mean value of the error squared.
- **Acceptability:** The percentage of frames for which the absolute error may be deemed 'acceptable', that is, less than or equal to 3.

2) Scalability: Ideally, the training data must cover a wide range of scenarios, similar to those which are expected to be found during operation. In the case of crowd counting, however, we may not have access to video footage of all possible scenarios. Excessive levels of over or under crowding may not be present in the training data because these events are abnormal, and this is the reason we wish to detect them. A system which cannot extrapolate in this context is of little practical use. We test the scalability of this system using two methods:

- Downscaling: The system is trained on large crowds, and tested on smaller crowds.
- Upscaling: The system is trained on small crowds, and tested on larger crowds.

3) Practicality: For a crowd counting system to be practical, it must be relatively easy to deploy. For real world deployment where the algorithm may be required run on several hundred different cameras within a single installation, being able to use a reduced training set is highly desirable. When training crowd counting algorithms, each

training frame requires ground truth to be supplied. If several hundred training frames are needed for each camera ([6] uses 150 frames, each taken 10 seconds apart for training; [4] uses 800 consecutive frames for training), then the process of training becomes very tedious and time consuming. To assess practicality, systems are evaluated using reduced training sets.

B. Systems Tested

Three crowd counting techniques are evaluated:

- **Proposed:** The system described here, in which local features are extracted for each blob and ground truth annotation is performed on a local level.
- **Equivalent Holistic System:** This is a system which utilises the same features as the proposed system, taken on a holistic rather than local level. Ground truth is also annotated on a holistic level.
- **Kong:** Blobs are sorted into six histograms of bin width 1500, as described in [9]. An edge angle histogram is also calculated, for which we use six histogram bins between 0° and 180°. This is also a holistic system.

For each system, two classifiers are tested: a neural network and linear model.

The results provided by [4] for this database can not be compared, as their estimate was calculated for pedestrians walking in either direction, rather than a total count. If the segmentation algorithm were changed from dynamic textures [5] to background subtraction, then the total count could be calculated. This would somewhat resemble the Equivalent Holistic System above, differentiated by the number of features.

C. Experimental Results

1) Accuracy: The accuracy of each system listed in Section V-B is tested. Frames 605, 610, ..., 1400 were designated for training (160 total) and testing was performed on frames 1-600 and 1401-2000. Those in the training set were annotated with ground truth counts for each blob, which was used to train the classifier. Neural network results differ slightly from test to test, therefore in order to determine a *typical* result for each system, the networks were retrained five consecutive times. The test which returned the median MSE for the filtered output was taken.

Results are tabulated in Table I. Results across the whole testing data set using the linear classifier are plotted in Figure 4.

By all three measures of accuracy, the proposed system significantly outperforms Kong and the equivalent holistic system. The mean error of the filtered estimate is 1.353 and the estimate is acceptable (within 3 of ground truth) 95.67% of the time (for the linear classifier). The linear classifier performs slightly better than the neural network, though similar performance trends are observed with the proposed system outperforming the other evaluated systems for a neural network classifier. The poorer performance of the neural network classifier can be attributed to the training data used. It is expected that for a larger training set, performance would equal or exceed that of the linear classifier.

2) Scalability: Scalability is tested in two steps, downscaling and upscaling. To test downscaling, frames 1205, 1210, ..., 1600 are designated for training (80 total), featuring crowds of size 30-45. These frames contain a mixture of large and small blobs. Testing is performed on frames 1-1200 and 1601-2000 (crowd sizes 11-40).

Due to the neural network's poor extrapolation capabilities, the holistic methods were unable to provide any meaningful results, as shown in Figure 5. The proposed system, trained on blobs of various sizes, was able to count smaller crowds.

The linear model is capable of superior extrapolation. The results in Table II indicate that all three systems can extrapolate downwards when linear fitting is used, however the proposed system is most accurate.

To evaluate upscaling, frames 805, 810, ..., 1100 were designated for training (60 total), featuring crowds of size 11-27[1]. Testing was performed on frames 1-800 and 1101-2000 (crowds 11-45). The blobs in the test set were larger than those in the training set, therefore all systems were unable to extrapolate when neural network classification was employed. As a result, evaluation results for the neural network classifier are not presented.

The linear model, however, is capable of extrapolation. Table III and Figure 6 illustrate the ability of the system to count crowds that are larger than those seen in the training set. It can be seen that the proposed algorithm is better equipped to deal with conditions that are unseen in the training set.

The superior performance on unseen conditions can be attributed to the manner in which the proposed algorithm counts crowds. As each blob is considered individually, the proposed algorithm only needs to have seen similar blobs in the training data. The holistic approaches however need to have seen a similar number of people overall in both and training and testing.

3) Practicality: The fewer training frames required of a system, the greater its practicality. While a neural network requires a large range of training data, the linear model can be calculated with very little. Given this, only a linear classifier is used in evaluating the systems practicality. The robustness of the proposed system is evaluated by testing the systems using only 10 training frames (640, 720, ..., 1360). For Kong [9], in order to supply all of the histogram bins with sufficient data, it was necessary to train the algorithm on 40 frames (620, 640, ..., 1400). Testing was performed on frames 1-600 and 1401-2000.

[1]The training range was widened for Kong $\{805, 810, ..., 1300\}$, so that the training data contained blobs large enough to contribute to each of the blob size histogram bins.

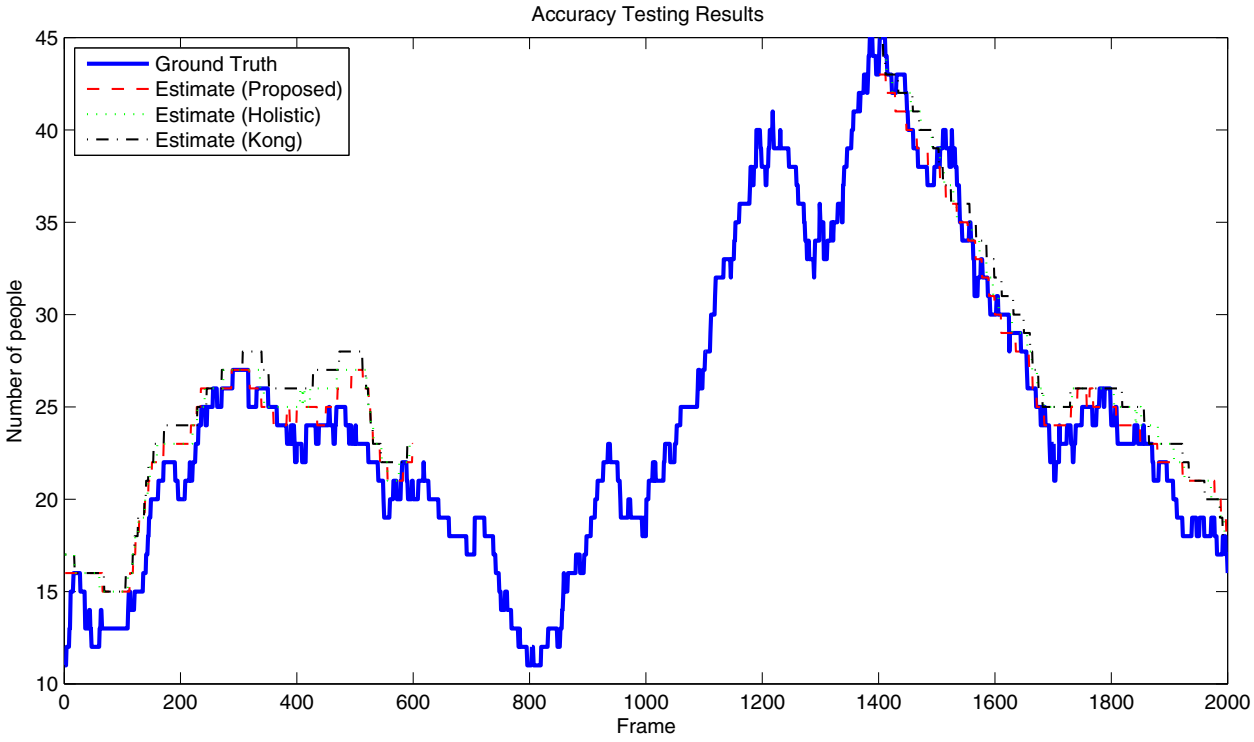

Figure 4. Accuracy testing results. Estimate is rounded and median filtered, and is shown for the test set only.

System	Classifier	Raw Estimate			Median Filtered		
		Error	MSE	Accept.	Error	MSE	Accept.
Proposed	NN	**1.889**	**5.646**	**86.75%**	**1.558**	**3.850**	**95.08%**
Kong	NN	2.976	15.158	68.08%	2.043	6.492	85.00%
Holistic	NN	2.570	9.962	74.08%	2.296	7.116	82.67%
Proposed	Linear	**1.525**	**3.666**	**88.00%**	**1.353**	**3.065**	**95.67%**
Kong	Linear	2.072	6.079	78.00%	2.013	5.447	88.83%
Holistic	Linear	1.798	4.720	84.42%	1.662	4.028	94.00%

Table I
ACCURACY TESTING RESULTS.

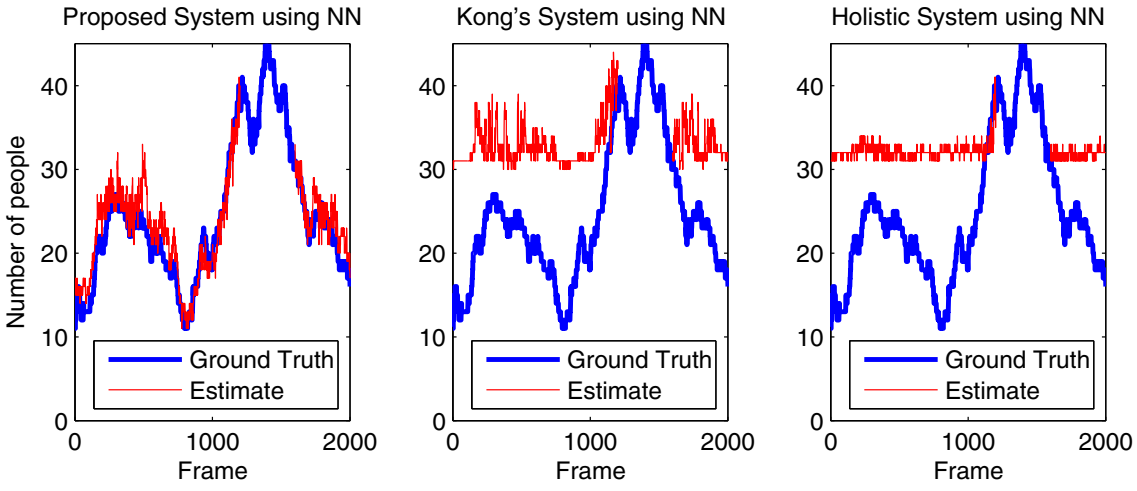

Figure 5. Downscaling testing results using neural network. Estimate has been rounded but not filtered.

System	Classifier	Raw Estimate			Median Filtered		
		Error	MSE	Accept.	Error	MSE	Accept.
Proposed	NN	**2.086**	**6.701**	**82.56%**	**1.881**	**5.532**	**86.63%**
Kong	NN	System failed. See Figure 5.					
Holistic	NN	System failed. See Figure 5.					
Proposed	Linear	**1.635**	**4.186**	**86.75%**	**1.537**	**3.674**	**92.81%**
Kong	Linear	2.659	10.074	59.31%	2.559	8.839	72.31%
Holistic	Linear	2.341	8.787	71.69%	2.194	7.938	80.44%

Table II
DOWNSCALING TESTING RESULTS USING LINEAR FITTING.

System	Training Set Size	Raw Estimate			Median Filtered		
		Error	MSE	Accept.	Error	MSE	Accept.
Proposed	60	**1.838**	**4.976**	**81.41%**	**1.654**	**4.075**	**93.65%**
Kong	100[1]	2.779	10.068	60.00%	2.749	9.34	73.53%
Holistic	60	2.524	8.581	63.47%	2.448	7.842	78.88%

Table III
UPSCALING TESTING RESULTS USING LINEAR FITTING.

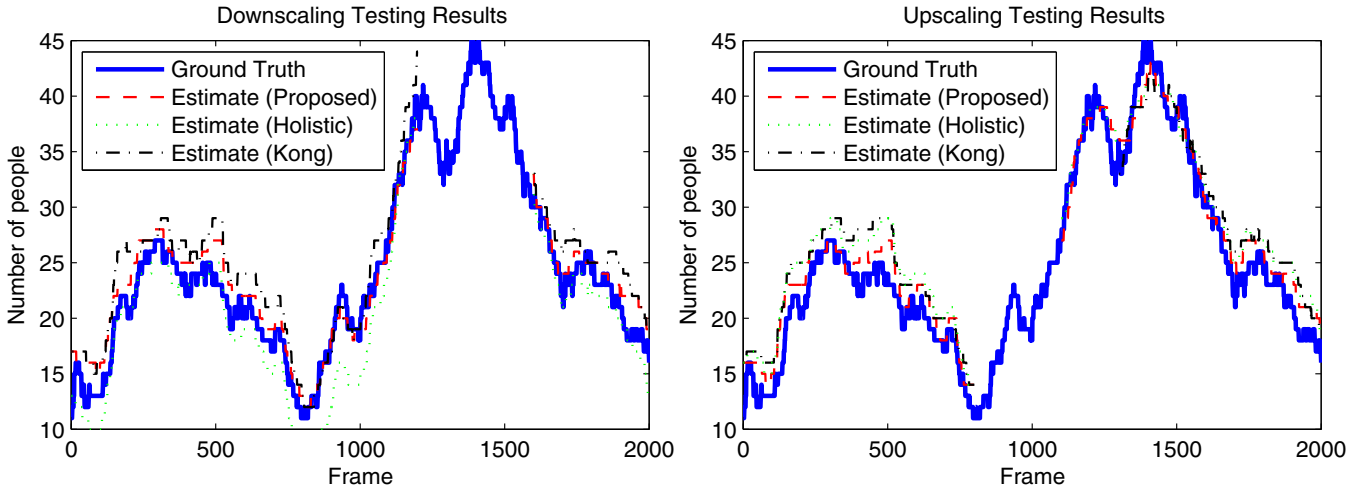

Figure 6. Downscaling and upscaling testing results (Linear Classifier Only).

Results are shown in Table IV. The proposed system outperforms the holistic systems using a limited training set, and achieves better results than when using a larger training set. The superior generalisation is likely due to the wider spacing of the training frames. These results indicate that the proposed system is highly practical, with accurate results obtained from as few as 10 frames of training data.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed the use of multiple local features for crowd counting. This approach reduces the task of crowd counting to the group level, so that the crowd estimate is the sum of its parts. By three standards (accuracy, scalability and practicality), the proposed system outperforms existing holistic methods of crowd counting. The proposed system is capable of extrapolating outside of the training range, and can also count crowds with minimal training (10 frames), demonstrating practicality. The ability to train the system from as few as 10 frames means it can be easily deployed in a real world setting consisting of a large number (possibly hundreds) of cameras with much greater ease than holistic approaches.

The use of local features also makes estimating local crowd density across the scene, and performing crowd counting across a network of multiple overlapping cameras possible. Analysing crowd densities at specific locations in a scene will enable the detection of local abnormalities. For example, a high-density crowd concentrated at one location may require attention, even if the holistic count for the scene is at a safe level. The use of multiple cameras will enable larger environments to be covered and monitored, as well as increasing accuracy in areas of overlap (due to the observations from multiple view points). Both these extensions will be investigated in the future. In addition,

System	Training Set	Raw Estimate			Median Filtered		
		Error	MSE	Accept.	Error	MSE	Accept.
Proposed	640,720,...,1360	**1.306**	**2.684**	**93.17%**	**1.047**	**1.902**	**99.25%**
Kong	620,640,...,1400	1.710	4.642	84.25%	1.352	3.200	93.75%
Holistic	640,720,...,1360	4.462	31.24	41.58%	3.538	17.788	57.83%

Table IV
PRACTICALITY TESTING RESULTS.

future work will also focus on capturing additional data for further testing, and evaluating the proposed algorithm in conditions where there is poor segmentation performance, reduced image resolution, and erroneous ground truth labelling.

REFERENCES

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–6, 2007.

[2] O. Arandjelović. Crowd detection from still images. *In Proc. British Machine Vision Conference*, 1:523–532, 2008.

[3] H. Celik, A. Hanjalic, and E. Hendriks. Towards a robust solution to people counting. *Image Processing, 2006 IEEE International Conference on*, pages 2401–2404, Oct. 2006.

[4] A. Chan, Z.-S. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. *CVPR 2008*, pages 1–7, June 2008.

[5] A. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926, May 2008.

[6] A. Davies, J. H. Yin, and S. Velastin. Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 7(1):37–47, Feb 1995.

[7] S. Denman, V. Chandran, and S. Sridharan. An adaptive optical flow technique for person tracking systems. *Elsivier Pattern Recognition Letters*, 28(10):1232–1239, 2007.

[8] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups. *Comput. Vis. Image Underst.*, 110(1):43–59, 2008.

[9] D. Kong, D. Gray, and H. Tao. A viewpoint invariant approach for crowd counting. *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 3:1187–1190, 2006.

[10] S.-F. Lin, J.-Y. Chen, and H.-X. Chao. Estimation of number of people in crowded scenes using perspective transformation. *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, 31(6):645–654, Nov 2001.

[11] A. Marana, L. Da Fontoura Costa, R. Lotufo, and S. Velastin. Estimating crowd density with minkowski fractal dimension. *ICASSP '99*, 6:3521–3524 vol.6, Mar 1999.

[12] A. Marana, S. Velastin, L. Costa, and R. Lotufo. Estimation of crowd density using image processing. *Image Processing for Security Applications (Digest No.: 1997/074), IEE Colloquium on*, pages 11/1–11/8, Mar 1997.

[13] O. Masoud and N. Papanikolopoulos. A novel method for tracking and counting pedestrians in real-time using a single camera. *Vehicular Technology, IEEE Transactions on*, 50(5):1267–1278, Sep 2001.

[14] N. Paragios and V. Ramesh. A mrf-based approach for real-time subway monitoring. In *2001 Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 1034–1040, Dec. 2001.

[15] H. Rahmalan, M. Nixon, and J. Carter. On crowd density estimation for surveillance. *Crime and Security, 2006. The Institution of Engineering and Technology Conference on*, pages 540–545, June 2006.

[16] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2:II–459–66 vol.2, June 2003.