# Statistical t+2D subband modelling for crowd counting

BHOWMIK, Deepayan <http://orcid.org/0000-0003-1762-1578> and WALLACE, Andrew

# STATISTICAL T+2D SUBBAND MODELLING FOR CROWD COUNTING

*Deepayan Bhowmik*[*]

Department of Computing,
Sheffield Hallam University,
Sheffield, S1 1WB, UK
deepayan.bhowmik@shu.ac.uk

*Andrew Wallace*

School of Engineering and Physical Sc.,
Heriot-Watt University
Edinburgh, EH14 4AS, UK
a.m.wallace@hw.ac.uk

## ABSTRACT

Counting people automatically in a crowded scenario is important to assess safety and to determine behaviour in surveillance operations. In this paper we propose a new algorithm using the statistics of the spatio-temporal wavelet subbands. A t+2D lifting based wavelet transform is exploited to generate a motion saliency map which is then used to extract novel parametric statical texture features. We compare our approach to existing crowd counting approaches and show improvement on standard benchmark sequences, demonstrating the robustness of the extracted features.

## 1. INTRODUCTION

With increases in population, mobility and urbanisation, there have been many fatal crowd related accidents *e.g.*, the Love Parade stampede, Germany (2010), the Santa Maria fire disaster, Brazil (2013) and the Hajj stampede (2015). Not surprisingly, crowd dynamics and behaviour analysis have received considerable attention from both social and lately the technical research disciplines, *e.g.*, signal and image processing. Various applications of crowd dynamics include *crowd management*, *surveillance*, *public space design*, and *virtual environments design* for simulation [1]. Crowds can be described with five fundamental characteristics [2], *i.e.*, *size, density, time* (acting together), *collectivity* (shared behaviour) and *novelty* (coherent action in unfamiliar situation). Many algorithms have been proposed for crowd analysis, *e.g.*, crowd segmentation, counting, abnormal behaviour detection and tracking. Here, we concentrate on the problem of counting the number of people in a crowd, which is important in safety and surveillance operations. The proposed approach does not track each individual member, as this is difficult, particularly due to occlusions and close proximity, complex in processing strategy, and not always necessary.

Previous algorithms proposed in the literature can be categorised into multiple groups [3], *e.g.*, counting by *1) detecting, 2) clustering*, *3) regression and 4) convolutional neural network (CNN)* of which algorithms consisting a) image features and regression and b) CNN exhibit better performances. Counting by detection implies that each individual has a distinct signature, for example partial detection, *e.g.*, head or shoulders [4]. Although these algorithms are tractable in relatively sparse scenes, they often fail in dense crowds. In *clustering* based approaches, a set of visual features are tracked to represent individuals or a group as independent moving entities, *e.g.*, Liang *et al.* [5] applied Speeded Up Robust Features (SURF). However, motion coherency of a moving group or crowd has been assumed which may not always be true due to variable direction or limb articulation. *Regression* based approaches [6, 7] avoid individual detection or tracking and rely on a holistic description to characterise a crowd. Therefore regression based methods are an option when detection and tracking fail in dense scenarios. Recently, CNN based approaches were proposed in the literature, *e.g.*, cross scene counting [8] and semantic informed dense feature mapped counting [9]. However, these rely on large annotated training datasets which are not often easy to acquire.

In this paper we propose a unique people counting method based on a spatio-temporal wavelet based saliency model, that segments the motion salient regions in the scene and extracts texture features by analysing the multi-resolution subbands using statistical models. The discrete wavelet transform (DWT) is a powerful tool for texture analysis [10]. The DWT decomposes an image into independent frequency subbands of multiple orientations at multiple scales demonstrating detail and structure. Recently, statistical modelling of wavelet transform coefficients has gained momentum in solving problems related to image texture analysis, *e.g.*, retrieval [11]. Our approach uses the DWT and identifies suitable statistical models to represent crowd texture patterns generated from a motion saliency map. The contributions of this paper are:

- Spatio-temporal wavelet decomposition of crowd scenes to segment motion salient regions; and
- A novel parametric approach using statistical subband modeling to extract unique texture based holistic features for crowd counting.

[*]We acknowledge the support of the UK Engineering and Physical Sciences Research Council (Grant Reference: EP K/009931/1).

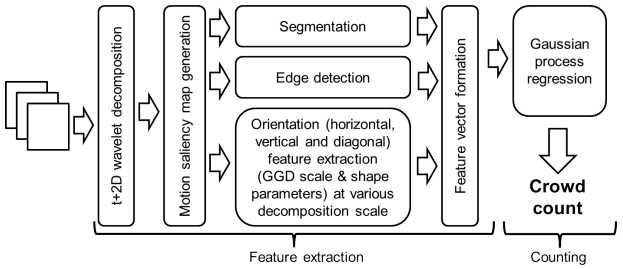


**Fig. 1**: Proposed algorithm.

## 2. MOTION SALIENT PARAMETRIC FEATURES

Our approach identifies subject movements such as directional movement, *e.g.*, walking or other small localised movements, *e.g.*, gesturing while standing. We detect and process such motions using hierarchical measurements of *pixel activity* in consecutive frames. Using a spatio-temporal wavelet transform, directional and localised motions of human subjects are derived in the high-frequency components of the temporal decomposition while the features are preserved in the spatial subbands. A block diagram of our algorithm is shown in Fig. 1. First, the input sequences are decomposed using t+2D wavelet transforms. A saliency model is then applied to the high frequency subbands to generate a motion saliency map which is used to segment and detect local edges in active regions of the scene. A set of parametric texture features (scale and shape of Generalised Gaussian Distribution (GGD)) are extracted at horizontal, vertical and diagonal orientations, at different resolutions of the spatial decomposed high frequency subbands. Traditional segmentation features such as area and local edges are also extracted and concatenated with the texture features to form a feature vector which is the input to a Gaussian process regressor (GPR). Finally, the regressor is trained and tested to count the crowd.

### 2.1. t+2D decomposition

The spatio-temporal wavelet decomposition can be either performed by a 3D wavelet transform or by temporal decomposition followed by a spatial transformation [12]. Inspired by the low complexity lifting schemes for wavelets [13], we use a lifting based spatio-temporal (3D) decomposition of the input frames. To enable multi-level 3D wavelet decomposition, the input frames are temporally decomposed and then organized in a hierarchical order followed by 2D spatial decomposition that allows us to identify the motion *active pixels* and related spatial texture information. Therefore we call this decomposition t+2D, where $t$ stands for temporal decomposition.

The formulation of the t+2D scheme follows a Haar wavelet decomposition. Let $I_t$ be the input video sequence, where $t$ is the time index. We consider two consecutive frames $I_t$ and $I_{t-1}$, as the current and reference frame, respectively. For the $[m, n]$ pixel location the prediction and update steps for temporal decomposition are defined in Eq. (1) and Eq. (2), respectively:

$$I'_{t-1}[m,n] = I_t[m,n] - I_{t-1}[m,n]. \quad (1)$$

$$I'_t[m,n] = I_t[m,n] + \frac{1}{2} I'_{t-1}[m,n]. \quad (2)$$

Finally lifting steps are followed by the normalization steps:

$$I''_t[m,n] = \sqrt{2} I'_t[m,n], \quad (3)$$

$$I''_{t-1}[m,n] = \frac{1}{\sqrt{2}} I'_{t-1}[m,n]. \quad (4)$$

The temporally decomposed frames $I''_t$ and $I''_{t-1}$ are the first level low and high pass subband frames, respectively. These steps are repeated for all the frames in the low pass subband frames to obtain the next level low and high pass subband frames, and are repeated to obtain the desired number of temporal decomposition levels. Similarly, the lifting based 2D transform is applied to obtain the desired spatio-temporal decomposition. We choose a bi-orthogonal 5/3 filter due to its proven decomposition performance within JPEG2000 image compression.

### 2.2. Motion saliency estimation and map generation

A saliency based model mimics human vision and helps to identify objects that visually stand out from the surroundings. In the proposed algorithm we generate a spatio-temporal wavelet based motion saliency map to extract a holistic feature set, used for counting people in a crowd. Wavelet transformation combines frequency domain analysis and scale-space decomposition to model visual saliency [14]. We use the higher frequency spatio-temporal wavelet decomposed subbands to generate the saliency map. First, the consecutive frames are wavelet decomposed in time to generate high frequency frames ($I''_{t-1}$). This captures the object and human motion within the scene. Next a multi level spatial 2D wavelet transform is applied to $I''_{t-1}$. An orientation map is then formed by combining centre-surround differences among horizontal, vertical and diagonal subbands across different resolutions of the spatial wavelet transform. As we are interested only in contrast at different orientations, absolute values of the coefficients are considered here. Finally, the saliency map $\hat{I}_t$ is produced by Eq. (5).

$$\hat{I}_t = \sum\nolimits_{l,\varnothing} f(C^l_\varnothing), \quad (5)$$

where $\hat{I}_t$ is the final saliency map at time $t$, $C^l_\varnothing$ represents higher frequency subbands of orientation $\varnothing \in$ Vertical (V), Horizontal (H), Diagonal (D) at resolution scale $l$ and $f()$ is an average filtering (to remove noise) and resize function.

### 2.3. Feature extraction

#### *2.3.1. Parametric texture features*

Texture features exhibit strong correlation with the number of people, particularly in high density regions. In this work

we extracted texture features using the generalised Gaussian distribution (GGD) of the spatial wavelet subbands, masked with the saliency map. The GGD is a parametric probability distribution that includes all Gaussian and Laplace distributions. The literature suggests that the histograms of the subband coefficients produced by various types of DWTs can be optimally modeled by adaptively varying the parameters of the GGD [15, 16]. The pdf of the GGD is defined as:

$$p(x;\mu,\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x-\mu|/\alpha)^{\beta}}, \tag{6}$$

where $-\infty < x < \infty$ is the detailed DWT coefficient value, $\Gamma(t) = \int_0^{\infty} u^{t-1}e^{-u}du$ is the Gamma function, $-\infty < \mu < \infty$ is the location parameter, $\alpha > 0$ the scale parameter that models the pdf peak and $\beta > 0$ is the shape parameter that is inversely proportional to the decreasing rate of the peak. As the detailed DWT transform coefficients theoretically sums to zero [15] we can comfortably define $\mu = 0$ in Eq. (6).

**Parameter estimation** The estimation of the GGD model parameters, *i.e.*, $\mu, \alpha$ & $\beta$, can be achieved by maximum-likelihood estimation (MLE). Varanasi and Aazhang [17] studied the accuracy of estimates using MLE for various samples with different sample sizes and shapes of the distribution and confirmed the usability of MLE for heavy-tailed distributions, *i.e.*, small $\beta$, as normally observed in the detailed coefficients of crowd images. We have briefly described here the MLE for GGD. Considering our sample $x = (x_1, x_2, ..., x_N)$, *i.e.*, the coefficients of the detailed DWT subbands, the likelihood function whose parameters $\alpha$ and $\beta$ are to be estimated, can be defined as:

$$L(x;\mu,\alpha,\beta) = \log \prod_{i=1}^{N} p(x_i;\mu,\alpha,\beta). \tag{7}$$

Considering $\mu = 0$, Eq. (7) can be modified using Eq. (6) as

$$L(x;\alpha,\beta) = N \log\left\{\frac{\beta}{2\alpha\Gamma(1/\beta)}\right\} - \sum_{i=1}^{N}(|x_i|/\alpha)^{\beta}. \tag{8}$$

A set of likelihood equations can be obtained using the partial derivatives of Eq. (8) that have a unique root to estimate the maximum likelihood parameters where $\Psi(.)$ is the digamma function ($\Psi(z) = \Gamma'(z)/\Gamma(z)$) [18]:

$$\frac{\partial L(x;\alpha,\beta)}{\partial \alpha} = -\frac{N}{\alpha} + \frac{\beta}{\alpha^{\beta+1}}\sum_{i=1}^{N}|x_i|^{\beta} = 0. \tag{9}$$

$$\frac{\partial L(x;\alpha,\beta)}{\partial \beta} = \frac{N}{\beta}\left\{\frac{\Psi(1/\beta)}{\beta} + 1\right\} - \sum_{i=1}^{N}\left(\frac{|x_i|}{\alpha}\right)^{\beta}\log\left(\frac{|x_i|}{\alpha}\right) = 0. \tag{10}$$

#### 2.3.2. Traditional features

In addition to statistical parametric features, we also take advantage of traditional segmentation features such as *Area ($A$)* and *edge ($G$)*.

**Area ($A$):** The saliency map provides motion active pixels in the scene. The most salient regions are then segmented into two levels using Otsu's [19] adaptive thresholding algorithm. The segmented area ($A$) is calculated by counting the number of pixels present within the motion active regions.

**Edge feature ($G$):** The edge map of the salient region is extracted by applying a Sobel operator to the motion map. The saliency map preserves the local characteristics of the moving objects and is thus used to extract edge features. Finally the edge feature ($G$) is calculated by counting the number of edge pixels.

#### 2.3.3. Feature vector formation

Along with area and edge features the scale ($\alpha$) and shape ($\beta$) parameters of individual wavelet subbands at each decomposition level are considered as features in this work. We advocate that the crowd density can be characterised by the parametric features of the oriented subbands at multiple resolutions. The features, $\mathcal{F}$, of the subbands, grouped by orientation, are defined in vector form as:

$$\mathcal{F}_{\mathrm{V}^{(\varnothing)}} = \left(\mathrm{V}_{\alpha}^{(\varnothing)}, \mathrm{V}_{\beta}^{(\varnothing)}\right), \quad \mathcal{F}_{\mathrm{H}^{(\varnothing)}} = \left(\mathrm{H}_{\alpha}^{(\varnothing)}, \mathrm{H}_{\beta}^{(\varnothing)}\right),$$
$$\mathcal{F}_{\mathrm{D}^{(\varnothing)}} = \left(\mathrm{D}_{\alpha}^{(\varnothing)}, \mathrm{D}_{\beta}^{(\varnothing)}\right). \tag{11}$$

Finally a feature vector was formed by concatenating the features, into $\mathcal{F} \in \mathbb{R}^d$, which is used as the input to the regression model described in Section 2.4. The final $d$-dimensional feature vector, where $d = 2 + 3 \cdot 2 \cdot L$ ($L$ is the number of spatial decomposition levels), considering area ($A$), edge ($G$) and three oriented subbands consisting of two parameters at each resolution scale, is expressed as:

$$\mathcal{F} = \left(A, G, \mathcal{F}_{\mathrm{V}^{(\varnothing)}}, \mathcal{F}_{\mathrm{H}^{(\varnothing)}}, \mathcal{F}_{\mathrm{D}^{(\varnothing)}}\right). \tag{12}$$

### 2.4. Gaussian process regression

The extracted feature set is a good predictor of the number of people in a selected region. Our basic assumption is that these features, captured at different orientations at multiple resolutions have local deviations which are not necessarily linear due to occlusions. This indicates the need for a regression framework that handles multiple features with local non-linearity in a high-dimensional space and can accurately model crowd counts. In this work we rely on the Gaussian process regression (GPR)framework proposed by Rasmussen and Williams [20].

GPR is a Bayesian approach and a distribution over functions that can construct a real process $f(\mathbf{x})$ of a feature vector

$\mathbf{x} \in \mathbb{R}^d$ from a training sample. A Gaussian process ($\mathcal{GP}$) is a collection of random variables that has a set of finite numbers with joint Gaussian distribution [20] and can be specified by its mean $m(\mathbf{x})$ and covariance functions $k(\mathbf{x}, \mathbf{x'})$:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x'})), \tag{13}$$

where

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x'}) &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x'}) - m(\mathbf{x'}))]. \end{aligned} \tag{14}$$

Finally, the target count $y$ can be a model for prediction by Eq. (15) considering that $f(\mathbf{x})$ is linear in the transformation space:

$$y = f(\mathbf{x}) + \epsilon, \tag{15}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is an independent identically distributed (i.i.d.) Gaussian noise.

The functions, approximated by GPR, rely on the covariance (also referred as kernel function). We have used a squared-exponential kernel $k_r(\mathbf{x}, \mathbf{x'})$ producing a regression model that can handle local non-linearities within the feature space. This covariance function is also called the Radial Basis Function (RBF) and can be expressed as:

$$k_r(\mathbf{x}, \mathbf{x'}) = \theta_1^2 e^{-(1/\theta_2^2)||\mathbf{x}-\mathbf{x'}||^2}, \tag{16}$$

where $\theta_i$ are the covariance hyperparameters.

## 3. EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithm we used the popular benchmark dataset *Mall*. The Mall pedestrian database was introduced by Chen *et al.* [6] and contains 2000 annotated frames captured inside a cluttered indoor shopping centre. A split of 800 vs 1200 frames were allocated between training and testing, respectively; following the original test conditions, *i.e.*, the first 800 frames for Mall database were used for training. In our experiment, we trained the regressor using the feature vector and corresponding GT and then evaluated the regressor on the unseen data. To handle perspective problems, frames were divided into four non-overlapping region.

Two different commonly used evaluation metrics are used here: 1) *mean absolute error* (MAE) and 2) *mean-square error* (MSE). These are defined as:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|; \quad \text{MSE} \quad = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2, \tag{17}$$

where $N$ is the total number of test frames, $y_n$ is the actual number of people and $\hat{y}_n$ is the estimated count.

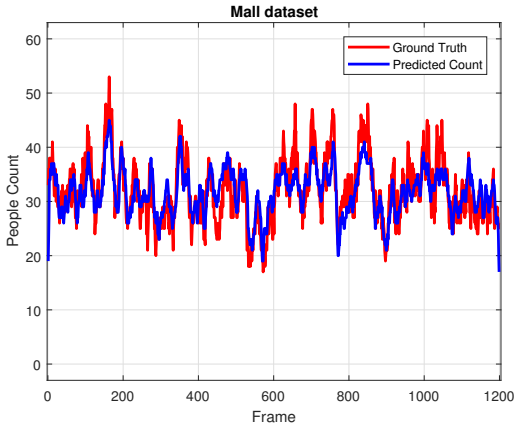


**Fig. 2**: Frame by frame crowd counting result on Mall dataset.

| | Mall | | | | |
|---|---|---|---|---|---|
| Metric | MORR [6] | IIS-LDL [7] | LAF+VALD [9] | CS-SLR [21] | Our |
| MAE | 3.15 | 2.69 | 2.86 | 3.23 | 2.72 |
| MSE | 15.7 | 12.1 | 13.05 | 15.77 | 12.28 |

**Table 1**: Comparison with state-of-the-art algorithms.

Frame by frame results are shown in Fig. 2. We also compare the performance for Mall dataset. The results are reported in Table 1. The results show either better or comparable performances over the existing methods. This is because regular textured structures are formed with higher people counts, which can be robustly represented by the parametric features estimated from the GGD. The motion saliency map provides reasonably accurate information on subject motions, resulting in better segmentation and edge pixel estimations. The proposed algorithm outperformed the state-of-the-art CNN LAF+VALD [9] approach. Only IIS-LDL [7] reported a better result but this relies on an additional label distribution from neighboring class labels.

## 4. CONCLUSIONS

We have proposed a new people counting algorithm which can be used in a crowded scenario. Unlike existing regression based methods, our approach focuses on a new set of low-level features derived from wavelet decomposition. First, we decompose the input frames using a lifting based spatio-temporal (t+2D) wavelet transform. Then, we segment motion salient regions by applying a frequency domain model. Our texture feature set is derived by using a statistical parametric approach. A Gaussian process regressor is used to train and estimate the number of people. The algorithm has been evaluated against existing algorithms and exhibits improved performance, especially for higher density crowds, demonstrating the advantage of using the features we extract.

## 5. REFERENCES


[1] J. C. S. Jacques-Junior, S. Raupp Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010.

[2] R. Challenger, C. W. Clegg, and M. A. Robinson, "Understanding crowd behaviours," *UK Cabinet office. Crown*, 2011.

[3] D. Ryan, S. Denman, S. Sridharan, and C. Fookes, "An evaluation of crowd counting methods, features and regression models," *Computer Vision and Image Understanding*, vol. 130, pp. 1–17, 2015.

[4] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Proc. IEEE Int'l Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, Sept 2012, pp. 470–475.

[5] R. Liang, Y. Zhu, and H. Wang, "Counting crowd flow based on feature points," *Neurocomputing*, vol. 133, pp. 377–384, 2014.

[6] K. Chen, C.-C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. British Machine Vision Conference (BMVC)*, 2012, pp. 21.1–21.11.

[7] Z. Zhang, M. Wang, and X. Geng, "Crowd counting in public video surveillance by label distribution learning," *Neurocomputing*, vol. 166, pp. 151–163, 2015.

[8] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE CVPR*, 2015, pp. 833–841.

[9] B. Sheng, C. Shen, G. Lin, J. Li, W. Yang, and C. Sun, "Crowd counting via weighted VLAD on dense attribute feature maps," *IEEE Trans. on Circuit and Systems for Video Technology*, 2016.

[10] E. de Ves, D. Acevedo, A. Ruedin, and X. Benavent, "A statistical model for magnitudes and angles of wavelet frame coefficients and its application to texture retrieval," *Pattern Recognition*, vol. 47, no. 9, pp. 2925 – 2939, 2014.

[11] R. Kwitt and A. Uhl, "Lightweight probabilistic texture retrieval," *IEEE Trans. on Image Processing*, vol. 19, no. 1, pp. 241–253, Jan 2010.

[12] S.-J. Choi and J.W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb 1999.

[13] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *Journal of Fourier Analysis and Applications*, vol. 4, no. 3, pp. 247–269, 1998.

[14] D. Bhowmik, M. Oakes, and C. Abhayaratne, "Visual attention-based image watermarking," *IEEE Access*, vol. 4, pp. 8002–8018, 2016.

[15] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, Jul 1989.

[16] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Trans. on Image Processing*, vol. 11, no. 2, pp. 146–158, Feb 2002.

[17] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1404–1415, 1989.

[18] M. Abramowitz and I. A. Stegun, "Handbook of mathematical functions with formulas, graphs, and mathematical tables. national bureau of standards applied mathematics series 55. tenth printing.," 1972.

[19] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[20] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[21] X. Huang, Y. Zou, and Y. Wang, "Cost-sensitive sparse linear regression for crowd counting with imbalanced training data," in *Proc. IEEE ICME*, 2016, pp. 1–6.