Statistical t+2D Subband Modelling for Crowd Counting

Deepayan Bhowmik[1] & Andrew Wallace[2]

[1]Department of Computing, Sheffield Hallam University, UK
[2]School of Engineering and Physical Sc., Heriot-Watt University, UK

d.bhowmik@ieee.org, 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 statistical texture features.

Introduction

With increases in population, mobility and urbanisation, there have been many fatal crowd related accidents *e.g.*, the Santa Maria fire disaster, Brazil (2013) and the Hajj stampede (2015). Crowd dynamics and behaviour analysis have received considerable attention from both social and 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. Here, we concentrate on the problem of counting the number of people in a crowd, which is important in safety and surveillance operations. This work 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.

Main Contributions

This paper proposes 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. Main contributions of this paper are:

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

Motion salient parametric features

The algorithm identifies subject movements, *e.g.*, directional movement or other small localised movements. We detect & process such motions using hierarchical measurements of *pixel activity* in consecutive frames. Using a spatio-temporal wavelet transform, directional and localised motions of subjects are derived in the high-frequency components of the temporal decomposition while the features are preserved in the spatial subbands.

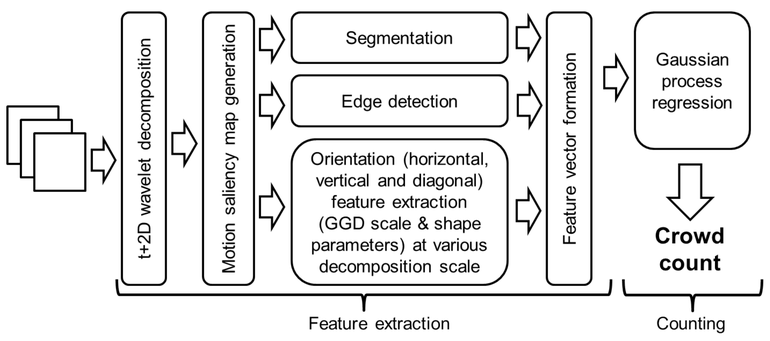

Figure 1: Flow diagram of the proposed algorithm.

t+2D 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 and the prediction & update steps are:

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

The final decomposition is achieved by a following spatial decomposition.

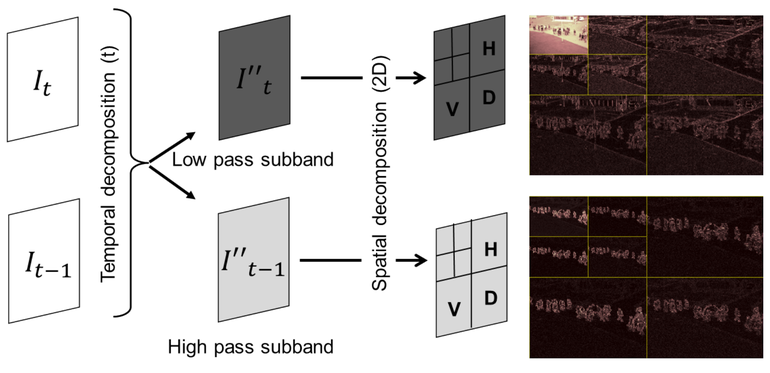

Figure 2: Example of wavelet based t+2D decomposition.

Feature extraction

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. The histograms of the subband coefficients can be optimally modeled by adaptively varying the parameters of the GGD. The pdf of the GGD is:

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

where $\alpha > 0$ is 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. We use α and β to form the feature vector. In addition to statistical parametric features, we also take advantage of traditional segmentation features such as *Area (A)* and *edge (G)*.

Along with area and edge features the scale (α) and shape (β) 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:

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

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:

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

Results

We evaluated the algorithm on the popular benchmark dataset *Mall*. The Mall pedestrian database 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. 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.

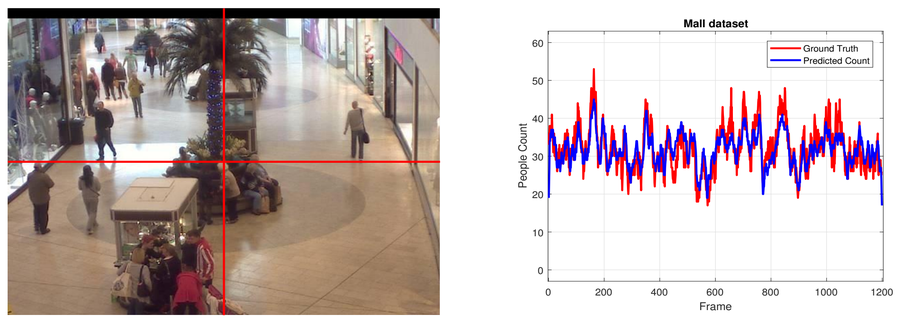

Figure 3: Sample frame from the dataset and frame by frame crowd counting result.

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.

	Mall				
Metric	MORR (Chen'12)	IIS-LDL (Zhang'15)	LAF+VALD (Sheng'16)	CS-SLR (Huang'16)	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.

Conclusions

A new people counting algorithm is proposed for crowded scenarios. Our approach focuses on a new set of low-level features derived from the wavelet decomposition. First, we decompose the frames using a 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 exhibits improved performance, especially for higher density crowds, demonstrating the advantage of using the features we extract.

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