Deep Residual Network with Subclass Discriminant Analysis for Crowd Behaviour Recognition

ICIP 2018

Bappaditya Mandal[1], Jiri Fajtl[2], Vasileios Argyriou[2], Dorothy Monekosso[3] and Paolo Remagnino[2]

[1] Keele University, Staffordshire, United Kingdom,

[2] Kingston University London, Surrey, United Kingdom

[3] Leeds Beckett University, West Yorkshire, United Kingdom

Outline

- **Introduction - Motivation**
  - Related work and systems
  - System overview of general Crowd Behavior Monitoring
- **Proposed System**
  - System overview
  - Deep Learning Features and Fine-tuning
  - Subclass Partitioning and Discriminant Analysis
  - Regularization of the Subclasses
  - Feature Selection and Dimensionality Reduction
- **Experimental Results**
  - Construction of violent crowd behavior dataset
- **Conclusions**
  - Conclusion and Summary

Introduction & Motivation

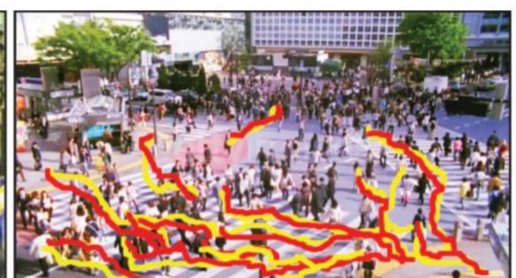
Crowd monitoring using videos plays an important role at public events such as concerts, sport matches, event celebrations and protests, public gatherings at stations.

A large number of people die every year in very crowded environments, such as the **Mumbai railway station** 2017 stampede which killed 22 people and injured 30 people [1] and the New Year's Eve 2015 celebration in **Shanghai**, where a stampede tragically left 36 people dead and nearly 50 others injured [2].

Introduction & Motivation

> For human observers, it is extremely difficult to monitor a very large number of individuals, their behaviours and activities from a large topology of cameras.

> The affected areas are generally highly congested urban areas and **extracting useful behaviour pattern information** has become of paramount importance for public security, safety, crowd management, providing timely critical decisions and support.

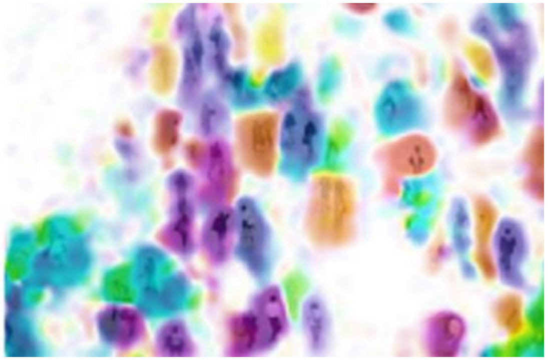


Introduction & Motivation

Existing research is mainly focused on **sparse** and mostly **staged** scenes, relatively little effort has been devoted to reliable classification and understanding of human activities in real and very crowded scenes.

In general, researchers have proposed two ways of analysing behaviour in such complex scenes.

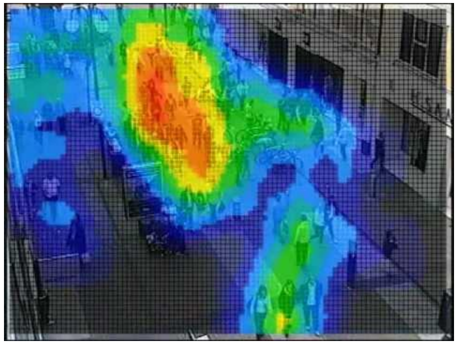
1. Holistic approaches
2. Object based approaches



<u>Introduction & Motivation</u>

The first approach considers the crowd and scene targets as a **whole**, where individual targets such as objects, places, scenes, their actions or interactions are **<u>not</u> identified or classified individually**, rather they are processed based on their whole appearance.

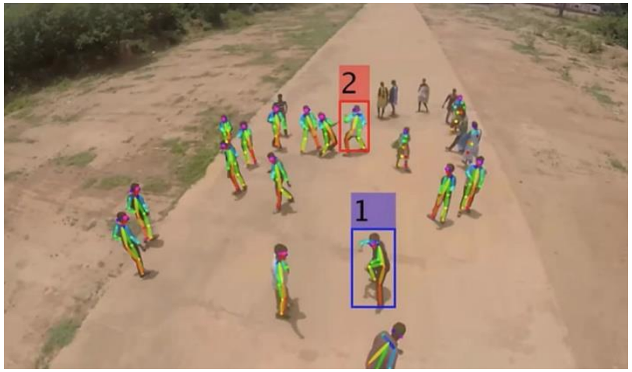
It is often advantageous and simpler to understand the crowd behaviour without knowing the actions of the individuals.



Introduction & Motivation

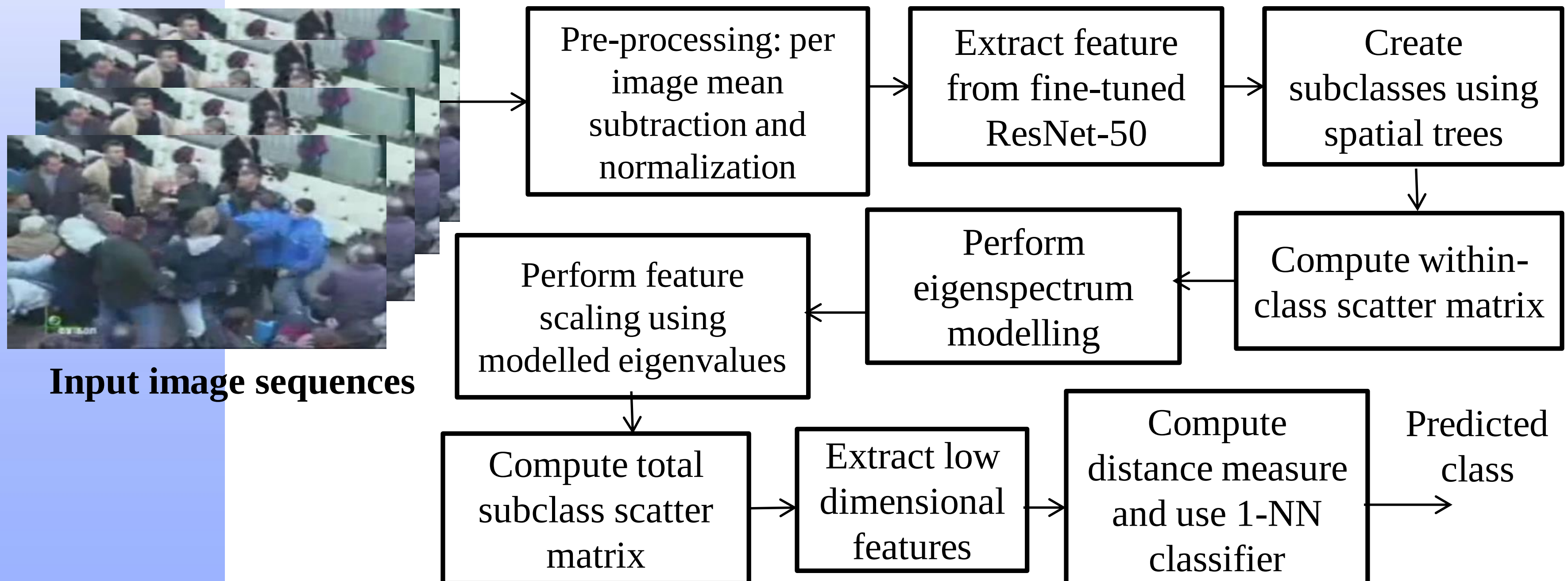
The object based approaches, where individuals (humans and/or objects) are detected and segmented to perform motion and/or behaviour analysis.

This kind of complex segmenting and tracking individuals in crowded videos is a challenging task.



Proposed System

In our work we use a holistic approach, where crowd behaviour patterns are perceived holistically



Proposed System

The motivation to use pre-trained deeply learned residual models for crowd behaviour analysis was that this kind of architectures solve both the vanishing gradient and over-fitting problems.

In our work, we **first fine tune the network** using crowd behaviour videos and then extract rich representations of the pattern of specific crowd behaviour.

Proposed System

The fine-tuned ResNet-50 is represented as

$$\begin{aligned}\Theta_R = I(224, 224, 3) \rightarrow C(7, 2, 64) \rightarrow P(2, 3) \rightarrow \\ 3 \times R(C(1, 1, 64) \rightarrow C(3, 1, 64) \rightarrow C(1, 1, 256)) \rightarrow \\ R(C(1, 2, 128) \rightarrow C(3, 2, 128) \rightarrow C(1, 2, 512)) \rightarrow \\ 3 \times R(C(1, 1, 128) \rightarrow C(3, 1, 128) \rightarrow C(1, 1, 512)) \rightarrow \\ R(C(1, 2, 256) \rightarrow C(3, 2, 256) \rightarrow C(1, 2, 1024)) \rightarrow \\ 5 \times R(C(1, 1, 256) \rightarrow C(3, 1, 256) \rightarrow C(1, 1, 1024)) \rightarrow \\ R(C(1, 2, 512) \rightarrow C(3, 2, 512) \rightarrow C(1, 2, 2048)) \rightarrow \\ 2 \times R(C(1, 1, 512) \rightarrow C(3, 1, 512) \rightarrow C(1, 1, 2048)) \rightarrow \\ P^*(1, 7) \rightarrow F(e) \rightarrow Softmax\end{aligned}$$

- C(filter size, strides, filter banks) represents the **convolutional layer**
- P(strides, subsampling) represents the **pooling layer**.
- Each convolutional layer is followed by a **batch normalization** layer and **RELU** as a non-linearity function.
- Summations at the end of each residual unit are followed by a ReLU unit.

Proposed System

The fine-tuned ResNet-50 is represented as

$$\begin{aligned}\Theta_R = I(224,224,3) \rightarrow C(7,2,64) \rightarrow P(2,3) \rightarrow \\ 3 \times R(C(1,1,64) \rightarrow C(3,1,64) \rightarrow C(1,1,256)) \rightarrow \\ R(C(1,2,128) \rightarrow C(3,2,128) \rightarrow C(1,2,512)) \rightarrow \\ 3 \times R(C(1,1,128) \rightarrow C(3,1,128) \rightarrow C(1,1,512)) \rightarrow \\ R(C(1,2,256) \rightarrow C(3,2,256) \rightarrow C(1,2,1024)) \rightarrow \\ 5 \times R(C(1,1,256) \rightarrow C(3,1,256) \rightarrow C(1,1,1024)) \rightarrow \\ R(C(1,2,512) \rightarrow C(3,2,512) \rightarrow C(1,2,2048)) \rightarrow \\ 2 \times R(C(1,1,512) \rightarrow C(3,1,512) \rightarrow C(1,1,2048)) \rightarrow \\ P^*(1,7) \rightarrow F(e) \rightarrow Softmax\end{aligned}$$

- Each **repetitive residual unit** is presented inside **R**.
- **F(E)** denotes the **fully connected layer** where E is the number of neurons. The length of F(E) depends on the number of categories E.
- P* refers to **average pooling** rather than max pooling as used else.
- The **softmax** function (or normalized exponential function) is used

Proposed System

Random projection (RP) and principal component analysis (PCA) trees are used to partition each crowd behaviour class into subclasses.

After the subclass creation, crucial intra-class variance information is learned by computing the within-subclass scatter matrix and optimization is performed using Fisher criterion.

$$J(\Psi) = \frac{tr(\Psi^T S_{bs} \Psi)}{tr(\Psi^T S_{ws} \Psi)}$$

In this work we propose to use Fisher objective Function as

$$J(\Psi) = \frac{tr(\Psi^T S_{ts} \Psi)}{tr(\Psi^T S_{ws} \Psi)}$$

Proposed System

S_{ws} is the within-subclass scatter matrix

$$S_{ws} = \sum_{i=1}^{E} p_i \sum_{j=1}^{H_i} \frac{q_{H_i}}{G_{ij}} \sum_{k=1}^{G_{ij}} (x_{ijk} - \mu_{ij})(x_{ijk} - \mu_{ij})^T$$

S_{ts} is the total subclass scatter matrix

H_i denotes the number of subclasses of the ith class

G_{ij} denotes the number of samples in jth subclass of ith class.

x_{ijk} is the kth image vector in jth subclass of ith class.

μ_{ij} is the sample mean of jth subclass of the ith class

pi =1/E and q_{Hi} =1/H_i aree the estimated prior probabilities.

Proposed System

The total scatter matrix S_{ts} of the regularized training data is employed to extract the discriminative features because of its greater noise tolerance as compared to S_{bs}.

$$\tilde{S}_{ts} = \sum_{i=1}^{E} \frac{p_i}{n_i} \sum_{j=1}^{n_i} (\tilde{y}_{ij} - \tilde{\mu})(\tilde{y}_{ij} - \tilde{\mu})^T$$

The transformed features $\tilde{y}_{ij} = \tilde{\mathbf{\Psi}}_l^{wsT} x_{ij}$

$$\tilde{\mathbf{\Psi}}_l^{ws} = [\tilde{\omega}_k^{ws} \psi_k^{ws}]_{k=1}^{l}$$

The scaling function is $\tilde{\omega}_k^{ws} = \frac{1}{\sqrt{\tilde{\lambda}_k^{ws}}}$

$$\tilde{\lambda}_k^{ws} = \begin{cases} \lambda_k^{ws}, & k < m \\ \frac{\alpha}{k+\beta}, & m \le k \le r_{ws} \\ \frac{\alpha}{r_{ws}+1+\beta}, & r_{ws} < k \le l \end{cases}$$

Proposed System

Selecting the eigenvectors with the **d** largest eigenvalues, $\tilde{\Psi}_d^{ts} = [\tilde{\psi}_k^{ts}]_{k=1}^{d}$

the proposed feature scaling and extraction matrix is given by $\mathbf{U} = \tilde{\Psi}_l^{ws} \tilde{\Psi}_d^{ts}$

which transforms a crowd behavior image vector x, into a feature vector z = U^T x.

>To compare two behaviour events of different lengths we use dynamic time warping (DTW).

>Cosine distance measure and the first nearest neighbourhood classifier (1-NNK) are applied for crowd behaviour recognition.

Experimental Results

We use WWW (where, who and why questions) Crowd Database comprising of 10,000 videos from 8,257 scenes. The WWW has 94 crowd-related annotated attributes, such as stadium, concert, stage, fight, mob, parade, and others, to describe each video in the database.

We selected a few normal crowd videos (like waking, skating, graduation, and others) and 4 violent crowd behaviour videos, such as **fight**, **protest**, **mob** and **protester** from this large database.

Attributes	Normal	Fight	Mob	Protest	Protester
# Images	$15,631$	$14,059$	$14,609$	$87,241$	$87,554$

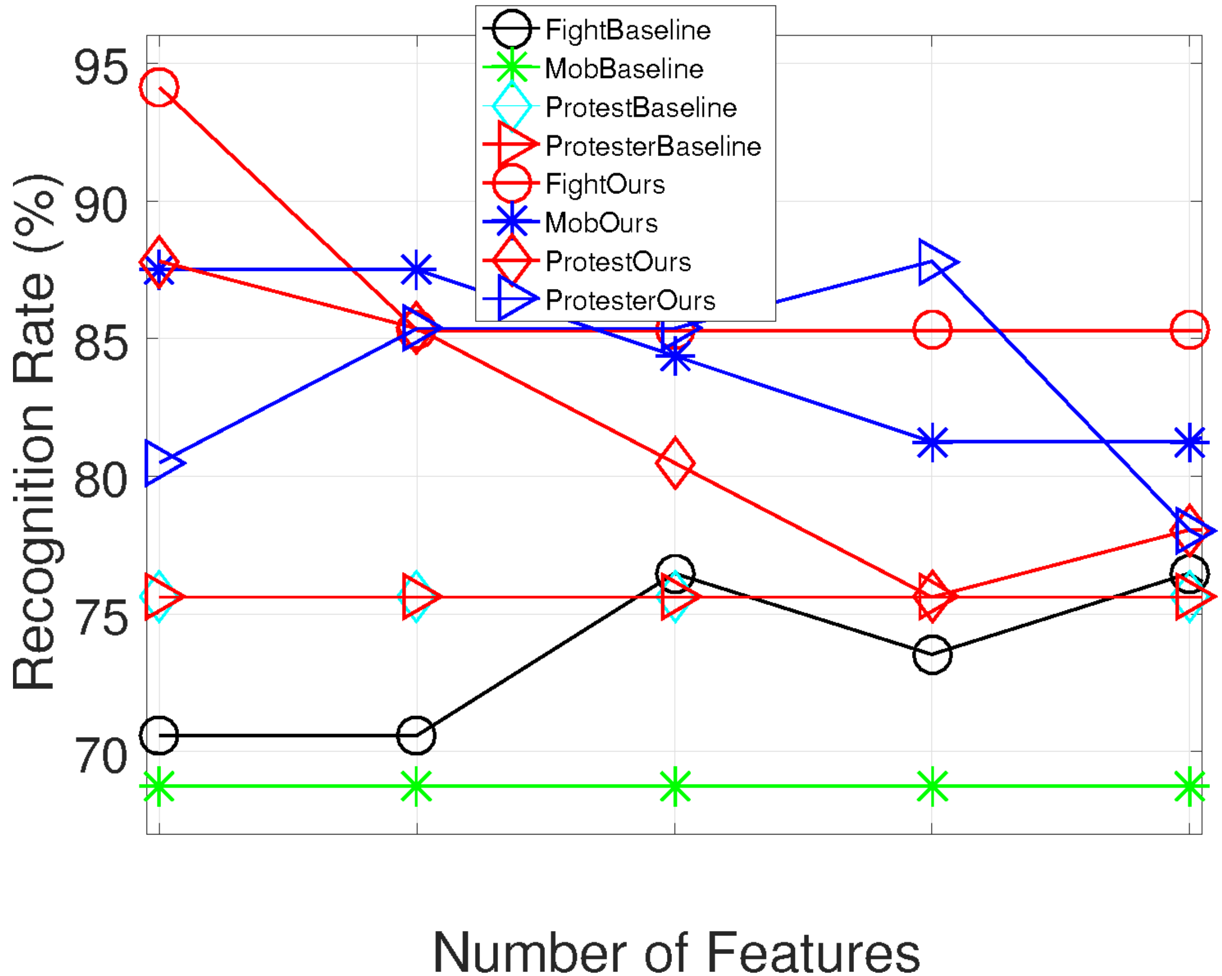
Experimental Results

Sample images from the violent crowd



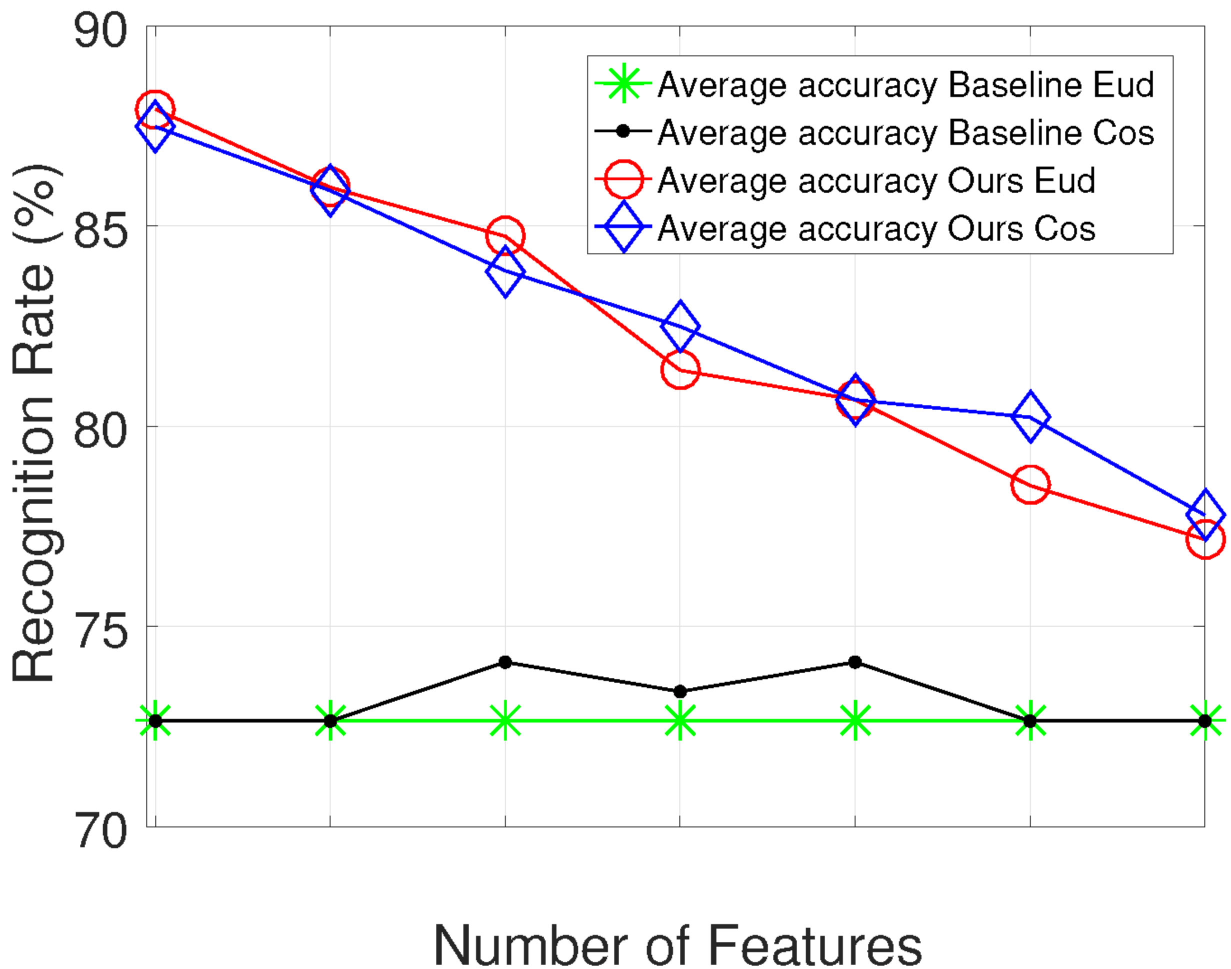
Experimental Results

Recognition rate (%) on WWW crowd database.



Experimental Results

Average recognition rate (%) on WWWcrowd database



Experimental Results

Crowd behavior recognition AUCs on WWW crowd database. ResnetCrowd[1] and ResnetCrowd[2] represent single task and multi-task respectively.

Methods	Fight	Mob	Protest	Protester	Average
Baseline	0.87	0.82	0.83	0.89	0.85
Shao *et al.* [8]	0.93	0.91	0.95	0.97	0.94
ResnetCrowd[1] [6]	0.62	0.68	–	–	0.65
ResnetCrowd[2] [6]	0.71	0.77	–	–	0.74
Our Proposed	0.95	0.94	0.96	0.96	0.95

Conclusions

- This paper proposes a fine-tuned deep convolutional neural residual network framework that creates subclasses in the feature maps of each of the crowd behaviour attribute classes using spatial partitioning trees.

- Eigen feature regularization using eigenmodel is used to weigh the features of the whole intra-subclass eigenspace of the crowd behaviour videos. This has helped to model the variance appearing from the intra-subclass variance information.

Conclusions

- Low dimensional discriminative features are extracted using total subclass scatter matrix along with dynamic time warping is used on the cosine distance measure to find the similarity measure between the two videos for crowd behaviour recognition task.

- Experimental results on a large crowd behaviour video database show the superiority of our proposed framework compared to the baseline and current state-of-the-art methodologies.