Counting in Dense Crowds using Deep Features

Karunya Tota, Haroon Idrees
(tota2@illinois.edu, haroon@cs.ucf.edu)

UCF Center for Research in Computer Vision

National Science Foundation
Research Experience for Undergraduates

Motivation:

Crowds occur in a variety of situations (i.e. concerts, political speeches, rallies, stadiums) and manually counting such dense crowds is a highly tedious task.

Applications:

- Crowd Management
- Safety
- Surveillance

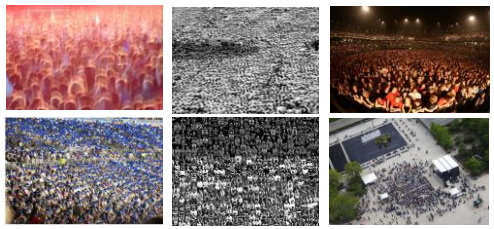
Difficulties:

- Occlusion
- Clutter
- Low Resolution
- Foreshortening
- Irregular Arrangement

Research Goals:

1. Extract features from a pre-trained convolutional neural network (CNN) and train a Support Vector Machine afterwards.
2. Extract dense SIFT features from images and encode Fisher Vectors using these features. Then, train another Support Vector Machine.
3. Future Work: possibly explore more options to further improve the performance of the counting algorithm (i.e. horizontal line estimation).

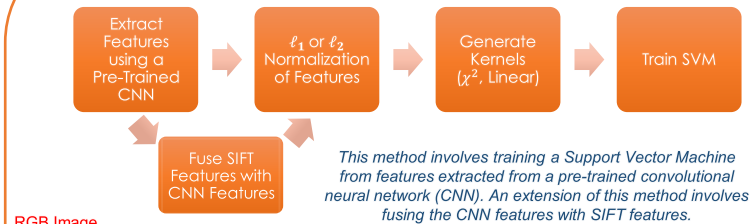
Crowd Dataset:



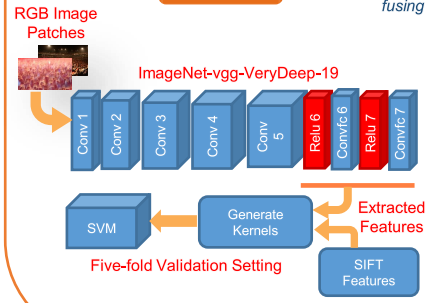
CVPR'13 UCF Crowd Counting Dataset

Our dataset includes 50 pre-annotated dense crowd images with counts ranging from 96 to 4633.

Method 1: CNN Features



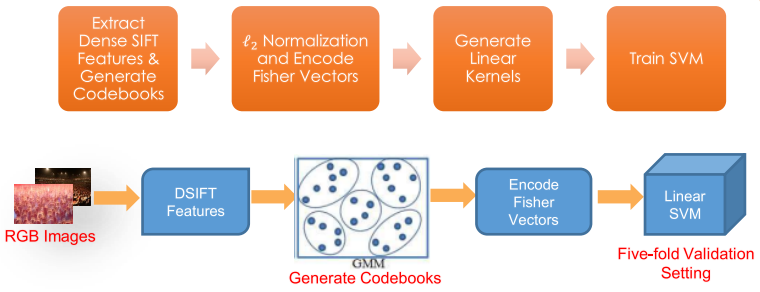
This method involves training a Support Vector Machine from features extracted from a pre-trained convolutional neural network (CNN). An extension of this method involves fusing the CNN features with SIFT features.



Network Structure:

- We used the ImageNET-vgg-verydeep-19 model [2] and the AlexNET model [3] as our pre-trained networks which are accessible through the MatConvNet toolbox. [4]
- This model is 19 layers deep.
- Features are extracted from Relu 6, Relu 7, and Fully Connected layers 6 and 7.
- Linear Kernels are used for the Fully Connected layers and χ^2 kernels for the Relu layers.

Method 2: Fisher Vectors



- Dense SIFT and Root SIFT features were extracted using vl_phow. [5]
- The Fisher Vectors were ℓ_2 normalized before and after encoding.
- The number of descriptors per image was set to 6000.
- We used Gaussian Mixture Models to generate the codebooks with 256 clusters.
- In a further approach, the Fisher Vectors were fused with CNN features before training a linear SVM.

Results



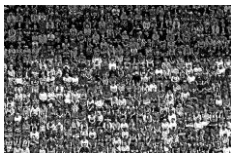
Minimum AD: 1
Ground Truth: 96 Estimated: 97



Maximum AD: 2042
Ground Truth: 4633 Estimated: 2591



Minimum NAD: 0.2 %
Ground Truth: 1601 Estimated: 1604



Maximum NAD: 79%
Ground Truth: 440 Estimated: 786

These are the results for features from CNN + SIFT + Fisher Vectors.

CVPR'13 Method	Absolute Difference	Normalized Difference
SIFT + Textual Features	324.42	30.91
SIFT Features	366.22	32.54

Method 1: CNN Features

	ImageNET-VeryDeep-19		AlexNET	
Layer	AD	NAD	AD	NAD
Relu 6	291.85	28.57	324.41	29.37
Relu 7	317.14	29.92	328.09	31.04
Convfc 6	308.49	25.49	336.55	29.75
Convfc7	326.86	27.45	357.53	30.00
Relu 6 + SIFT	269.55	29.93	295.48	27.14
Relu 7 + SIFT	291.64	31.82	303.25	27.93
Convfc 6 + SIFT	281.48	23.86	361.05	31.22
Convfc 7 + SIFT	286.54	24.59	362.26	31.20

Method 2: Fisher Vectors

Type	Absolute Difference	Normalized Difference
Dense SIFT	337.7	28.93
Root SIFT	337.25	29.52
Dense SIFT + CNN	286.60	25.46
Root SIFT + CNN	292.76	26.25

Combined Results

Type	Absolute Difference	Normalized Difference
CNN + FV SIFT + SIFT	273.89	23.62

References:

[1] Idrees, Haroon, et al. "Multi-source multi-scale counting in extremely dense crowd images." Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.
[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[3] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.
[4] Vedaldi, Andrea, and Karel Lenc. "MatConvNet-convolutional neural networks for MATLAB." arXiv preprint arXiv:1412.4564 (2014).
[5] Vedaldi, Andrea, and Brian Fulkerson. "VLFeat: An open and portable library of computer vision algorithms." *Proceedings of the international conference on Multimedia*. ACM, 2010.