Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images

Haroon Idrees[1] Imran Saleemi[1] Cody Seibert[2] Mubarak Shah[1]

[1]Center for Research in Computer Vision
University of Central Florida
{haroon, imran, shah}@eecs.ucf.edu

[2]Department of EECS
University of Central Florida
seibert_cody@knights.ucf.edu

Abstract

We propose to leverage multiple sources of information to compute an estimate of the number of individuals present in an extremely dense crowd visible in a single image. Due to problems including perspective, occlusion, clutter, and few pixels per person, counting by human detection in such images is almost impossible. Instead, our approach relies on multiple sources such as low confidence head detections, repetition of texture elements (using SIFT), and frequency-domain analysis to estimate counts, along with confidence associated with observing individuals, in an image region. Secondly, we employ a global consistency constraint on counts using Markov Random Field. This caters for disparity in counts in local neighborhoods and across scales. We tested our approach on a new dataset of fifty crowd images containing 64K annotated humans, with the head counts ranging from 94 to 4543. This is in stark contrast to datasets used for existing methods which contain not more than tens of individuals. We experimentally demonstrate the efficacy and reliability of the proposed approach by quantifying the counting performance.

1. Introduction

The problem of counting the number of objects, specifically people, in images and videos arises in several real-world applications including crowd management, design and analysis of buildings and spaces, and safety and security. In certain scenarios, obtaining the people count is of direct importance, e.g., in public rallies, marathons, public parks, and transportation hubs, etc. The manual counting of individuals in very dense crowds is an extremely laborious task, but is performed nonetheless by experienced personnel when needed [18].

Computer vision research in the area of crowd analysis has resulted in several automated and semi-automated solutions for density estimation and counting. Practical application of most existing techniques however, is constrained by two important limitations: (1) inability to handle crowds of *hundreds or thousands* (Fig. 1) rather than a few tens of individuals [4, 5]; and (2) reliance on temporal constraints in crowd videos [20], which are not applicable to the more prevalent *still images*.

Figure 1: This figure shows five arbitrary images from the dataset used in this paper. On average, each image in the crowd counting dataset contains around 1280 humans. The bottom row shows four patches from different images at original resolution.

Most existing methods can be categorized by the application scenario and experimental setup. Some methods proposed in literature for crowd detection perform image segmentation without actual counting or localization [1], while others simply estimate the coarse density range within local regions [24]. In terms of experimental data, most of the existing algorithms for exact counting have been tested on low

to medium density crowds, e.g., USCD dataset with density of $11-46$ people per frame [4], Mall dataset with density of $13-53$ individuals per frame [5], and PETS dataset containing $3-40$ people per frame [9]. In contrast to these images and videos, our algorithm has been tested on still images containing between 94 and 4543 people per image, with an average of 1280 people over fifty images in the dataset. Such high density implies that an individual may occupy so few pixels that it can neither be detected, nor can its presence be verified given the location, which are key requirements in existing techniques.

The proposed approach is motivated by the fact that in extremely dense crowds of people, no single feature or detection method is reliable enough to provide an accurate count due to low resolution, severe occlusion, foreshortening, and perspective. Indeed even the state-of-the-art human, head, or face detectors perform poorly in such scenarios. We observe however that densely packed crowds of individuals can be treated as a texture, albeit irregular and inhomogeneous at a coarse scale. And this texture begins to correspond to a harmonic pattern, as is the case in regular textures, at a finer scale. Furthermore, there does exist a spatial relationship that is expected to constrain the counting estimates in neighboring local image regions in terms of similarity of counts.

We also observe that, in derived intensity spaces such as image derivative, or edges, groups of individuals are likely to exhibit an increased level of similarity. Therefore, in addition to supervised training of human or head detectors, appearance based feature descriptors like SIFT are also useful to estimate the so called texture elements or textons [25]. This observation has been used successfully for crowd detection in [1], although not for counting or localization. Our goal in using appearance based descriptors for localized patches is to estimate repeating structures in the image, but with the important distinction that such image patches are not expected to fully contain a person, rather the textons can represent a single part of a person, multiple parts, or multiple people and their parts.

Another main contribution of the proposed framework is the use of frequency-domain analysis in crowd counting. Fourier transform has been used extensively in texture analysis [2], and specifically in crowd analysis [17]. Given geometrically arranged texture elements, the Fourier transform can provide reliable estimates of the texton counts [14]. In the domain of crowd counting however, the application of frequency analysis is severely limited due to two main reasons: (1) the spatial arrangement of texture elements is very irregular; and (2) the Fourier transform is not useful in localizing the repeating elements.

We propose novel solutions to overcome these limitations. First, we employ Fourier analysis along with head detections and interest-point based counts in local neighborhoods on multiple scales to avoid the problem of irregularity in the perceived textures emanating from images of dense crowds. The count estimates from this localized multi-scale analysis are then aggregated subject to global consistency constraints. Secondly, in order to leverage multiple estimates from distinct sources, the corresponding confidence maps need to be comparable and in the same space. For instance, the Fourier transform is not directly useful in this regard since it cannot be combined with count estimate maps in the image domain. We therefore reconstruct the low to medium frequency component of image region and the reconstructed image is then compared with the original image after alignment. This process provides two important pieces of information: the estimated count per local region, and a measure of error relative to the original image.

Combining the three sources, i.e., Fourier, interest points and head Detection, with their respective confidences, we compute counts at localized patches independently, which are then globally constrained to get an estimate of count for the entire image. Since the data terms are evaluated independently at different scales, the smoothness constraint has to be applicable to spatial neighborhoods as well as immediate neighbors at different scales. We propose a solution to obtain counts from multi-scale grid MRF which infers the solution simultaneously at all scales while enforcing the count consistency constraint.

The organization of the rest of the paper follows. We review relevant literature in §2, present detailed problem formulation and solution in §3, and finally, the experimental evaluation is reported in §4. The paper is concluded in §5.

2. Related Work

Some of the existing literature relevant to the proposed approach and application is briefly reviewed in this section. Person detection for counting individuals, present in an image or video, has been employed in [10, 15]. This category of methods however is not useful for the kind of images we deal with, because human, or even head and face detection in these images is difficult due to severe occlusion and clutter, low resolution, and few pixels per individuals due to foreshortening. We demonstrate this fact by reporting quantitative results of detection on our crowd image dataset.

Brostow and Cipolla [3] and Rabaud and Belongie [19] count moving objects by estimating contiguous regions of coherent motion. Computation of such patterns of motion were also proposed in [22, 23, 12], but not with explicit application to the problem of crowd counting. These algorithms require video frames as input, with reasonably high frame rate for reliable motion estimation, but are not suitable to still images of crowds, or even videos if the individuals in the crowd show nominal or no motion, e.g., political gatherings and concerts.

Another category of techniques proposed for crowd

counting rely on estimation of direct relationships between low level or local features and counts, by learning regression functions. Such a function can be global [4, 6, 11, 21] where a single function's parameters are learned for the entire image or video. These methods have the implicit assumption that the density is roughly uniform regardless of the location where the feature is computed. This assumption is largely invalid in most real world scenarios due to perspective, changes in viewpoint, and changes in crowd density.

The problems associated with global feature regression can be alleviated by relaxing this assumption. Methods such as [16] propose to divide an image into cells and perform regression individually for each cell. These methods [16, 13] aim to compensate for problems associated with foreshortening, and local geometric distortions due to perspective. One key problem with this approach however is that the local context, or spatial consistency constraints are ignored as information across local regions is not shared.

Chen et al [5] have recently proposed that information sharing among regions should allow more accurate and robust crowd counting. They propose a single multi-output model for joint localized crowd counting based on ridge regression. Their proposed framework employs inter-dependent local features from local spatial regions as input and people count from individual regions as multi-dimensional structured output. The proposed algorithm however was not applied to scenarios with crowds of more than a few tens of people.

We now describe our proposed approach in detail which puts forth several novel ideas to overcome limitations in existing work. We also collected, annotated, and tested on a large dataset of real world crowd images.

3. Framework

Given an image, our goal is to estimate the number of people in the image. The density of people, i.e., the number of people per unit area, in an arbitrary crowded image is rarely uniform, and varies from region to region. This variation in density may be inherent to the scene that the image captures (different distribution of individuals in different parts of the scene) or it may arise due to the viewpoint and perspective effects of the camera. Therefore, a crowded scene cannot be analyzed in its entirety for counting. Thus, the proposed framework begins by counting individuals in small patches uniformly sampled over the image. But, even though the density varies across the image, it does so smoothly, suggesting the density in adjacent patches should be similar.

We handle the issues of variation in density and smooth variation separately. When counting people in patches, we assume the density is uniform but implicitly assume that the number of people in each patch is independent of adjacent patches. Once we estimate density or counts in each patch, we remove the independence assumption and place them in multi-scale Markov Random Field to model the dependence in counts among nearby patches.

Figure 2: Results of Head Detection: Image on the left is one of the few images where head detection gives reasonable results. False negatives and positives are still evident in both images.

3.1. Counting in Patches

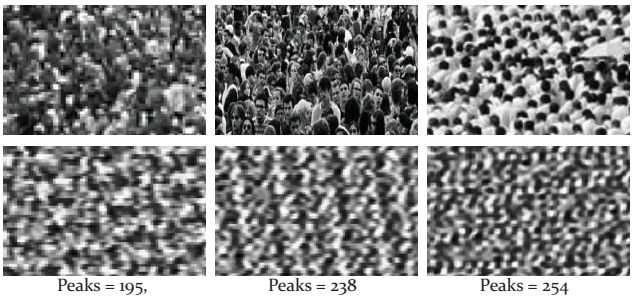
Given a patch P, we estimate the counts from three different and complementary sources, alongside confidences for those counts. The three sources are later combined to obtain a single estimate of count for that patch using the individual counts and confidences.

3.1.1 HOG based Head Detections

The simplest approach to estimate counts is through human detections. However, a quick glance at images of dense crowds reveals that the bodies are almost entirely occluded, leaving only heads for counting and analysis. We, therefore, used Deformable Parts Model [7] trained on INRIA Person dataset, and applied only the filter corresponding to head to the images. Often, the heads are partially occluded, so we used a much lower threshold for detection. There are many false negatives and positives since the images are inherently difficult (see Fig. 2). The detections are accompanied with scale and confidence. For each patch, we use number of detections, η_H, mean and variance of scale $\mu_{H,s}$, $\sigma_{H,s}$ and confidence $\mu_{H,c}$, $\sigma_{H,c}$. The consistency in scale and confidence is a measure of how reliable head detections are in that patch.

3.1.2 Fourier Analysis

When a crowd image contains thousands of individuals, with each individual occupying only tens of pixels, especially those far away from the camera in an image with perspective distortion, histograms of gradients do not impart any useful information. However, a crowd is inherently repetitive in nature, since all humans appear the same from a distance. The repetitions, as long as they occur con-



Peaks = 195, GT Count = 54 | Peaks = 238 GT Count = 102 | Peaks = 254 GT Count = 134

Figure 3: Counting through Fourier Analysis: The first row shows three original patches, while the second row shows corresponding reconstructed patches. The positive correlation is evident from the number of local maximas in the reconstructed patch, and the ground truth counts shown at the bottom.

sistently in space, i.e., crowd density in the patch is uniform, can be captured by Fourier Transform, $f(\xi)$, where the periodic occurrence of heads shows as peaks in the frequency domain. Specifically, for a given patch, we compute the gradient image, $\nabla(P)$, and apply a low-pass filter, $f(\xi) > f(\xi_o) = 0$, to remove very high frequency content. Next we discard low amplitude frequencies, which is followed by reconstruction, P_r, through inverse Fourier Transform. We find the number of local maximas in the reconstructed image (Fig. 3) after alignment and non-maximal suppression which serves as an estimate for the Fourier-based count, η_F. In addition, we compute several other measures, such as entropy as well as statistical measures related to first four moments - mean, variance, skewness and kurtosis for both the reconstructed image and difference image $|P_r - \nabla(P)|$. The count is normalized for the size of the patch.

3.1.3 Interest Points based Counting

We use interest points not only to estimate counts but also to get a confidence whether the patch represents crowd or not. Since sky, buildings and trees naturally occur in outdoor images, and the fact that head detection gives false positives in such regions (Fig. 2) and Fourier Analysis is crowd-blind, it is important to discard counts from such patches. For both counting and confidence, we obtain SIFT features, and cluster them into a codebook of size c. In order to obtain counts or densities using sparse SIFT features, we use Support Vector Regression using the counts computed at each patch from ground truth.

From the perspective of Statistics, the number of individuals in a particular patch can be seen as spatial Poisson Counting Process with parameter (corresponds to density), λ, i.e., $N(P) \sim \texttt{Poisson}(\lambda|P|)$, and expected value of

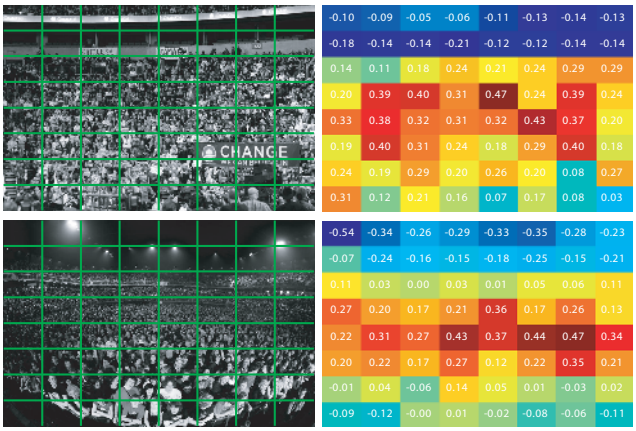


Figure 4: Images with their confidence maps: The images on the left have confidence of crowd likelihood obtained through Eq. 2. In the top image, the gap between stadium tiers gets low confidence of crowd presence. Similarly, patches containing the sky and flood lights in bottom image have low probability of crowd.

$N(P)$ is simply $\lambda|P|$. Since we assumed the density is uniform in the patch, the process is homogenous and λ is not a function of location (x, y). Moreover, the independence assumption among patches gives, for the image, I:

$$\begin{aligned} N(I) &= N(P_1 \cup P_2 \ldots P_n) \\ &= N(P_1) + N(P_2) + \ldots + N(P_n), \end{aligned} \tag{1}$$

where $P_1, P_2, \ldots P_n$ form a disjoint partition of I.

Furthermore, due to sparse nature of SIFT features, the frequency γ of a particular feature i in a patch can also be modeled as a Poisson R.V., $p(\gamma_i|crowd) = exp(-\lambda_i^+).(\lambda_i^+)^{\gamma_i}/\gamma_i!$ with expected value, λ_i^+. Given a set of positive$(+)$ and negative examples$(-)$, the relative densities (frequencies normalized by area) of the feature vary in positive and negative images, and can be used to identify crowd patches from non-crowd ones. Assuming independence among features, the log-likelihood $\varphi(P)$ of the ratio of patch containing crowd to non-crowd is [1]:

$$\begin{aligned} &log(\gamma_1, \gamma_2, \ldots \gamma_c|crowd) - log(\gamma_1, \gamma_2, \ldots \gamma_c|\neg crowd) \\ &\qquad = \sum_i^c \left(\lambda_i^- - \lambda_i^+ + \gamma_i(log\lambda_i^+ - log\lambda_i^-)\right). \end{aligned} \tag{2}$$

The above equation gives us a confidence for presence of crowd in a patch. The resulting confidence maps are shown in Fig. 4 for two images.

3.2. Fusion of Three Sources

For learning and fusion at the patch level, we densely sample overlapping patches from the training images and

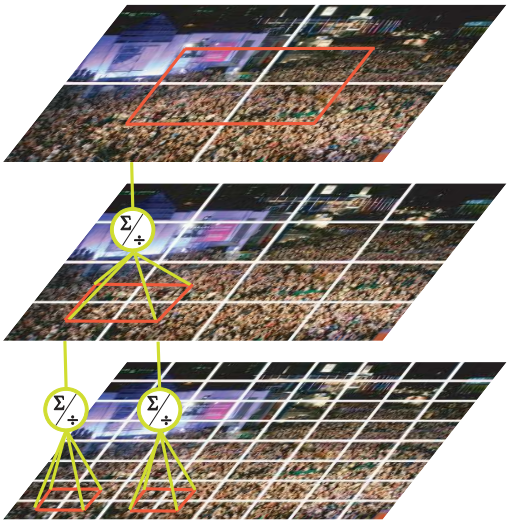


Figure 5: The figure shown multi-scale Markov random Field for inferring counts for the entire image. The patches in each layer have independent data terms, thus requiring a simultaneous solution for all layers.

using the annotation, obtain counts for the corresponding patches. Computing counts and confidences from the three sources, we scale individual features and regress using ϵ-SVR, with the counts computed from the annotations.

3.3. Counting in Images

In order to impose smoothness among counts from different patches, we place them in an MRF framework with grid structure. Furthermore, although small patches have consistent density, they have fewer repetitions or periods and can easily be affected by low-frequency noise. Larger patches, if they have consistent density, have more people, and therefore more periods and better relevant-to-irrelevant frequency ratio. Moreover, it is difficult to ascertain in advance the right scale for analysis for a particular image. This problem lends itself to a multi-scale MRF, an example of which is shown in Fig. 5. The graph can be represented with $(\mathcal{V}, \mathcal{E})$ and $\mathcal{N}$ are the four neighbors at the same level and intermediate nodes that connect a patch to layers above and below it. Note that, this multi-scale MRF is different from other hierarchical models used for images, in that the data term (unary cost) for a patch is evaluated independent of the patches at layers above and below it, whereas in image restoration and stereo, data cost for patch at higher level is computed from layer directly below. The energy function is thus given by:

$$E(\ell) = \sum_{p \in \mathcal{V}} D_p(\ell_p) + \sum_{(p,q) \in \mathcal{N}} V(\ell_p - \ell_q), \tag{3}$$

where labeling ℓ assigns a label $\ell_p \in \mathcal{L} = \{0, 1, 2, ..., C_{max}\}$ for every every patch $p \in P$. The data term is quadratic, $D_p(\ell_p) = \lambda(\eta_p - \ell_p)^2$ and smoothness term is truncated quadratic, $V(\ell_p - \ell_q) = \min\left((\ell_p - \ell_q)^2, \tau\right)$.

The graph is inferred using Max-Product/Min-Sum BP on grid structure [8]. At any time t, the message that node p sends to q for a label ℓ_q is given by, $m^t_{p \to q}(\ell_q)$:

$$\min_{\ell_p} \left(V(\ell_p - \ell_q) + D_p(\ell_p) + \sum_{s \in \mathcal{N}_p \setminus q} m^{t-1}_{s \to p}(\ell_p) \right), \tag{4}$$

and the belief for a label ℓ_q of node q at time t can be obtained as:

$$b^t_q(\ell_q) = D_q(\ell_q) + \sum_{p \in \mathcal{N}_q} m^t_{p \to q}(\ell_q). \tag{5}$$

The inference starts by sweeping in four directions at the bottom level using Eq. 4, the beliefs are then evaluated for each patch using Eq. 5. Then, the beliefs in the groups of 2×2 are added giving the beliefs for the intermediate nodes b^t_i above the bottom layer. After four sweeps at the middle layer, the fifth sweep of messages goes from intermediate nodes to the middle layer. This is followed by computation of beliefs at the middle layer. This step repeats for the top layer, and the whole process corresponds to one time step t. Then, the process repeats but from top to bottom. The beliefs at the intermediate nodes are divided for each of the patch below, i.e., for each patch q in 2×2 group below the intermediate node, its share of beliefs from the layer above is given by: $b^{t+1}_{i,q}(\ell_q) = b^t_q(\ell_q).b^{t+1}_i(\ell_q)/b^t_i(\ell_q)$. After a fixed number of iterations, the final beliefs can be computed using Eq. 5, and the labels which have minimum cost in the belief vectors are selected as the final labels. The sum of labels (counts) at the bottom layer gives the count for the image.

Fig. 6 shows three instances where the estimated count of patch was improved based on neighbors (both spatial and layer). In all cases, the patch under consideration lies in the center of 3×3 patch set. In the first two columns, after imposing the smoothness constraint using MRF, the overestimated counts are reduced, becoming closer to ground truth. A special case is shown in the last column. The patch in the middle had a much lower count than neighbors which after inference increased becoming similar to its neighbors. Although the new estimate is closer to ground truth, the increase is not necessarily correct since the lower count was due to presence of a non-human object (an ambulance). The last column belongs to the image which had the highest count in the dataset.

4. Experiments

We collected the dataset from publicly available web images, including Flickr. As mentioned in the introduction, it

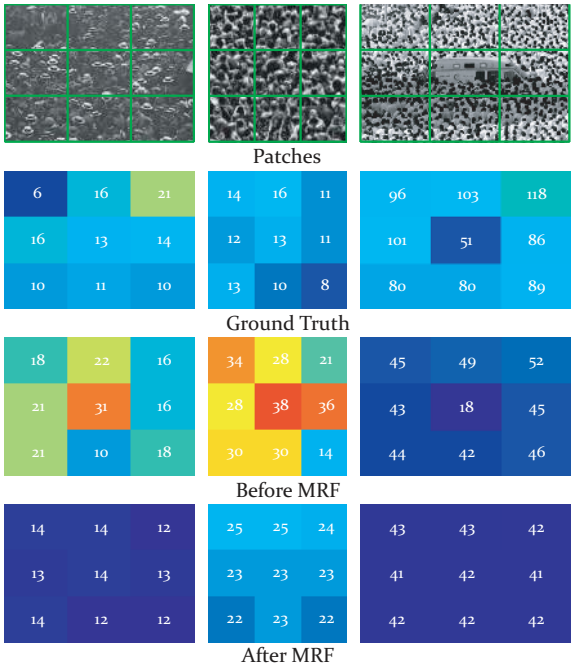
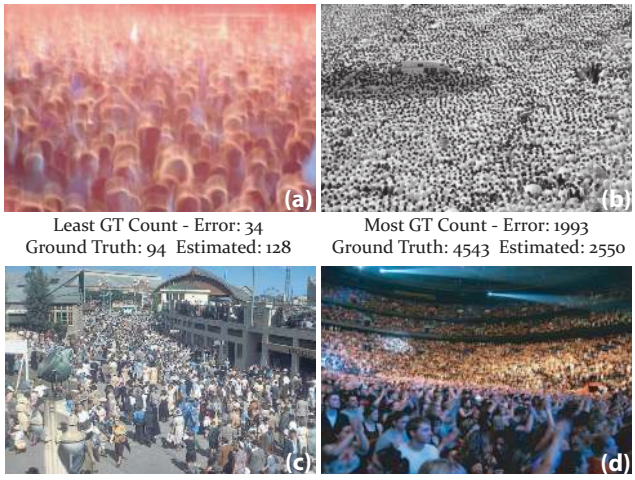


Figure 6: Results after MRF-based inference: Three nonets from different images are shown in first row. The second row shows the ground truth counts, and the estimated counts before and after MRF inference are shown in third and fourth rows, respectively. The patches from only one layer are shown in this figure.

consists of 50 images with counts ranging between 94 and 4543 with an average of 1280 individuals per image. Much like the range of counts, the scenes in these images also belong to a diverse set of events: concerts, protests, stadiums, marathons, and pilgrimages. One of the images is a painting while another is an abstract depiction of a crowd (the one with the least count, shown in Fig. 7a). Using a simple tool for marking the ground truth positions of individuals, we obtained 63705 annotations in the fifty images. Some examples of images with the associated ground truth counts can be seen in Fig. 7.

For experiments, we randomly divided the dataset into sets of 10, reduced the maximum dimension to 1024 for computational efficiency, and performed 5−fold cross-validation. We used two simple measures to quantify the results: mean and deviation of Absolute Difference (AD), and mean and deviation of Normalized Absolute Difference (NAD), which was obtained by normalizing the absolute difference with the actual count for each image. Since we divide the image into patches, we report our results for both patches and images. The quantitative results are presented in Table 1.

The first row in Table 1 shows the results of using counts from Fourier Analysis only, giving AD of 703.9 and NAD



Least GT Count - Error: 34
Ground Truth: 94 Estimated: 128

Most GT Count - Error: 1993
Ground Truth: 4543 Estimated: 2550

Minimum Error - Error: 2
Ground Truth: 426 Estimated: 428

Maximum Error - Error: 2046
Ground Truth: 3333 Estimated: 1287

Figure 7: Selected images with their respective counts and errors: The first row shows the extreme ends of the dataset in terms of counts. The second row shows the images with lowest and highest error.

of 84.6. Supplementing it with confidences from various sources including Eq. 2 improves AD by 181.8 and reduces NAD by almost one-half. Including counts from head detections improves AD marginally to 510.9. Adding counts from regression on sparse SIFT features reduces error in both measures, giving values of 468.0 and 32.2, respectively. Finally, inferring counts for complete images using counts from patches through multi-scale MRF further improves AD taking it to 419.5. It can be observed from the table, that standard deviation follows the same trend as mean, the values reducing as we add more sources.

Figs. 8a-b shows AD and NAD for patches in the individual images, respectively. The mean per patch are shown with black asterisks, deviations with red bars, and olive dots in Fig. 8a show average of actual counts per patch in that image. For easier analysis, the x-axis shows images sorted with respect to actual counts in both plots. It can be seen that AD per patch increases as the actual counts increases, except for the images in the range 25 to 45 with corresponding actual counts in the range of $1000-2500$ per image. Not only does this range boast lowest mean in AD and NAD, but lowest deviations as well, which means the approach consistently predict correct counts for patches in this range. The reason for better performance in the middle range is obvious: the counts range from $94-4543$, so the largest count is a tremendous 4832% of the smallest count. Forcing the learning algorithm to predict correct estimates at both ends simultaneously, makes it overestimate the lower end and underestimate the higher end, thereby working in favor of the middle range, even though, we used RBF kernel for regres-

Method	Error			
	Per Patch		Per Image	
	AD	NAD	AD	NAD
Fourier	13.8 ± 21.3	96.4 ± 200.4	703.9 ± 682.0	84.6 ± 157.3
F+**c**onfidence	11.0 ± 19.7	58.7 ± 74.9	522.1 ± 610.1	41.0 ± 31.0
Fc+**H**ead	11.1 ± 19.3	63.3.0 ± 84.0	510.9 ± 587.3	41.8 ± 30.9
FHc+**S**IFT	10.2 ± 18.9	53.3.0 ± 69.5	468.0 ± 590.3	32.2 ± 27.1
FHSc+**M**RF **(Proposed)**	-	-	**419.5 ± 541.6**	**31.3 ± 27.1**
Rodriguez et al.	-	-	655.7 ± 697.8	70.6 ± 102.1
Lempitsky et al.	-	-	493.4 ± 487.1	61.2 ± 91.6

Table 1: Quantitative results of the proposed approach and comparison with Rodriguez et al. [20] and Lempitsky and Zisserman [13] using mean and standard deviation of Absolute Difference and Normalized Absolute Difference from ground truth. The influence of the individual sources is also quantified. The proposed approach outperforms the other two methods.

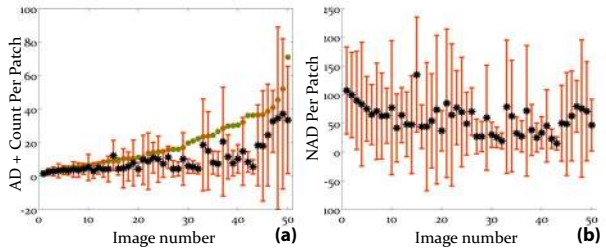


Figure 8: This figure shows analysis of patch estimates in terms of absolute and normalized absolute differences. The x-axis shows image number sorted with respect to actual count. Means are shown in black asterisk, standard deviations with red bars, and ground truth counts with olive dots.

sion on three sources.

For comparison, we used the methods of Rodriguez et al. [20], and Lempitsky and Zisserman [13], which were suitable for this dataset since other methods for crowd counting mostly deal with videos or use human detection, and cannot be used for testing on this dataset. The method presented in [20] relies on head detections, while [13] requires annotated ground truth points for training, and learns a regression model using dense SIFT features on randomly selected patches. The quantitative results are shown in Table 1. Fig. 9 breaks these numbers according to counts. The results using [20] are in red, those in green use [13], and the results of the proposed approach are shown blue. In Fig. 9b, the black curve represents the ground truth. In Fig. 9a, we show NAD for ten groups of five images each, which are sorted according to ground truth counts. The x-axis shows the average counts of each of the 10 groups. Density aware person detection [20] performs best around counts of 1000, but its error increases as we move away. The reason becomes obvious when we look at the absolute counts output by the method in Fig. 9b, as they are fairly steady across the entire dataset and do not respond well to change in density. It overestimates at lower end and then underestimates at the higher end, resulting in increased absolute errors on both ends. The MESA-distance [13] on the other hand, performs fairly well at higher counts, but gives high NAD at lower counts. The reason lies in the algorithm itself, as it is designed to minimize the maximum AD across images when training, and since images with higher counts tend to have higher AD, the learning focuses on such images. The learner gets biased towards high density images, thus, producing a lower AD overall, but overestimating at lower counts (Fig. 9b), thus giving higher NAD. The proposed approach, on the other hand, performs well across the whole range, giving steady NAD's across all ten groups.

Finally, all methods underestimate the tenth set and this can be due to several reasons. First, images in this group are very high resolution and therefore it is less likely to miss individuals while annotating. Since we fixed the maximum image size for experiments, the images in this group had correct and therefore, more annotations than their low-resolution counterparts. Second, a careful look at Fig. 8a reveals that patch density increases super-linearly for this group, which otherwise is linear for first nine groups. Since there are few such images, their patch instances could have been treated as outliers (have higher slack weights) for regression. The last reason may be associated with histograms of features that capture relative frequencies. At very high density, the relative frequencies across patches with different density may become similar, resulting in a loss of discriminative power.

5. Conclusion

We presented an approach to count number of individuals in extremely dense crowds, on a scale not tackled before. We fuse information from three sources in terms of counts, confidences and different measures at the patch level, and then enforce smoothness constraint on nearby patches to improve estimates of incorrect patches, thereby

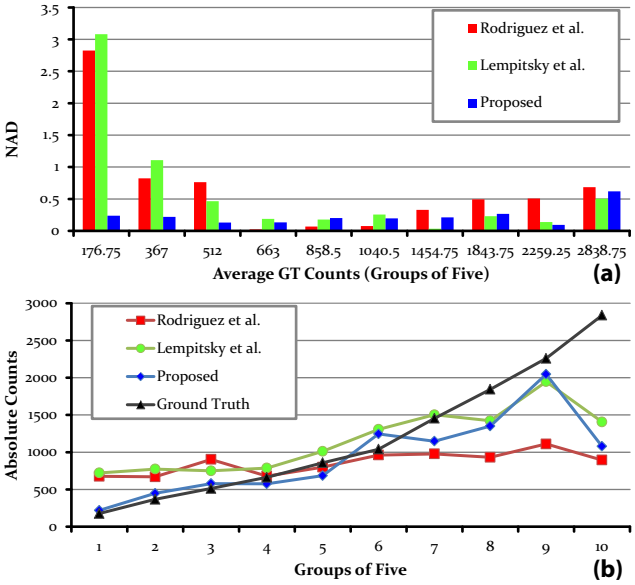


Figure 9: Analysis of comparison: Bars and lines in red depict [20], green show [13], blue shows the results using proposed approach, while ground truth is shown in black. (a) shows Normalized Absolute Difference (an error measure) and (b) shows the actual and estimated counts.

producing better estimates at the image level. We showed that the proposed approach scales well to different densities producing consistent error rates across images with diverse counts. Possible improvements include explicit preprocessed estimation of crowd density, and making regression an explicit function of density so that it better adapts to various crowd sizes. Furthermore, texton detection to recognize repetitions can supplement frequency-domain analysis.

Acknowledgments. This material is based upon work supported in part by, the U.S. Army Research Laboratory, the U.S. Army Research Office under contract/grant number W911NF-09-1-0255. Cody Seibert was supported on National Science Foundation's REU Program.

References

[1] O. Arandjelovic. Crowd detection from still images. In *BMVC*, 2008.

[2] R. Azencott, J.-P. Wang, and L. Younes. Texture classification using windowed fourier filters. *PAMI*, 19(2):148–153, 1997.

[3] G. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[4] A. Chan, Z. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[5] K. Chen, C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012.

[6] S. Cho, T. Chow, and C. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 29(4):535–541, 1999.

[7] P. Felzenszwalb, D. McAllester, and D. Ramaman. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *Int. J. Comput. Vision*, 70(1):41–54, Oct. 2006.

[9] J. Ferryman and A. Ellis. Pets2010: Dataset and challenge. In *AVSS*, 2010.

[10] W. Ge and R. Collins. Marked point processes for crowd counting. In *CVPR*, 2009.

[11] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*, 2005.

[12] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009.

[13] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*, 2010.

[14] T. Leung and J. Malik. Recognizing surface using three-dimensional textons. In *ICCV*, 1999.

[15] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *ICPR*, 2008.

[16] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In *ICCIT*, 2010.

[17] A. Marana, S. Velastin, L. Costa, and R. Lotufo. Automatic estimation of crowd density using texture. In *IWSIP*, 1997.

[18] R. Melina. How is crowd size estimated? In *Life'sLittleMysteries.com*, 2010.

[19] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, 2006.

[20] M. Rodriguez, J. Sivic, I. Laptev, and J. Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011.

[21] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications*, 2009.

[22] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *CVPR*, 2007.

[23] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *IJCV*, 67(1):21–51, 2006.

[24] B. Zhou, F. Zhang, and L. Peng. Higher-order svd analysis for crowd density estimation. *CVIU*, 116(9):1014–1021, 2012.

[25] S. Zhu, C. Guo, Y. Wu, and Y. Wang. What are textons? *IJCV*, pages 121–143, 2002.