A Comparative Study for Object Detection and Estimation using Multi-Task Convolutional Neural Network

Pankaj Sharma[1], Harish Sharma[2], Sunita Singhal[3]
[1]*Student, Manipal University, Jaipur, India*
[2]*Assistant Professor, Manipal University, Jaipur, India*
[3]*Professor, Manipal University, Jaipur, India*

***Abstract*: In object detection, the detection is done by finding similar object is same as your specimen object. In our case we are performing this using image as digital images. Yes, we are using the image processing which belongs to computer vision field. Apart from object (in our case crowd) detection other ways are clustering and regression based for fulfilling the desired output efficiently. Regression based models and detection-based models are performing very well nowadays in crowd counting using Deep Convolutional Neural Network. Due to different scaled, occlusion, clutter, low resolution, sensor noise, perspective, illumination, similarity, deformity, color are the factors which affect mostly our crowd counting accuracy. The target representation were mainly doing by blob, rectangle, ellipse, dot, coloured area. In this we will provide a comparative study for multi-task CNN in image processing. We will see how much Neural network have changed our way of getting the desired result from images or videos.**

***Keywords*: Multi-Task Convolutional Neural Network**

1. Introduction

Some applications for object (crowd) counter,

- Mall, Marts for optimizing the things location, managing the high crowd in peak times.
- Safety of people in public areas, political rallies, sports events, concerts.
- Traffic controlling for determining driver alcohol influence or drowsiness detection for preventing accidents.

In dense crowd it is hard to detect the desired object. And there are other factors that also affect its detection like occlusion, clutter of two objects, perspective, noise in image. But this crowd counting topic also gained its interest among researchers to develop as good as they can. Analyzing crowd have various aspects like estimating the crowd, [12] tracking and monitoring, determining crowd behaviour. If the image have less dense crowd then human can do that count by his bare eyes but if dense crowd is there and image have some imperfections, then it is hard to count. Therefore, we needed image processing which is part of computer vision. Convolutional Neural Network (CNN) played comparatively very well in many other field applications such as Natural Language Processing (NLP) [10], chat bots etc. So, in crowd counting or density estimation have also influenced by Convolutional Neural Network. Crowd counting mainly practiced by crowd detection, crowd clustering and regression also. Scalability also plays crucial role in obtaining the better accuracy for crowd counting, which is considerably enlighten by some models like multi-column and resolution models. Here in this study we will discuss past few multi-tasks Convolutional Neural Network applied models for counting objects to the state of art model currently. Multi-tasks Convolutional Neural Network models are able to multiple tasks simultaneously like determining local, global or coarse count from the images and then from those multiple features on an image we will try to fuse them together or add them up to get better accurate count of objects.

2. Literature Survey

For object detection a model first needs a specimen and object representation ways. Objects are represented by a contour, template, blob, skeleton or some geometric shapes. These representation helps model to locate or track the desired object easily. [13], [11] Geometric shapes representation is mostly followed by the researchers in their respective models for those desired objects which can look like some geometric shape. But, for human like objects which are of typical shape, they prefer skeleton, point, blobs and contour(area) like representation.

Convolutional Neural Networks uses convolution mathematical formula,

$$(f * g)(t) = \int_{-\infty}^{\infty} f(x) \cdot g(t-x)\,dx$$

$$(f * g)(t) = \sum_{x=-\infty}^{\infty} f(x) \cdot g(t-x)$$

Here we take f as intensity for the observing pixel and g is the weight function or kernel. This kernel is our feature map of desired object detection in the image. This kernel is set of matrices of intensity of each pixel in the that. For every layer of CNN model we have image as in greyscale or RGB images. For making n number layer of convolution neural network we use this formula

$$(I * \omega)(n, x, y) = \sum_{C=0}^{c} \sum_{p=-\frac{k}{2}}^{k/2} \sum_{q=-\frac{k}{2}}^{k/2} I(c, x+p, y+q)^{*} w(c, n, p, q)$$

Here I is the pixel at location x and y in particular channel. And k is size of kernel.

The current state of art CNN model for object detection or crowd count is cascaded Multi-Task with high prior learning is proposed by Sindagi et al. [1]. They have been able to count the crowd with three different techniques which are high prior stage and density estimation stage. The high prior stage will give coarse count of crowd from the provided image and the density estimation stage comprised of local and global count on that particular frame. Then those results are fused together to get the accurate count. This model by them proposed in year 2017. They have performed shanghai dataset (part_A and part_B) and UFC_CC_50 dataset. Their model has performed exceptionally good on those datasets. Graph below show you the MAE (Mean Absolute Error) and MSE (Mean squared Error) on those datasets.

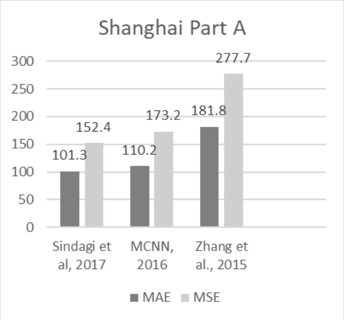

Fig. 1. Comparison of MAE and MSE for Shanghai_Part_A dataset

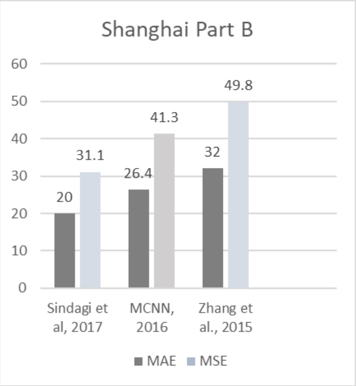

Fig. 2. Comparison of MAE and MSE for Shanghai_Part_B dataset

As you see Sindagi et. al. model outperforms previous models very well. Those both errors MAE and MSE are less than previous state of art methods such as Zhang et al [2] and MCNN by Wang et al. [3] on same dataset. Same numbers difference is also shown in Shanghai Part B dataset also.

Here in shanghai part B dataset, Sindagi et al. [1] also have low error values than other two crowd count CNN-based model.

Now we will see on UFC_CC_50 dataset of images.

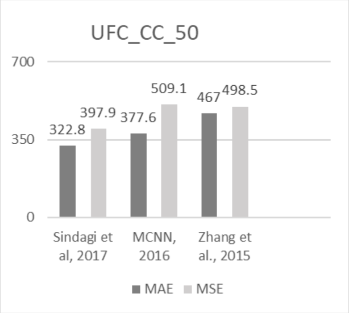

Fig. 3. Comparison of MAE and MSE for UFC_CC_50 dataset

Their MAE and MSE both are low than the previous other models on UFC_CC_50 dataset.

These both error calculation formula is,

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |Y_i - Y_i'|$$

$$\text{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'|^2}$$

Apart from density estimation multi task CNN model were implemented by [4] for estimating the food dish and calories in that.

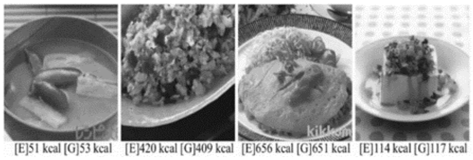

Fig. 4. Examples on successful estimation of food calories

[E] and [G] represents the estimated calorie values and the ground-truth values, respectively. And in [5] they developed multi-task CNN model for action recognition of people like clapping, smile, walking, hand waving etc. But multi task C3D +LSTM their accuracy rate was less than their model like TSN [6].

In [7], they have proposed a light weight (means small data size) multi-task CNN model for estimation of age and gender of people for mobile applications. They have used convolution depth wise so that its size gets optimized. Their LMTCNN models are 8 to 30 Mb only.

Table 1
Size in mb(Megabytes) of [7] variants

Model	Size (in mb)
LMTCNN-1-1	8 mb
LMTCNN-2-1	30 mb

But, they have still programming limitations for images. In [8] They have improved the object detection multi-task CNN model by doing semantic segmentation for better structural information, overlapping in 2-D the loss function for better local count and making hierarchical feature maps.

Table 2
Improvement in percentage of model

Feature	Result
overlap loss	By 0.9% (in mAP)than previous .
System	By 4.8% than basic RCNN model

In [6], they have proposed a model which represent an object by collection bag, a single object suppose watch, its images bag collection is being made means at a time there will be so many samples of watches that exactly means multiple instances. And by adding multi task methodology in this we detect object on different scales to detect accurately. They have a very big data which is Allen Developing Mouse Brain Atlas data which have less labelled data, from which they have drained fine features related to desired object to be detected. Here they are able to establish multiple to multiple relationships. They have outperformed the bag of word representation in performance. In [9], they have outperformed the previous models in hand detection. They have used multi task learning methodology for better detection of hands and tracking the hands simultaneously.

3. Methodology

Used a simple method to find out the crowd density. Implemented the [1] model to see what result will give in our machine. Here the ground truth density map, Di is determined by,

$$D_i(x) = \sum_{x_g \in S} N(x - x_g, \sigma)$$

Here D_i is the density map for ith patch in the frame. xg the location of the kernel for matching. σ can change scale value for kernel and S is locations set for kernel. The training and evaluation was performed on latest intel i7 processor, with Geforce GTX 1050 Ti with 4/2 GB GDDR5 using pyTorch framework.

To train model, learning rate, λ was set to 0.0001 and momentum to 0.9. Training took a lot of time like few hours. For validation 10% data is used. The processing time training of shanghai dataset took 5 to 6 hours.

4. Implementations and Results

A. Convolution layers

For base there are two convolution layers with activation function as pReLu. The convolution layers have 32 and 16 density maps for kernel sizes 7x7 and 9x9 respectively. These feature map then further propagated to density estimation layers and High-level prior layers.

For classification purpose high-level prior stage, checking the datasets for minimizing the loss by determining the samples for each class.

Results on Shanghai dataset were parametrised on basis of MAE and MSE.

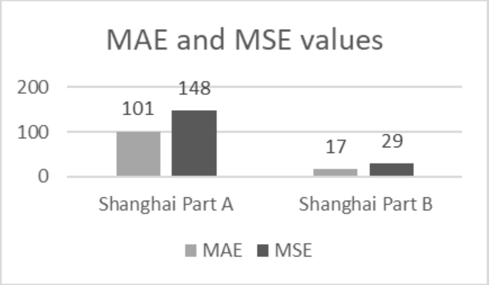

Fig. 5. Mean absolute Error and Mean Squared Error for MTCNN applied model [1]

B. Count Results

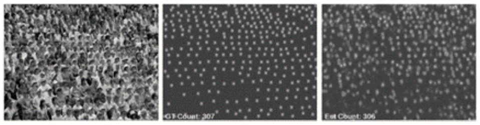

Fig. 6. Crowd Count Result with first(left) image as input and second(middle) image as ground truth and final(right) one as crowd count output

For High-level prior stage, cross entropy loss function is stated below, where N is samples, θ is model parameters, X_i is the ith sample, Fc is the function for classification Y_i is the sample class for 'i' and M is the sum of classes.

$$L_c = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{M}[(y^i = j)F_c(X_i, \theta)]$$

Density loss function is,

$$L_d = \frac{1}{N}\sum_{i=0}^{N} || F_d(X_i, C_i, \theta) - D_i ||$$

Where F is the estimated feature map, Di is ground truth feature map. Di is calculated as,

$$i = \sum_{x_g \in s} N(x - x_g, \sigma)$$

Net loss for model calculated as,

$$L = \lambda L_c + L_d$$

Where λ is the weight factor. Detection of objects and their tracking is necessary for public safety and infrastructure also. Being applied in many applications in industry it is necessary for technology be multi-tasking for giving us better insights so that we improve in our environment. Getting insights from images/ video for surveillance [14] is very important nowadays. On shanghai dataset MTCNN (Multi Task Convolutional Neural Network) have shown better performance and continuously evolving as [6] for different datasets, scenarios.

5. Conclusion and future work

Current Multi Task CNN learning models are performing well in detection, tracking and monitoring. Tried cascading multi column CNN model with MTCNN [1] but got poor results so far, will try to improve results in further study. Researchers have started using multi-instance and multi-task CNN models which are able to take samples or feature/kernels as bag collection for better performance in detection. Industries also wants high tech models only to do work efficiently. Further We will explore hierarchical multi-task learning models with multiple-instance on various datasets.

References

[1] Sindagi, V., Patel, V., 2017. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: Advanced Video and Signal Based Surveillance (AVSS), 2017 IEEE International Conference on, IEEE.

[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE CVPR, pages 589– 597, 2016.

[3] Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE CVPR, pages 833–841, 2015.

[4] Takumi Ege and Keiji Yanai. Simultaneous Estimation of Food Categories and Calories with Multi-task CNN. In 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA) Nagoya University, Nagoya, Japan, May 8-12, 2017.

[5] X. Ouyang *et al*., "A 3D-CNN and LSTM Based Multi-Task Learning Architecture for Action Recognition," in *IEEE Access*, vol. 7, pp. 40757-40770, 2019.

[6] Tao Zeng, SHuiwang Ji, Deep Convolutional Neural Networks for Multi-Instance Multi-Task Learning IEEE Trans. 2015

[7] Jia-Hong Lee, Yi-Ming Chan, Ting-Yen Chen, and Chu-Song Chen Institute of Information Science, Academia Sinica, Taipei. Joint Estimation of Age and Gender from Unconstrained Face Images using Light Weight Multi-task CNN for Mobile Applications. in 2018 IEEE Conference on Multimedia Information Processing and Retrieval

[8] Yingxin Lou, Guangtao Fu, Zhuqing Jiang, Aidong Men, Yun Zhou.Improve Object Detection via a Multi-feature and Multi-task CNN Model. In IEEE 2017.

[9] Yizhang Xia, Shiyang Yan, Bailing Zhang. Combination of ACF Detector and Multi-task CNN for Hand Detection. IEEE 2016.

[10] Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria. Recent Trends in Deep Learning Based Natural Language Processing, 2018

[11] Jianbo Shi and Carlo Tomasi. Good Features to Track. IEEE Conference on Computer Vision and Pattern Recognition, pages 593a600, 1994.

[12] Marcus A. Brubaker, Leonid Sigal and David J. Fleet, "Video Based People Tracking", hand book of ambient intelligence under smart environments 2010.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014

[14] Yiwei Wang, John F. Doherty and Robert E. Van Dyck, "Moving Object Tracking in Video", in proceedings of 29th applied imagery pattern recognition workshop, page 95, 2000.