JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CROWD COUNTING USING SURVEILLANCE FOOTAGE

[1]Savitha SJ,[2]Pravinraj C ,[3]Sakthiswarup JP, [4]Venkatesh Krishna E

[1]Assistant Profesor,[2]Student,[3]Student, [4]Student
Department of Computer Science and Engineering,
Sri Ramakrishna Institute of Technology (An Autonomous Institution), Coimbatore,India.

Abstract: Crowd counting and density estimation is an important and challenging problem in the visual analysis of the crowd. Estimating crowd density or estimating the number of people attending the event can substantially reduce the cost by deploying an exact number of security personnel required for public safety and security

Most of the existing approaches use regression on density maps for the crowd count from a single image. However, these methods cannot localize individual pedestrian and therefore cannot estimate the actual distribution of pedestrians in the environment. On the other hand, detection based methods detect and localize pedestrians in the scene, but the performance of these methods degrade when applied in high-density situations. In our project we have used Faster R-CNN for the detection of pedestrians in the low-to-medium density crowd images that can help in overcoming the drawbacks of existing method.

***IndexTerms* – Convolutional Neural Networks(CNN), Region-based Convolutional Neural Network(R-CNN)**

I. INTRODUCTION

Crowd Counting is a task to count people in image. It is mainly used in real-life for automated public monitoring such as surveillance and traffic control. Different from object detection, Crowd Counting aims at recognizing arbitrarily sized targets in various situations including sparse and cluttering scenes at the same time. Convolutional neural networks, or CNNs for short, form the backbone of many modern computer vision systems. This post will describe the origins of CNNs, starting from biological experiments of the 1950s. Throughout the 1990s and early 2000s, researchers carried out further work on the CNN model. Around 2012 CNNs enjoyed a huge surge in popularity (which continues today) after a CNN called AlexNet achieved state-of-the-art performance labeling pictures in the ImageNet challenge. Alex Krizhevsky et al. published the paper “ImageNet Classification with Deep Convolutional Neural Networks” describing the winning AlexNet model; this paper has since been cited 38,007 times. These image classification neural network can be used to identify images and this technology helps in monitoring the crowd in a video footage.

You are asked to analyze and estimate the number of people who attended each session. This will help your team understand what kind of sessions attracted the biggest crowds (and which ones failed in that regard). This will shape next year’s conference, so it’s an important task! There were hundreds of people at the event – counting them manually will take days That’s where your data scientist skills kick in. You managed to get photos of the crowd from each session and build a computer vision model to do the rest. Crowd counting is essential to serve many real-world applications, such as resource management (such as water, food supply), traffic control, security, disaster management etc. The traditional methods for crowd-counting such as manual counting, using registers to maintain records of each person, and counting through use of sensors are time consuming and tedious, and may produce fallible results due to dynamic movements. This has led to the evolution of crowd-counting methods which rely on CCTV video feeds.

II. LITERATURE SURVEY

In paper [1] authors proposed a method, in which they provide a comprehensive survey of recent Convolutional 12 Neural Network (CNN) based approaches that have demonstrated significant improvements over earlier methods that rely largely on hand-crafted representations. First, then they briefly review the pioneering methods that use hand-crafted representations and then they delve in detail into the deep learning-based approaches and recently published datasets. Furthermore, they discuss the merits and drawbacks of existing CNN-based approaches and identify promising avenues of research in this rapidly evolving field.

In paper [2] authors proposed a method In which the MCNN, they proposed a multi-column parallel convolutional neural network structure that generates population density maps by adapting crowd changes caused by camera view-points and resolution using filters with different size receptive fields. In Switch-CNN, they added a density classifier to the MCNN to enable the use of local density changes in the crowd. In CSRNet, they abandoned the structure of a multi-column convolutional neural network, using the first ten layers of VGG-16 as the front part and the convolutional neural network as the latter part. From the analysis results, CSRNet shows advanced performance.In

Paper [3] the author used the global density features are extracted and added to the MCNN through the cascaded learning method. Because some detailed features during the down-sampling process will be lost in the MCNN and it will affect the accuracy of the density map, an improved MCNN structure is proposed.

In paper [4] the technique that author implements in this paper describes a viewpoint invariant learningbased method for counting people in crowds from a single camera. Our method takes into account feature normalization to deal with perspective projection and different camera orientation.

In paper [5] the authors uses the Big data applications that are consuming most of the space in industry and research area. Among the widespread examples of big data, the role of video streams from CCTV cameras is equally important as other sources like social media data, sensor data, agriculture data, medical data and data evolved from space research. Surveillance videos have a major contribution in unstructured big data.

III. MODULES OF PROPOSED SYSTEM

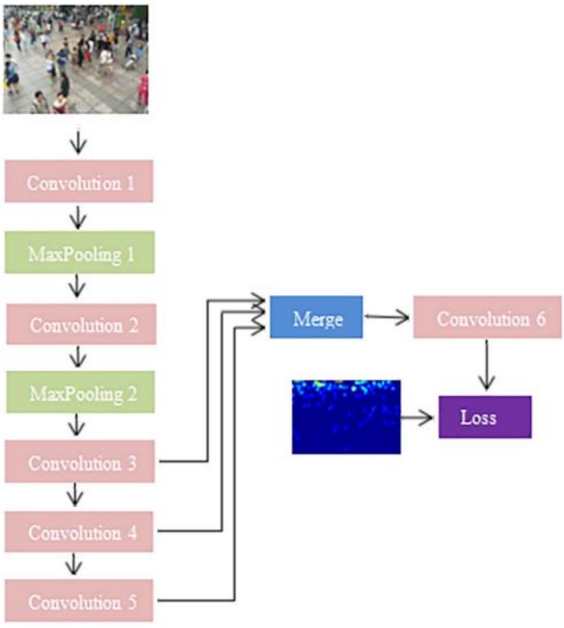
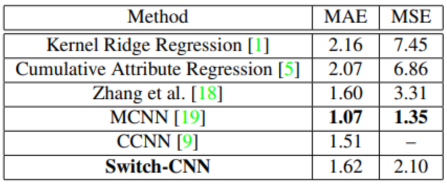


Fig. 1. Overview of the proposed rich convolutional features fusion CNN architecture for crowd counting

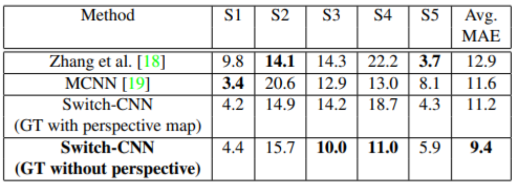
3.1 Dataset

The UCSD dataset The UCSD dataset crowd counting dataset consists of 2000 frames from a single scene. The scenes are characterized by sparse crowd with the number of people ranging from 11 to 46 per frame. A region of interest (ROI) is provided for the scene in the dataset. We use the train-test splits used by [4]. Of the 2000 frames, frames 601 through 1400 are used for training while the remaining frames are held out for testing. Following the setting used in [19], we prune the feature maps of the last layer with the ROI provided. Hence, error is backpropagated during training for areas inside the ROI. We use a fixed spread Gaussian to generate ground truth density maps for training Switch-CNN as the crowd is relatively sparse. At test time, MAE is computed only for the specified ROI in test images for benchmarking Switch-CNN against other approaches. Table 3 reports the MAE and MSE results for SwitchCNN and other state-of-the-art approaches. Switch-CNN performs competitively compared to other approaches with an MAE of 1.62. The switch accuracy in relaying the patches to regressors R1 through R3 is 60.9%. However, the dataset is characterized by low variability of crowd density set in a single scene. This limits the performance gain achieved by Switch-CNN from leveraging intra-scene crowd density variation.



Method	MAE	MSE
Kernel Ridge Regression [1]	2.16	7.45
Cumulative Attribute Regression [5]	2.07	6.86
Zhang et al. [18]	1.60	3.31
MCNN [19]	**1.07**	**1.35**
CCNN [9]	1.51	–
Switch-CNN	1.62	2.10

Table 3. Comparison of Switch-CNN with other state-of-the-art crowd counting methods on UCSD crowd-counting dataset [4].



Method	S1	S2	S3	S4	S5	Avg. MAE
Zhang et al. [18]	9.8	**14.1**	14.3	22.2	**3.7**	12.9
MCNN [19]	**3.4**	20.6	12.9	13.0	8.1	11.6
Switch-CNN (GT with perspective map)	4.2	14.9	14.2	18.7	4.3	11.2
Switch-CNN (GT without perspective)	4.4	15.7	**10.0**	**11.0**	5.9	**9.4**

Table 4. Comparison of Switch-CNN with other state-of-the-art crowd counting methods on WorldExpo'10 dataset [18]. Mean Absolute Error (MAE) for individual test scenes and average performance across scenes is shown.

3.2 Data Pre-processing

In the past, there were much works dealing with crowd counting and crowd density estimation. These techniques used can be classified into two categories. They are pixel counting based and feature-based. In pixel counting based methods [1, 2, 3], background segmentation is first performed to extract the foreground, mainly made of moving people. Then people number is computed by a function of the number of foreground pixels, the function is obtained by learning, such as least squares method. In feature based methods [4, 5, 6], features are computed for the whole image, and segmentation may or may not be required, e.g. texture features are used on the observation that crowded areas exhibit textures and the higher the crowd density the stronger the texture features (contrast, homogeneity, energy, entropy). No matter the technique adopted, the estimation results were only for each individual camera. We can’t analyze crowd density for an area where there are a lot of cameras at different locations in spatial perspective. There have been some attempts to simulate the crowd density and pedestrian behavior in spatial and temporal. [7] presents a system to simulate the movement of individual agents in large-scale crowds performing the Wawaf. This approach uses a finite state machine to specify the behavior of the agents at each time step in conjunction with a geometric, agent based algorithm to specify how an agent interacts with its local neighbors to generate collision-free trajectories. [8] propose a framework for modeling lower-level pedestrian navigational behaviors. In this framework, spatial-temporal patterns are used to represent the situational perception. They construct a computational model to simulate pedestrian behaviors in a corridor with medium to relatively high density of pedestrians. But the simulation results are based on static or virtual data. These spatial-temporal analysis methods not take into account other spatial factors (spatial distribution of facilities, real-time road traffic conditions etc.), but confine their analysis region to a single area or only a large building.

3.3 Component Analysis

In UCSD dataset, our approach employs the mean and variance first layer CNN features in pre-trained offline ImageNet to compute the number of people by SVM regression, and we employ the kernel function of SVM. Then experiment proves that the statistic first layer CNN features is useful and promising performance. Thus the result of the experiment shows that our approach has achieved promising performance by employing the statistic convolutional neurol network (CNN) based features. The advantages of our approach are that we use statistical first layer features in the pre-trained offline ImageNet, and can directly count them by SVM regression. However, our limitations are two parts. First, the performance improvement of our approach is not obvious. Second, the accuracy of our approach has declined in complex environment.We compare the statistic first layer CNN features with the first layer CNN features, and discover that the MAE of the first layer CNN based features to compute the number of people is 9.7. So we employ the statistic first layer CNN features to count crowd.

3.4 Model 1: CNN

Pedestrians in videos pass through a wide range of variations in pose, clothing, lighting and occlusions. This wide range of intra-class variability has a negative effect on the detector's performance. In some cases, the detector missed detection for a particular person in subsequent frames of video. In other cases, the detector ends up with many false positives which results in low recall rates and high Mean Absolute Error (MAE) of a detector. CNN based detectors are designed to learn features from raw image pixels and cannot leverage the temporal information existed across the frames of the video

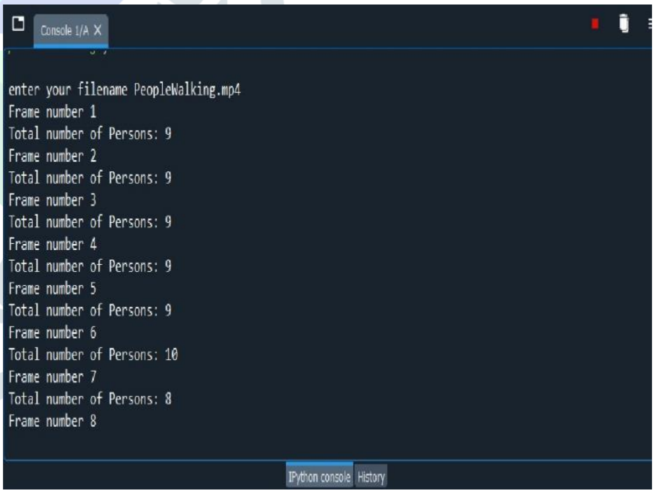
3.5 Model 2: RCNN

In proposed project we have used Faster RCNN for fast and accurate detection of people in crowded areas. And overcome the pitfalls in the previously propped system or mode 1.

IV . RESULT AND CONCLUSION

Crowd counting is essential to serve many real-world applications, such as resource management (such as water, food supply), traffic control, security, disaster management etc. The traditional methods for crowd-counting such as manual counting, using registers to maintain records of each person, and counting through use of sensors are time consuming and tedious, and may produce fallible results due to dynamic movements. This has led to the evolution of crowd-counting methods which rely on CCTV video feeds.

By further enhancement of this project, the major benefit of using counting methods on video feed is that the dynamism of people's movement cannot be incorporated in any of the previous ways of crowd counting. This requires a modern outlook into the problem. An accurate crowd counting system provides solutions for emergency situations such as fire outbreaks, earthquakes and many other disaster situations. In these conditions, an estimate of the crowd would allow the concerned authorities to make the correct decisions regarding supplies of resources.



```
enter your filename PeopleWalking.mp4
Frame number 1
Total number of Persons: 9
Frame number 2
Total number of Persons: 9
Frame number 3
Total number of Persons: 9
Frame number 4
Total number of Persons: 9
Frame number 5
Total number of Persons: 9
Frame number 6
Total number of Persons: 10
Frame number 7
Total number of Persons: 8
Frame number 8
```

Figure 6.1: console output

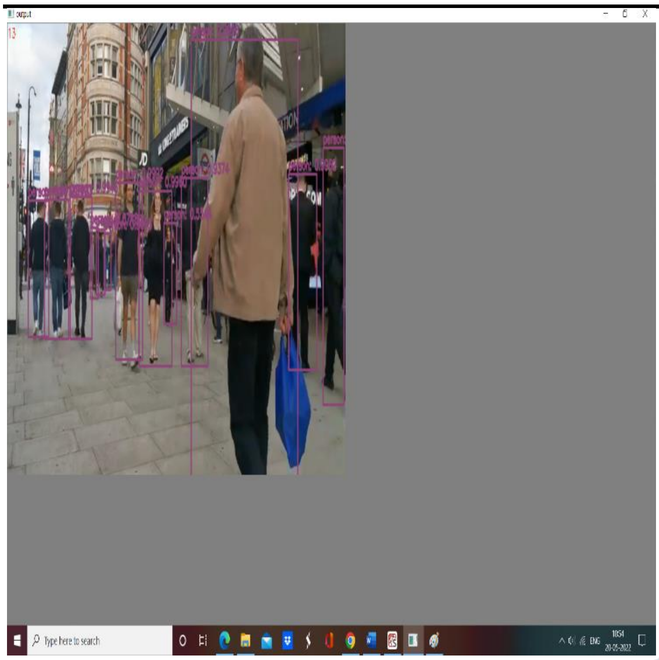
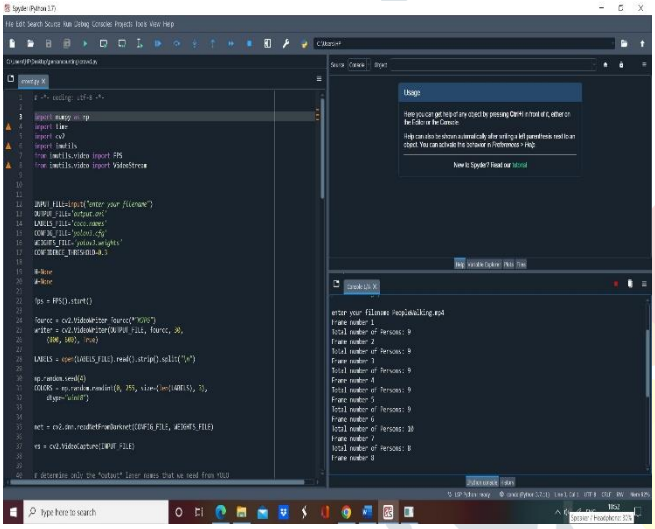


Figure 6.2: video output



```
# -*- coding: utf-8 -*-

import numpy as np
import time
import cv2
import imutils
from imutils.video import FPS
from imutils.video import VideoStream



INPUT_FILE=input("enter your filename")
OUTPUT_FILE='output.avi'
LABELS_FILE='coco.names'
CONFIG_FILE='yolov3.cfg'
WEIGHTS_FILE='yolov3.weights'
CONFIDENCE_THRESHOLD=0.3

H=None
W=None

fps = FPS().start()

fourcc = cv2.VideoWriter_fourcc(*"MJPG")
writer = cv2.VideoWriter(OUTPUT_FILE, fourcc, 30,
    (800, 600), True)

LABELS = open(LABELS_FILE).read().strip().split("\n")

np.random.seed(4)
COLORS = np.random.randint(0, 255, size=(len(LABELS), 3),
    dtype="uint8")


net = cv2.dnn.readNetFromDarknet(CONFIG_FILE, WEIGHTS_FILE)

vs = cv2.VideoCapture(INPUT_FILE)


# determine only the *output* layer names that we need from YOLO
```

Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Help.

New to Spyder? Read our tutorial

```
enter your filename PeopleWalking.mp4
Frame number 1
Total number of Persons: 9
Frame number 2
Total number of Persons: 9
Frame number 3
Total number of Persons: 9
Frame number 4
Total number of Persons: 9
Frame number 5
Total number of Persons: 9
Frame number 6
Total number of Persons: 10
Frame number 7
Total number of Persons: 8
Frame number 8
```

Figure 6.3: console application output

Fig 6 : Result

REFERENCES

[1] A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation Vishwanath A. Sindagia,∗∗, Vishal M. Patelb aDept. of Electrical and Computer Engineering, 94 Brett Road, Piscataway, NJ 08854, USA bDept. of Electrical and Computer Engineering, 94 Brett Road, Piscataway, NJ 08854, USA

[2] The Comparison of Crowd Counting Algorithms based on Computer Vision Zhaoqing Wang1, * , Qishu Deng2, a , Yusheng Zhao2, b 1School of Information & Communication Engineering, Beijing Information Science& Technology University, Beijing, China 2 International School, Beijing University of Posts and Telecommunications, Beijing, China.

[3] The global density features are extracted and added to the MCNN through the cascaded learning method. Because some detailed features during the down-sampling process will be lost in the MCNN and it will affect the accuracy of the density map, an improved MCNN structure is proposed.

[4] A Viewpoint Invariant Approach for Crowd Counting Dan Kong, Doug Gray and Hai Tao Department of Computer Engineering University of California, Santa Cruz Santa Cruz, CA 95064 (kongdan,dgray,tao)@soe.ucsc.edu

[5] Intelligent video surveillance: a review through deep learning techniques for crowd analysis G. Sreenu* and M. A. Saleem Durai *Correspondence: gsreenug@gmail.com VIT, Vellore 632014, Tamil Nadu, India

[6] Zhaoqing Wang et al 2019 J. Phys.: Conf. Ser. 1187 042012. ISPECE. IOP Conf. Series: Journal of Physics: Conf. Series 1187 (2019) 042012. IOP Publishing. doi:10.1088/1742- 6596/1187/4/042012

[7] This research is partially supported by Key Projects in the National Science & Technology Pillar Program during the Twelfth Five-Year Plan Period under Grant Grant No. 2012BAH35B02 and the Major Basic Research Funded Projects of Jiangsu Province under GirantNo.10KJA420025.