Abstract

In today's world, the exponential growth of the worldwide population has caused a considerable increase in crowd densities in places like rallies, public speeches, stadiums, mass transit, and tourist or pilgrim sites. Such places are vulnerable to crowd disasters. Crowd disasters are controlled by efficiently analyzing crowd scenes. The crowd analysis (CA) is beneficial for drawing better crowd management strategies, and public space design. It helps in developing an AI-based visual surveillance system to provide security and safety to the crowd. However, the CA is a tedious task that comprises several correlated tasks, out of which crowd counting and density estimation (CCDE), crowd congestion-level analysis (CCA), and crowd behavior analysis (CBA) are the minimum tasks required to control crowd disasters. Most existing solutions for crowd analysis are done manually in real world scenario, which are very complex and prone to error. Recently, deep learning techniques have significantly improved the performance of several computer vision tasks and contributed considerably to the development of AI-based solutions. Thus, the drawback of the manual process for the CA can be overcome by developing automated AI-based solutions using computer vision and deep learning solutions.

Further, CA using several single-task AI models will incur computational complexity overheads, which can be minimized by drawing better multitasking CA models. So, to control crowd disasters and provide security and safety to the crowd, efficient solutions using computer vision and deep learning approaches for significant tasks of CA such as CCDE, CCA, CBA, and multitasking CA, are required. However, as per the literature review reported later, the performance of several computer vision-based CA models is mainly affected by cluttered background, varying crowd densities, crowd

shape changes due to perspective distortion, illumination changes, and lack of availability of largescale CA datasets. Therefore, to address the issues mentioned above, the problem statement of the thesis is defined as the design and development of some methods and models for crowd analysis using computer vision and deep learning techniques.

This thesis mainly focuses on studying and analyzing the state-of-the-art CA techniques, finding their advantages and limitations, and proposing new methods and models to accomplish the objectives. This thesis aims to conduct a comprehensive literature review on four tasks of CA, i.e., CCDE, CCA, CBA, and multitasking CA. For each task, various methods and models concerning current research trends have been analyzed by mentioning their pros and cons and identifying possible research scopes. Various models using computer vision and deep learning approaches have been proposed in this thesis to fulfill the research scopes of each of the four tasks of CA.

The first contribution in the thesis is related to the task of Crowd Counting and Density Estimation (CCDE), where two models using deep learning techniques have been proposed. The first proposed model for CCDE is an **A**ttentive **M**ulti-**S**tream **CNN** (AMS-CNN) for video-based crowd counting. The main objective behind the AMS-CNN is to enhance the feature representation for crowd videos, minimize the effect of the cluttered background, and design attention mechanism for each stream to improve the counting performance. The second proposed model for CCDE is a novel cascaded deep architecture with weak supervision for video crowd counting. This model comprises two deep models named **L**ocal **D**ensity-map **R**egressor (LDR) and **G**lobal **C**rowd **C**ounting **R**egression (GCCR) modules. The LDR focuses on extracting multiscale spatial-temporal features using a multicolumn 3D Atrous (Dilated) CNN to tackle crowd shape changes due to perspective distortion. It also minimizes the effect of the cluttered background using a **H**ead **A**ttention **M**odule (HAM). The LDR considers the local distribution of

crowds and generated crowd density maps. On the other hand, the GCCR model is trained in a weakly supervised manner to exploit global crowd properties from the predicted density maps to obtain final crowd counting using a multi-layer perceptron neural network.

The second contribution to the thesis is related to the task of Crowd Congestion Level Analysis (CCA), where deep learning-based real-time **T**wo **I**nput **S**tream **M**ulti-**C**olumn **M**ulti-**S**tage **CNN** (TIS-MCMS-CNN) is proposed. In the proposed method, each of the two streams of TIS-MCMS-CNN has been built with three columns of multi-layers of CNN with different receptive fields to extract multiscale spatial and temporal features from two cues of video frames, i.e., frame and the flow magnitude of the frame. The extracted multiscale features are also known as scale-invariant features, which can handle crowd shape change due to perspective distortion. For experimental analysis, a dataset for the CCA is prepared using three publicly available benchmark crowd datasets. It is observed that the TIS-MCMS-CNN can process the frames in real-time.

The third contribution to the thesis is related to the task of Crowd Behavior Analysis (CBA). Under this contribution, two models have been proposed out of which the first model i.e., a **Mu**ltiscale **S**patial-**T**emporal 3D Atrous-Net with **P**CA-guided **O**C-**S**VM (MuST-POS) is designed using conventional machine learning and deep learning approaches to classify normal and panic crowd behaviors. On the other hand, the second model is a **T**wo-**S**tream **M**ultiscale **D**eep **A**rchitecture (TS-MDA) is developed using deep learning techniques for multiclass crowd behavior prediction. Both the models resolve the issue of human shape variation in the crowd videos, while the second model takes measured to minimize the effect of cluttered backgrounds in the crowd video.

The fourth contribution to the thesis is related to multitasking Crowd Analysis (CA), where an efficient multitasking deep model is proposed using a backbone structure

of multi-layer **D**epth-wise **S**eparable **CNN** (DSCNN) to predict crowd behaviors with crowd counting. The proposed model exploits spatial-temporal features with **F**low **A**ttention **B**locks (FABs) to provide optical flow attention to different scales features of the backbone network. The FABs focus on the moving pixels, thereby minimizing the effect of cluttered background. The multiscale flow attentive features are fused to handle crowd shape variation and perform multitask CA using a feed-forward network. In addition, a largescale multitasking CA dataset is also developed from the available benchmark crowd behavior datasets. Around 1,20,000 frames have been annotated to obtain ground truth crowd counting information.

The proposed models have been implemented, experimented and evaluated on several publicly available datasets. Experimental results and analysis show that the proposed models perform better than state-of-the-art methods reported in the literature. Extensive ablation studies for each of the proposed models have also been conducted to show the influence of several components of the proposed models.

The thesis concludes with an overall conclusion of the proposed research work, followed by a discussion of possible future research scopes in the areas of CA.