Pedestrian detection and gender recognition utilizing YOLO and CNN algorithms

Zixuan Zhong

Software and Systems engineering, Lappeenranta University of Technology, Kela, 53851, Finland

Zixuan.Zhong@student.lut.fi

Abstract. As crowd-based activities continue to surge in locales such as markets and restaurants, the significance of understanding pedestrian flow is increasingly evident. Over recent years, advancements in dynamic pedestrian detection, facilitated by the YOLO (You Only Look Once) algorithm, have seen widespread application in areas like crowd management and occupancy estimation. The YOLO algorithm has demonstrated high accuracy and efficiency in real-time object tracking and counting. However, for specific use cases, data derived solely from monitoring pedestrian flows may prove inadequate. This study presents YOLO-Gender, a system leveraging YOLO and Convolutional Neural Network (CNN) for pedestrian tracking and gender classification. The objective is to enhance the richness of data extracted from surveillance camera footage, thus rendering it more valuable for societal applications. The YOLO suite of algorithms, hailed for their superior performance and rapid iteration speed, is among the most extensively utilized tools in the field. The proposed system is predicated on YOLO v8, the most advanced iteration of the YOLO algorithm, released in 2023, which boasts its highest accuracy to date.

Keywords: YOLO v8, pedestrian tracking, gender classification, Convolutional Neural Network, transfer learning.

1. Introduction

As society's demand for entertainment functionality and publicity steadily increases, understanding public behavior and the underlying social tendencies has become critical [1]. The emergence of computer vision and the enhanced reliability of surveillance technology have revolutionized our ability to monitor and interpret human movement across various settings. The YOLO algorithm [2] has been adopted widely due to its proven effectiveness in real-time object tracking and counting, delivering remarkable accuracy and efficiency.

However, the information derived solely from people detection is restricted. To enrich data extracted from surveillance cameras by incorporating additional gender-relevant information, and further assist in understanding social behavior and gender-based patterns, this study introduces the YOLO-Gender system. This system adds value to urban planning and sociological studies by providing gender-based insights [3]. The system refines detection by integrating a gender classification and counting system into the pipeline of a YOLOv8-based people flow tracking system, to implement secondary processing on the CCTV footage, divided into frame images according to the frame rate. The successful implementation of gender recognition and classification using transfer learning models, trained on

ImageNet, represents a breakthrough in terms of efficiency and speed, surpassing the traditional CNN model on the UTK Faces dataset. Furthermore, a comprehensive evaluation of the YOLO-Gender system has been carried out, including the impact of secondary classification on the inference speed of the entire system and the challenges of collecting facial data and extracting gender information. The evaluation results show a higher accuracy in face detection and gender classification (percentage) than the original YOLOv8 on the MOT20 dataset. This demonstrates that the secondary classification exhibits a higher generalization ability in face recognition and gender classification and is beneficial for enhancing the information content of data extracted from surveillance cameras.

The paper is organized as follows: the second section reviews and analyzes existing YOLO-based people flow tracking systems and CNN-based gender classification models; the third section primarily discusses the methodology adopted in this study; the fourth section presents the implementation of the YOLO-Gender system, including the YOLO-based people flow tracking system and CNN-based gender classification system, and their integration; the fifth section focuses on the evaluation standards and results; the sixth and final section provides the conclusion and suggests future improvements.

2. Literature review

2.1. Development and application of YOLO-based people flow detection systems

You Only Look Once is computer vision algorithm that has gained much popularity, especially in the field of multiple object tracking [4], since its first version was introduced by Redmon et al. in 2015 [5]. Differs from technics based on Convolutional Neural Network, YOLO based object detectors considers to dimensionally separate the bounding boxes and interpret their class probabilities with a basis of regression [6]. In years of development, the more advanced version of YOLO including YOLOv1 to YOLOv8, and YOLOnas et al has been developed and published sequentially, with each version was improved and specialized in different tasks. Due to its small size and short inference time, YOLO algorithms has always been a widely adopted method in people flow detection tasks, and especially suitable for detecting dynamic footage. Meanwhile, the straightforward output of YOLO including the coordinates and category of the bounding boxes through its neural network [4] simplifies the adaption process within the secondary processing pipeline.

2.2. Limitations of existing people flow tracking models

The limitations of existing model mainly comes from the following aspects:

Occlusion: The detector may lose track on people when they are fully or partially occluded by some objects [6]. Resolution: Depend on the the resolution (either too high or too low) of row videos are going to be processed the results of detection are affected to some extent. Light condition: Excessive or dim light condition impedes feature extraction on objects to some extent, thus affecting accuracy. Processing speed: The processing speed of the detection model affects efficiency of the actual work effects (e.g. real-time surveillance camera needs fast processing speed to track objects in time) [7].

2.3. Gender classification models

Through difference features like appearances, voices and texts, multiple approaches are applied in gender classification tasks, this paper will only discuss image based classification tasks, namely, no voice or texts information will be taken into consideration. Image based gender classification are divided into two technics. Two technics are geometry-based classification and appearance-based classification. Geometry-based method relies on measuring the geometric properties of faces in images like the distance between the eyes, the width of the nose, the shape of the jaw, and other structural or proportional aspects of the face.

Appearance-based method focuses on the texture and color of the face, which can be thought of as the "appearance" of the face. It often involves extracting features from the pixel values of the image directly, without explicitly measuring geometric properties. This method is wider applied than geometry-based one due to the abilities of automatic learning and feature extracting of the deep learning models build on it.

The algorithms usually applied to gender classification tasks is Convolutional Neural Networks, which works with constructed layers to progressively extract and learn the features on images. Its ability to learn complex patterns and high-level features directly from image data makes it competent in complex real-time cases processing. The derivatives of CNN includes Visual Geometry Group Network (VGGNet), which is possessed with simple architecture with a focus on using many convolutional layers instead of a larger and more complex network. Residual Neural Network (ResNet), uses "shortcut connections" or "skip connections" to address the problem of training very deep neural networks. Transfer learning, instead of training a model from scratch, it involves a pre-trained model (often trained on a large-scale image recognition task like ImageNet) and fine-tuning it for a specific task like gender classification, namely, store the knowledge gained in pre-train tasks and apply them to the following tasks to significantly improve the efficiency and performance of learning.

3. Methodology

3.1. YOLO v8 algorithm

There are 8 versions of Yolo and various models that have been created to date. Yolo v8 is the most recent and most capable of them all. In addition Yolo v8 is particularly suitable for the task of people or facial detection regarding its speed (about 6.3% to 33% advantage when compared to Yolo v5 on many datasets) and accuracy (about 1% more accurate on large datasets than Yolo v5). Compared to R-CNN, the Yolo algorithm in general is more than 50% more accurate when testing on the same dataset with the same hardware condition [7].

In experimental conditions, Yolo v8n or Yolo nas (nano) could achieve up to 525 FPS, making it an ideal fit for people flow tracking based on videos provided by surveillance cameras. The Yolov8 framework provided by roboflow is used. This framework provides pertain weights [8]. In this project, yolov8 with 43.7M parameters is applied.

3.2. Limitations of traditional CNN model

A traditional CNN model was built first to deal with gender classification tasks. However, the experiment result shows it struggles with images that have variations in lighting, occlusion, expression, and other factors that don't change the gender of the subject, thus lowers the classifying accuracy.

3.3. Transfer learning

Attempted has been paid to train data through a traditional CNN model, but the ultimate accuracy in real detection was not ideal, since video footage contain scenes under various conditions and textures. Therefore, transfer learning was used. Here, the VGG-16 model, which has been pre-trained on a large dataset ImageNet, was used as a base model. VGG-16 is a deep learning model, which is a Convolutional neural network structure proposed by the Visual Geometry Group (VGG) of Oxford University in 2014. It consists of multiple convolutional layers and fully connected layers, totaling 16 weight layers, hence the name "VGG-16". The model is straightforward and relatively easy to modify to be trained for current tasks. The model has learned numerous feature detectors from pre-trained datasets, including colors, textures, shapes, etc., which can be used for numerous visual tasks [9].

3.4. Integration of transfer learning in the model

Generally, the system works as: the face detection system detects human faces from provided videos and passes the coordinates of the bounding boxes to the gender classification system, which will then recognize the gender and annotate on the output video [10].

4. System implementation

4.1. Implementation of YOLO people tracking system

The Yolov8 framework provided by roboflow is used. This framework provides pertain weights. In this project, yolov8 with 43.7M parameters is applied.

A dataset of human face detection from roboflow universe is used when training yolov8. This dataset contains 1386 images, each labeled with bounding boxes of human faces. This dataset is chose because its amount is manageable, showcasing a good performance after training. Moreover, this dataset contains pictures of faces viewed from different angles and pictures containing multiple faces, including some faces very far from the camera.

4.2. Implementing the transfer learning model

When training the VGG-16 model on custom dataset, its pre-trained layers were frozen, so as to preserve the pre-learned features. After that, the final fully connected layers that are used in the original ImageNet tasks are replaced with a new trainable fully connected layer, which matches the binary output dimensions for gender classification (i.e. 1 for women and 0 for men).

The custom final layer was trained on UTK Face dataset, which consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

It is necessary to address that an ideal data set for classifying gender would be based on the image of the entire human body, including a person viewed from different angles and a person that is blocked by some obstacle. However, the only gender classifying data set found is based on faces and this dataset only contains human faces viewed from front with no obstacles like masks. This is the main problem that is affecting the final result. Nevertheless, to overcome this problem, data augmentation like image cropping is applied to address the cases where obstacles were present.

4.3. Integration of the YOLO and transfer learning systems

The sequence of execution for each frame is as follows:

a. A YOLO model is initialized for face detection.

b. Within a looping structure, each video frame is read and then passed to the YOLO model.

c. Once YOLO detects the objects, the coordinates of the bounding boxes are utilized to draw bounding boxes around each face.

d. Within the function that draws these boxes, for each bounding box (or Region of Interest - ROI), the image is resized and normalized. Subsequently, the script performs gender classification using the secondary system and labels the bounding box with the predicted gender.

e. This frame is then written to the output video, and the process proceeds to the next iteration of the loop.

Counting is performed every fifth frame, with IDs present in five consecutive frames being placed into a set (which does not allow duplicate keys), along with their corresponding class (either man or woman). The final count is computed at the end of the video.

4.4. Overall functioning of the integrated system

The footage from the CCTV would be used as original material for sorting. It would be divided into a series of images or frames depending on the frame rate of the footage. Each frame would be firstly processed by the detection system, and then the bounding box of detected human faces would be passed to gender classification model for secondary classification. Finally, the output footage would be a tagged video that could be further analyzed and thus return the people flow and the number of males or females present in the crowd.

5. Experiments and results

5.1. Evaluation metrics

a. Precision: Precision is the proportion of true positive predictions in the total predicted positives. It shows how many of the examples that the model predicted as positive are actually positive. A high precision means a low false-positive rate.

b. Recall: Recall, also known as sensitivity or true positive rate (TPR), is the proportion of true positive predictions in the total actual positives. It shows how many of the actual positive examples the model was able to identify. A high recall means a low false-negative rate.

c. F1 Score: The F1 score is the harmonic mean of precision and recall, and it tries to balance the two. It is particularly useful in scenarios where one is interested in the optimal balance between precision and recall or when dealing with imbalanced datasets. The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

d. mAP@.50: This metric computes the mAP where the Intersection over Union (IoU) between the predicted bounding box and the ground truth bounding box is at least 0.50. The IoU is a measure of the overlap between two bounding boxes. If the overlap is greater than or equal to 0.5, the prediction is deemed a 'hit'. This metric is less stringent about the location accuracy of the bounding box, and it's more about 'object detection' than 'object location'.

e. mAP@.50-.95 (also written as mAP@[.50:.05:.95] or mAP@IoU=.50:.05:.95): This metric computes the average mAP for IoU from 0.5 to 0.95 with step size 0.05. This means it calculates the mAP at different levels of IoU from 0.5 to 0.95 (0.5, 0.55, 0.6, ..., 0.95), and then averages the results. By incorporating a range of IoU thresholds, it is more stringent and provides a more comprehensive view of how the model performs not only in 'object detection' but also 'object location'.

5.2. Analysis of experiment results

The integrated model is experimented on several testing sites, including DELL G3 i7-10750+2060, ROG16 i9-12700+3080 etc., the experiment was implemented on MOT-20 dataset. The dataset contains a total of eight video clips from three different scenarios, and each video clip is provided in the form of video frames [8]. Four of the video, 4,479 frames in total are for testing. the precision score, recall score, precision-recall score and F1 score was measured. The integrated model has shown good and consistent performance, results are shown below:

According to the confusion matrix and result curve, the gender classification model using transfer learning has reached a precision score of 0.94 and a recall score of 0.96, an F1 score of 0.95 (better than traditional CNN: precision 0.91, recall 0.91, F1 score 0.91), which indicates the model has a good balance between precision and recall. The overall experiment result indicates the gender classification model is possessed with a high accuracy in recognizing gender, and a relatively low possibility of false alarm rate. According to the confusion matrix and result curve, the face detection model based on YOLO has reached a precision value of 0.9 and a recall value of 0.47, and an F1 score of 0.62, indicating the model has a high accuracy in recognizing faces while some flaws in predicting false negative samples, namely, mistakenly recognize those objects that are not faces as faces. This flaw is mainly caused by the limitation of dataset as there are not enough comprehensive face and body images, and the time limit for model optimization.

According to the mAP50 metrics, the IoU (Intersection over Union) threshold is set at 0.5. This means that as long as the overlap of the predicted bounding box and the actual bounding box is 50% or more, the prediction is considered correct. The mAP@.50 score is 0.86, indicating that the model performs very well with this relatively lenient IoU threshold. According to the mAP50-95 metrics, which includes a range of IoU thresholds, from 0.5 to 0.95, incremented by 0.05. The mAP@.50-.95 score is 0.504, indicating that the model's performance is moderate under stricter IoU thresholds. In summary, both the gender classification model and the face detection model have demonstrated good performance in their respective tasks, with the gender classification model demonstrating excellent precision and recall. The face detection model, while performing well in terms of precision, does show some limitations in its recall. Object detection metrics further confirm that while the face detection model is effective in detecting the general position of the target, there is room for improvement in accurately defining the target boundaries, especially at stricter IoU thresholds. Future work should focus on addressing these limitations to further improve the performance of the face detection model.

6. Conclusion

This study introduces an integrated system that fuses a YOLO-based people flow tracking and face detection mechanism with a CNN transfer learning approach for gender classification. Our goal was to improve people management strategies and foster the development of gender-focused functionalities.

The integrated model demonstrated remarkable performance in evaluation experiments. The gender classification module, leveraging CNN transfer learning, exhibited outstanding accuracy and efficiency. However, the YOLO-based face detection component demonstrated acceptable accuracy only under more stringent threshold settings, highlighting potential areas for future enhancement. Specifically, the need for a more comprehensive dataset featuring full-body images and a wide range of facial angles, along with further fine-tuning of parameters, has become evident. The primary strength of YOLO lies in its ability to carry out three tasks within a single network: classification, localization, and confidence level estimation. The model's loss function is the summation of these three components, facilitating multitasking via loss function minimization. However, we have limited insights into the mathematical underpinnings of this process. The extent to which the segregation of the classification task could improve performance is a subject that demands substantial experimentation and testing, which exceeded the scope of our two-week project.

References

[1] Vallimeena P, Gopalakrishnan U, Nair B B, et al. 2019 Detection of Human Face Attributes using CNN Algorithms – A Survey Proc. 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (IEEE) pp. 576-581

[2] Hou C. 2023 Human Detection Based on YOLOv5 Application Highlights in Science, Engineering and Technology vol. 34 pp. 203-208
Karahan M, Lacinkaya F, Erdonmez K, et al. 2022 Age and Gender Classification from Facial Features and Object Detection with Machine Learning Journal of Fuzzy Extension and Applications vol. 3(3) pp. 219-230

[3] Sumit S S, Awang Rambli D R, Mirjalili S, et al. 2022 Improving the Performance of Tiny-YOLO-Based CNN Architecture for Human Detection Applications Applied Sciences vol. 12(18) pp. 9331

[4] Priya K P L, Jyothirmai I, Akshaya G, et al. 2023 Identification of Autism in Children Using Static Facial Features and Deep Neural Networks Turkish Journal of Computer and Mathematics Education (TURCOMAT) vol. 14(2) pp. 704-715

[5] Jabraelzadeh P, Charmin A, Ebadpore M. 2020 Hybrid Method for Face Detection, Gender Recognition, Facial Landmarks Localization and Pose Estimation Using Deep Learning to Improve Accuracy Journal of Artificial Intelligence in Electrical Engineering vol. 8(32) pp. 1-14

[6] Vishwakarma H, Verma G, Singh S, et al. 2019 Single Shot Multi-Face Detection & Gender Recognition Proc. 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)

[7] Gawande U, Hajari K, Golhar Y. 2022 SIRA: Scale Illumination Rotation Affine Invariant Mask R-CNN for Pedestrian Detection Applied Intelligence vol. 52(9) pp. 10398-10416

[8] Lee S, Hwang J, Kim J, et al. 2023 CNN-Based Crosswalk Pedestrian Situation Recognition System Using Mask-R-CNN and CDA Applied Sciences vol. 13(7) pp. 4291

[9] Nagajyothi D, Charan P S, Zeeshan M, et al. 2023 Image Enhancement for Pedestrian Detection at Night Time Proc. 2023 2nd International Conference for Innovation in Technology (INOCON) (IEEE) pp. 1-7