Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230113

Advanced people flow monitoring and gender classification: A joint application of YOLO, DeepSORT, and convolutional neural networks

Qianjun Li

University of California San Diego, San Diego, California, CA 92093, USA

qil008@ucsd.edu

Abstract. In an era of increasingly crowd-focused activities, understanding the dynamics of people's flow is of paramount importance. However, data derived solely from detecting and counting individuals can prove inadequate for certain use-cases. Addressing this deficiency, this study introduces a robust method to enrich data extraction, thereby enhancing its value to a wide range of stakeholders. Presented herein is a novel system dubbed YOLO-Gender, an innovative integration of YOLO and CNN, designed to deliver comprehensive people tracking and gender classification. This gender recognition component provides a much-needed edge in crowd management and facilitates efficient planning of gender-specific services. The core foundation of the system is built upon YOLOv8, the apex of the YOLO model series, renowned for its unparalleled accuracy and efficiency. Through the use of transfer learning models pre-trained on ImageNet, gender recognition is achieved, showcasing a marked enhancement over conventional CNN models. Assessments of this system validate its robust performance, underlining its potential for large-scale deployment. This study represents a significant step forward in AI-powered surveillance, offering a solution that effectively enriches and analytically processes extracted data.

Keyword: YOLO, CNN, YOLO-Gender.

1. Introduction

The ever-increasing demand for public entertainment and visibility, as societies emerge from the pandemic era, has underscored the importance of comprehending public behavior and the underlying social trends [1]. The introduced system incorporates a secondary classification stage - gender recognition, offering substantial benefits for refined crowd management, planning gender-specific events, and bolstering security measures [2]. This facility empowers businesses and public spaces to fine-tune their services, safety protocols, and amenities to align accurately with visitor demographics. Furthermore, the system offers an avenue for understanding social behavior and gender-based patterns, yielding valuable insights for urban planning and sociological research [3].

The proposed system comprises a face detection module, grounded on YOLO, and a gender recognition and classification module based on a CNN model employing transfer learning. Typically, the face detection module identifies human faces and transmits the coordinates of the bounding boxes to the gender classification module, which subsequently determines the gender and annotates it on the

DOI: 10.54254/2755-2721/30/20230113

output video. Assessments underscore that the integration of a secondary classification stage does not undermine the system's overall efficiency. The outcomes substantiate the feasibility of such comprehensive systems in real-world environments, showcasing their ability to manage large scale data and deliver gender insights in extensive applications [4].

2. Background

Eight versions of Yolo have been developed to date, with Yolo v8 being the most recent and capable iteration. Yolo v8 excels in the task of people or facial detection, outperforming Yolo v5 in both speed and accuracy. Speedwise, Yolo v8 holds an advantage of around 6.3% to 33% over Yolo v5 across various datasets, while also achieving approximately 1% greater accuracy on large datasets. Compared to R-CNN, Yolo's overall accuracy surpasses it by over 50% when tested under the same hardware conditions and dataset [5].

In experimental setups, Yolo v8n or Yolo nas (nano) can achieve up to 525 FPS, positioning it as an ideal choice for tracking people flow based on CCTV footage. Despite these advantages, Yolo does have its limitations. It struggles with detecting small objects or fragments of objects. This issue can be observed in the output video, where individuals partially obscured by others may not be detected. Nevertheless, attempts to find better alternatives have yielded inferior results, with Yolo maintaining its superior balance of speed and accuracy. As for the gender classifier, two principal methods exist: appearance-based and geometry-based. To construct the most effective CNN and achieve the highest accuracy, the choice falls on the appearance-based classification. The CNN layer is designed to follow the Yolo detection module. However, one factor that may limit the accuracy of the classifier is the gender imbalance within the datasets used, as there are more images of men than women.

3. Approach

3.1. YOLOv8 algorithm

The YOLOv8 framework, facilitated by Roboflow, is implemented in this endeavor. It utilizes pre-trained weights and houses a staggering 43.7 million parameters [6]. For the purpose of training YOLOv8, a human face detection dataset sourced from the Roboflow Universe serves as the primary resource. This specific dataset boasts 1386 images, each marked with bounding boxes to highlight human faces. The selection of this dataset is attributed to its manageable size, ensuring promising performance post a training period of 1 to 2 hours. Notably, the dataset comprises images of faces captured from various angles and visuals featuring multiple faces. It even includes faces positioned considerably distant from the camera. Despite the low-confidence faces (those too far from the camera) not being forwarded to the classifier, their inclusion in the dataset reinforces the robustness of the YOLO model, rendering it compatible with a wider array of classifiers.

3.2. Gender classifier

Two approaches are applied to train models that classify gender based on face. The first one is a traditional CNN, and the second one is a transfer learning based on VGG-16 with one trainable fully connected layer. The results will be shown in the evaluation section.

3.3. Dataset for gender classifier

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. And the yolo should detect the entire human body as well. However, the only gender classifying data set we found on the internet is based on faces and this data set only contains human faces viewed from front and there are no obstacles like masks. This is the main problem that is affecting our final result. For example, if the person with a mask or showing only a part of the face is detected and passed into the classifier, the result will be unexpected. Nevertheless, to overcome this problem, data augmentation like image cropping is applied to address the cases where obstacles were present [7].

DOI: 10.54254/2755-2721/30/20230113

3.4. Integration of the two systems

For the purpose of this project, object detection and classification is done sequentially (no parallelism) . Because this project is used to count numbers of men and women, it is not required to perform this calculation in real time. The specific order of execution for each frame is the following [8]:

a. A yolo is initialized for detecting faces.

b. In a while loop, each frame of the video is read and passed to yolo.

c. Once the objects are detected by YOLO, the bounding boxes coordinates are used to draw bounding boxes around each face [9].

d. These bounding boxes are passed to deepsort and deepsort returns the updated bounding boxes and tracking IDs.

e. Inside the draw_boxes function, for each bounding box (ROI), the image is resized and normalized and then the script performs gender classification using the GenderClassifier and annotates the bounding box with the predicted gender.

f. Write this frame to output the video and enter the next iteration of the loop.

Counting is done every 5 frames, ids that are present in consecutive 5 frames are put into the set(no duplicate key), as well as their corresponding class(man or woman). Final number is calculated at the end of the video [10].

3.5. How the integrated system works

The footage from the CCTV would be used as original material for sorting. It would be divided into numbers of images or frames depending on the frame rate of the footage. Each frame is being processed with the method mentioned before and 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.

4. Experiments and evaluation

4.1. Explanation of evaluating metrics

Precision: Precision represents the ratio of correctly predicted positive instances out of the total predicted positive instances. It gauges how many of the cases identified as positive by the model are indeed positive. High precision correlates with a low false-positive rate.

Recall: Also referred to as sensitivity or true positive rate (TPR), recall is the proportion of correct positive predictions relative to the total actual positives. It measures how many of the genuine positive instances the model managed to correctly identify. High recall correlates with a low false-negative rate.

F1 Score: This is the harmonic mean of precision and recall, seeking to establish an optimal balance between the two. The F1 score proves particularly valuable in scenarios where an ideal equilibrium between precision and recall is sought or in situations dealing with imbalanced datasets. The F1 score's best potential value is 1 (signifying perfect precision and recall), with the worst being 0.

mAP@.50: This metric calculates the mean Average Precision (mAP) where the Intersection over Union (IoU) between the forecasted bounding box and the actual bounding box is at least 0.50. The IoU is a measurement of the overlap between two bounding boxes. If this overlap equals or exceeds 0.5, the prediction is considered a 'hit'. This metric focuses less on the precise location accuracy of the bounding box, emphasizing more on 'object detection' rather than 'object location'.

mAP@.50-.95 (also denoted as mAP@[.50:.05:.95] or mAP@IoU=.50:.05:.95): This metric computes the average mAP for IoU ranging from 0.5 to 0.95 with step size increments of 0.05. This means it determines the mAP at various levels of IoU from 0.5 to 0.95 (0.5, 0.55, 0.6, ..., 0.95), subsequently averaging these results. By incorporating a broad spectrum of IoU thresholds, this metric becomes more stringent, offering a comprehensive perspective of the model's performance, not only in terms of 'object detection' but also 'object location'.

DOI: 10.54254/2755-2721/30/20230113

4.2. Experiment result analysis

The integrated model has been put through rigorous testing on multiple platforms, including DELL G3 i7-10750+2060, ROG16 i9-12700+3080, etc. The experiment, executed on the MOT-20 dataset, measured precision score, recall score, precision-recall score, and F1 score. The model showed impressive and consistent results, as documented below: Per the confusion matrix and result curve, the gender classification model - utilizing transfer learning - achieved a precision score of 0.94 and a recall score of 0.96. It also accomplished an F1 score of 0.95, which indicates an optimal balance between precision and recall. This outperformed traditional CNN methods which had a precision of 0.91, recall of 0.91, and F1 score of 0.91. It is thus evident that the gender classification model offers high accuracy in gender recognition, with a significantly low rate of false positives.

Referring to the confusion matrix and result curve, the face detection model, based on YOLO, attained a precision value of 0.9 and a recall value of 0.47, and an F1 score of 0.62. This suggests that while the model has high accuracy in face recognition, it shows some deficiencies in predicting false negatives. In other words, the model might mistake objects that are not faces as faces. This limitation is primarily due to the constraints of the dataset, including an insufficient range of face and body images, and the limited time allocated for model optimization. Considering the mAP50 metrics, the IoU (Intersection over Union) threshold is set at 0.5, implying that predictions are correct as long as the overlap between the predicted and actual bounding box is 50% or more. A mAP@.50 score of 0.86 indicates that the model performs impressively under this lenient IoU threshold. The mAP50-95 metrics, which encapsulate a range of IoU thresholds from 0.5 to 0.95, incremented by 0.05, resulted in a mAP@.50-.95 score of 0.504. This signifies moderate model performance under stricter IoU thresholds. To summarize, the gender classification model and the face detection model have both demonstrated strong performance in their respective tasks. The gender classification model showcased superb precision and recall. The face detection model, though effective in terms of precision, showed some limitations in recall. Object detection metrics further corroborate that while the model is adept at identifying the general position of the target, it has room for improvement in delineating target boundaries, particularly at stricter IoU thresholds. Future endeavors should be focused on rectifying these limitations to enhance the face detection model's performance.

5. Conclusion

This groundbreaking study outlines the creation of an integrated system, marrying a YOLO-based people flow tracking system with a face detection and CNN transfer learning-based gender classification system. The aim is to revolutionize people management and facilitate the development of gender-oriented functionalities. Experimental evaluations demonstrated commendable performance of the integrated model. Although the gender classification model exhibited outstanding accuracy and efficiency, the YOLO-based face detection system showcased moderate accuracy, particularly under stringent threshold situations. This highlights potential areas for future development, including the necessity for a more comprehensive dataset encompassing entire human bodies and the observation of human faces from various angles. Further fine-tuning of parameters also emerged as a necessity. The implementation of the system has led to a remarkable reduction in training time costs. While training a YOLO model required approximately four hours, the training of a classifier took a mere ten minutes. This allows for greater time investment in the training of a robust, generalized YOLO model and significantly less time in its adaptation to a specific classification task.

Simultaneously, data collection costs have also been substantially minimized. Changing the task for the classifier now only necessitates the procurement of a dataset for the classifier. The appeal of YOLO lies in its ability to amalgamate three tasks within a single network - Classification, localization, and confidence level determination. The loss function of this network is simply the sum of these three tasks. The network achieves multitasking through minimizing this loss function, albeit without providing any mathematical insight. The extent to which separating the classification task may yield additional benefits

DOI: 10.54254/2755-2721/30/20230113

remains a topic for extensive experimentation and testing. However, such investigations are beyond the scope of the current two-week project and are earmarked for future research endeavors.

References

[1] Amin, M. S., Wang, C., & Jabeen, S. (2022). Fashion sub-categories and attributes prediction model using deep learning. The Visual Computer, 1-14.

[2] Huu, P. N., Tien, D. N., & Thanh, K. N. (2022). Action recognition application using artificial intelligence for smart social surveillance system. J. Inf. Hiding Multim. Signal Process., 13(1), 1-11.

[3] Gündüz, M. Ş., & Işık, G. (2023). A new YOLO-based method for social distancing from real-time videos. Neural Computing and Applications, 1-11.

[4] Sugianto, N., Tjondronegoro, D., Stockdale, R., & Yuwono, E. I. (2021). Privacy-preserving AI-enabled video surveillance for social distancing: Responsible design and deployment for public spaces. Information Technology & People.

[5] Rezaei, M., & Azarmi, M. (2020). Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic. Applied Sciences, 10(21), 7514.

[6] Tsai, J. K., Hsu, C. C., Wang, W. Y., & Huang, S. K. (2020). Deep learning-based real-time multiple-person action recognition system. Sensors, 20(17), 4758.

[7] Genaev, M. A., Komyshev, E. G., Shishkina, O. D., Adonyeva, N. V., Karpova, E. K., Gruntenko, N. E., ... & Afonnikov, D. A. (2022). Classification of fruit flies by gender in images using smartphones and the YOLOv4-tiny neural network. Mathematics, 10(3), 295.

[8] Zhu, X., Xu, H., Zhao, Z., Wang, X., Wei, X., Zhang, Y., & Zuo, J. (2021). An environmental intrusion detection technology based on WiFi. Wireless Personal Communications, 119(2), 1425-1436.

[9] Cojocea, E., Hornea, S., & Rebedea, T. (2019, October). Balancing between centralized vs. edge processing in IoT platforms with applicability in advanced people flow analysis. In 2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet) (pp. 1-6). IEEE.

[10] Ward, T. (2023). Development of Detection and Tracking Systems for Autonomous Vehicles Using Machine Learning (Doctoral dissertation, Morehead State University).