Machine Learning with Applications 14 (2023) 100500

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Real-time AI-based inference of people gender and age in highly crowded environments☆

Jasseur Abidi, Fethi Filali *

Qatar Mobility Innovations Center, Qatar University, P.O Box. 210531, Doha, Qatar

ARTICLE INFO

Keywords:
Intelligent video analytics
Multi-person tracking
Gender and age inference
Crowd analytics

ABSTRACT

Gender identity is one of the most fundamental aspects of life. Automatic gender identification is increasingly being used in areas such as security, marketing, and social robots. The objective of this paper is to address the challenges of gender and age identification in very crowded/noisy environments where faces are unclear and/or people are moving in relatively random directions. It presents an end-to-end real-time intelligent video analytics solution for instant people counting, gender and age estimation in crowded and open environments. The proposed solution includes a complete pipeline for training vision deep learning models and deploying them to edge devices connected to a distributed streaming analytics server. Our final Deep Learning architecture is an extended version of FairMOT, a multi-object tracking model, with two additional layers for multi-class gender classification and age regression. The training phase is performed using an enhanced and enriched version of the CrowdHuman dataset, a public dataset for human detection, with gender and age annotations added. The overall system has been validated for various movies and has shown state-of-the-art performance in terms of people tracking, gender and age inference. Our code, models, and data can be found at https://github.com/jasseur2017/people_gender_age.

1. Introduction

In recent years, intelligent video analytics has attracted increasing interest from industry and academia. It has proven be an essential need for many merchants in shopping malls, watch centers, healthcare facilities. Video analytics is becoming a fully automated activity, from data collection to insight detection, thanks to significant breakthroughs in deep learning. This has led to the development of a number of apps that provide a variety of services across many disciplines.

Some examples of video analytics are video motion detection, fault detection, intrusion detection, line crossing, people counting, gender and age prediction. Each application introduces specific requirements for the quality and timelessness of the video steam and requires different pipelines, from data collection, to model building, to actual deployment.

In this paper we propose an end-to-end real-time solution that counts people and estimates their gender and age from videos captured by 2D cameras in stores, malls, streets, stadiums, etc. In its edge (preferred) version, the inferred information is then sent in real-time to the back-end server for further processing.

Our final Deep Learning architecture is an extended version of FairMOT (Zhang, Wang, Wang, Zeng and Liu, 2021), which is a multi-object tracking model, with two additional layers for gender multi-class categorization and age regression. These layers have been added to the model. During the training phase, an improved and extended version of the CrowdHuman dataset (Shao et al., 0000), which is a publicly available dataset for person detection, is used. This version also includes gender and age annotations. The performance of the system as a whole has been validated for a number of different movies, and it has been proved to be at the cutting edge of technology in terms of people tracking, gender inference, and age estimation.

The proposed solution provides a comprehensive end-to-end real-time intelligent video analytics framework for real-time/online people counting as well as gender and age estimation in crowded and open areas. A complete pipeline for training deep learning visual models and deploying them to edge devices connected to a distributed streaming analytics server is integrated into the proposed system. There are two primary components to the solution:

☆ This work was made possible by NPRP Grant No: NPRP12S-0304-190212 from the Qatar National Research Fund (a member of The Qatar Foundation). Open Access funding provided by the Qatar National Library. The statements made herein are solely the responsibility of the authors.

* Corresponding author.
Linkedin: jasseur-abidi-44bb55138 (J. Abidi), Linkedin: fethifilali (F. Filali).
E-mail addresses: jasseura@qmic.com (J. Abidi), filali@qmic.com (F. Filali).
URLs: http://www.qmic.com (J. Abidi), http://www.qmic.com (F. Filali).

https://doi.org/10.1016/j.mlwa.2023.100500
Received 31 May 2022; Received in revised form 1 August 2023; Accepted 19 September 2023
Available online 26 September 2023

- Perception layer: a combination of cameras installed in different locations and embedded boards such as Nvidia Jetson, Raspberry PI, and Google Coral. These cards integrate the intelligence necessary to process videos captured by connected cameras to detect passengers or customers and extract certain related information such as age and gender. For all detections the system sends the time of detection, relative location, absolute location, gender, and age of detected person to a back-end server asynchronously.
- Analytics layer: a sequence of back-end servers performing three main tasks: big data analytics, log analytics and web visualization. It transforms raw data from perception layer into human-readable information.

The remainder of the paper is organized as follows. Section 2 gives an overview of some of the background aspects and selected related work. The end-to-end multi-task deep learning model is detailed in Section 3. Section 4 provides and analyzes important results obtained from the model. Section 5 concludes the paper and outlines future directions.

2. Background and related works

One of the main enablers considered in this work is the cameras which are the main source of data steams. Many types of cameras may leveraged for people detection depending on each specific use case: 2D vision is applied for simple front views while 2D 360° vision is for overhead views. In general, front views are best suited for outdoor applications while top views are best suited for indoor applications.

Additionally, 2D 360° vision requires a dewarping step in the perception pipeline. The dewarping step consists of dividing each frame of the input video into four undistorted frames representing the view from all four sides around the camera location.

For indoor applications, 3D active stereo vision gives accurate estimation of objects locations without adding too much complexity in the processing algorithm. To accomplish its task at least 3 cameras surrounding the specific location are needed. Cameras instantly generate n images representing the same location from different views. Each image is an RGBD image where D channel represents depth information of each pixel. RGBD image is then transformed into 3D array structure called Point Cloud representing real 3D view of the location. Point clouds can have complementary bleeds. To have a complete view or point cloud of the specific location, these point clouds are merged and then smoothed by some filtering techniques like Voxel filtering. Merging point clouds requires a calibration step which consists of finding the landmark transformations that generate the landmarks of the n cameras from a reference landmark.

For outdoor applications, 3D passive stereo vision gives sufficient estimation of objects locations, but it adds some complexity in the processing algorithm when merging the two RGB images generated by the camera system. 3D passive stereo cameras have low cost compared to 3D active stereo cameras.

Similarly to 2D 360° vision, 3D 360° vision provides a top view of a location but more with some 3D information like depth estimation. One the most common applications of 3D 360° vision is in retail where 2D 360° vision fails to keep up with people who are in a rush and therefore fails to count them. 3D 360° vision overcomes crowd congestion with additional spatial features.

Regardless of whether the applications are indoor or outdoor and the type of vision cameras used, the majority of previous work on gender and age estimation was based on detecting people faces, and they obtained, generally good results with clear face images.

We can cite for example InsightFace (InsightFace, 0000), which is the most famous 2D&3D face analysis library. It implements state-of-the-art algorithms for facial recognition, face detection, face alignment and gender & age estimation (An et al., 2021; Deng, Guo, Liu, Gong and Zafeiriou, 2020; Deng, Guo, Ververas, Kotsia and Zafeiriou, 2020; Deng, Guo, Xue and Zafeiriou, 2019; Deng, Roussos et al., 2019; Guo, Deng, Lattas, & Zafeiriou, 2021; Guo, Deng, Xue, & Zafeiriou, 2018; Somaldo & D., 2021). Despite performing well in clear face datasets like MegaFace, IJB, TrillionPairs, and NIST, it performed poorly in unconstrained face images and videos. Another previous interesting work (Garg, Jain, Kotecha, Goel, & Varadarajan, 2022) has tackled the task of face detection from partial features, but it is not dedicated to small body people classification. In the same category, we can cite (Holkara, Walambe, & Kotecha, 2022) for face recognition task which solved the problem from few-shot learning perspective. Despite its advantages for multi-task learning, it is not directly applicable for people classification task.

We also cite N. Sharma, R. Sharma & N. Jindal in their work (Sharma, Sharma, & Jindal, 2022) Face-Based Age and Gender Estimation Using Improved Convolutional Neural Network Approach, Wenzhi Cao, Vahid Mirjalili, Sebastian Raschka in their work (Cao, Mirjalili, & Raschka, 2020) Rank consistent ordinal regression for neural networks with application to age estimation, Gil Levi, Tal Hassner in their work (Levi & Hassncer, 2015) Age and Gender Classification using Convolutional Neural Networks, Jia-Hong Lee, Yi-Ming Chan, Ting-Yen Chen, and Chu-Song Chen in their work (Lee, Chan, Chen, & Chen, 2018) Joint Estimation of Age and Gender from Unconstrained Face Images using Lightweight Multi-task CNN for Mobile Applications that are based on face aspects. Although some of them are performed on unconstrained face images, they are still not applicable to full-body images where gender and age estimation can use additional features of different body parts.

With CCTV cameras captured persons appear with a small body shape and unclear face, which makes face-based methods unable to achieve acceptable inference accuracy. This is why some of the most recent works use various human aspects such as face, body image, head-shoulder, head, clothes, and 3D body shape. Some of them are based on classical computer vision feature engineering, such as Histogram of Oriented Gradients (HOG) feature extraction and Support Vector Machine (SVM) for face classification (Carletti, Greco, Saggese, & Vento, 2020). Although it solves whole body classification, it shows poor results in crowded environments.

Slightly distant from our work, more complex techniques such as 3D human body shape were used in Tang, Liu, Cheng, and Robinette (2011) to retrieve 3D human data by laser scanning and then applied SVM as a classifier. Fourier Descriptor (FD) method has been applied to location of breast regions and has proven to be robust for 3D imaging applications. Most of the time, surveillance cameras use 2D images and are installed at different angles. Therefore, a method to recognize gender and age from 2D images from all angles can be preferred.

Unlike previously discussed works, our focus in the paper is directed towards full-body person detection with gender and age inference. All of this is done using 2D images to make it more suitable for use with existing surveillance cameras and increase the chances of its adoption by potential end users.

Although a number of techniques like those described in Marathe, Walambe, and Kotecha (2021), Walambe, Marathe, Kotecha, and Ghinea (2021) and Walambe, Marathe, and Kotecha (2022) present good ideas to improve performance of people detection and classification task taking separately, our focus in this paper is focused on single multi-tasks models.

3. Multi-task deep learning model design

Online detection, tracking and classification of people appearing in real-time streaming videos from cameras is a complex task. It is well established that among computer vision methods, only those based on Deep learning can accurately perform the required tasks.

3.1. Modeling phase for multi-task execution

Overall, there are two families of models that could solve the three tasks together: detection, tracking and classification. The first is SDE (Separate Detection Embedding), such as DeepSort (Wojke & Paulus, 2017). The second is JDE (Joint Detection Embedding), such as FairMOT (Zhang, Wang et al., 2021) and ByteTrack (Zhang, Sun et al., 2021). SDE models perform detection and tracking in sequence while JDE models do it in parallel.

3.1.1. Detecting people

For detecting people, two families of models are used to solve detection tasks; anchor-based models such as YoloV3 (Redmon & Farhadi, 2018) and YoloV5 (Ultralytics, 2020) as well as anchor-free models such as CeneterNet (Duan et al., 0000) and YoloX (Zheng, Songtao, Feng, Zeming, & Jian, 2021). Both families perform a pixel-wise prediction of the existence of the bounding box and the regression of its dimensions, but the former use base anchors as a reference in the prediction unlike the latter which make a direct prediction.

Tracking people in video is a well-known challenge because video is a sequence of images that can change quickly depending on the scene. Consecutive frames may share common people. Re-identification of the same person in the frames leads to the correct count. The count will be the result of the detection and tracking steps. The method consists in predicting new states of tracklets (fragments of the track followed by a moving object/person) and in associating new detections with the current tracklets. The status of the tracklets consists of the position of the person and their speed. The prediction of new tracklet states is driven by Kalman filtering. Detections far from a predicted track will be rejected in its association.

We used a modified version of Kalman filtering algorithm to manage the variability of the time difference between consecutive frames. The equation of state is therefore the following:

$$x_{t|s} = x_{s|s} + (t-s).v_{s|s}$$
$$v_{t|s} = v_{s|s}$$

where x is person's position, v is person's velocity, t is current time, and s is last previous time. Another issue that needs to be addressed in the people detection phase is association, which involves associating new detections with already captured tracklets using an optimal cost. The cost is usually a predefined distance. In our case, the association is done in three steps:

- Embedding matching: by applying Linear assignment algorithm based on cosine distance on embedding features of new detections and old tracklets.
- First Intersection Over Union (IOU) matching: by applying Linear assignment algorithm based on IOU distance on boxes of new detections and old tracked tracklets.
- Second IOU matching: by applying Linear assignment algorithm based on IOU distance on boxes of new detections and old lost tracklets.

The above three-step method enables effective tracking of people even in a noisy/dynamic environment.

The combination of detection, tracking and association techniques makes it possible to process the video in real-time frame by frame and to keep the link with them.

3.1.2. Classifying people

Each person has a trajectory of detections and each detection has a prediction related to gender and age. To reduce this multitude of predictions to a single prediction, we apply one of three strategies:

- Voting: by taking most frequent class in the list of predictions.
- Average of predictions: by averaging predictions.

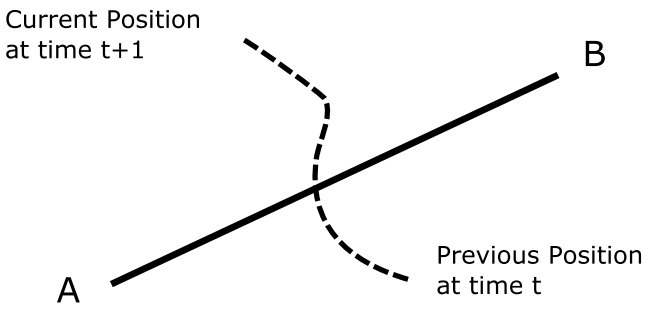

Fig. 1. Illustration of the cross line method used for people counting.

- Weighted average of predictions: by giving weights to the list of predictions based on certain factors such as distance from the person at the time of the prediction to the camera. The near prediction is more important than the far prediction. For instance: $weight = e^{\frac{-distance(person,camera)^2}{v}}$ where v is a regulation hyperparameter.

The first and second method are used in simple scenarios where people appear clearly in the scene of the detections. In our case people appear less clear at least in first frames of detections and gender or age prediction in this case is not reliable, so to overcome this point we give low weights to the unreliable predictions and high weights to the reliable predictions.

3.1.3. Counting algorithm

Anyone who crosses a predefined line towards the entrance or the exit will be counted positively or negatively respectively. Otherwise s/he will be omitted. A virtual cross-line, as shown in Fig. 1, is used to count people in the scene.

The simple counting algorithm works as follow:
if $(det(\vec{AC}, \vec{AB}) < 0)$ and $(det(\vec{AB}, \vec{AD}) \geq 0)$: return negative
elif $(det(\vec{AC}, \vec{AB}) \geq 0)$ and $(det(\vec{AB}, \vec{AD}) < 0)$: return positive
Where (AB) is the predefined line, C is the previous position, and D is the current position.

3.2. Training phase

Training is intended for parametric tasks only. In the previous section, we list three parametric tasks: detection, classification and re-identification that allow integrating the association into the tracking task. Non-parametric tasks like Kalman filtering and IOU association are excluded from training stage.

Thanks to the work of A. Kendall (Cipolla, Gal, & Kendall, 2018), multi-tasking training has become more efficient. We jointly train person detection, person classification and person re-identification together in a model having a common feature encoder and three separate decoders for each task. The multitasking loss is as follows:

$$loss = \frac{1}{2}\left(\frac{1}{e^{w_1}}L_{det} + \frac{1}{e^{w_2}}L_{id} + \frac{1}{e^{w_3}}L_{cls}\right) + w_1 + w_2 + w_3$$

where L_{det} represents the detection loss task, L_{id} represents the re-identification loss task, L_{cls} represents the classification loss task and w_1, w_2 and w_3 represent learnable parameters.

The **first task**, person detection, is a combination of two sub-tasks: person objectness and person bounding box regression. Person objectness is a pixel-wise binary classification problem that uses binary cross entropy criterion as the loss function. Person bounding box regression depends on the nature of the model whether it is anchor-based or anchor-free and generally we use the mean squared error criterion as the loss function.

The **second task**, person re-identification, is a metric learning task that associates the same person detected in consecutive frames with a cluster and different people with different clusters. This task can be solved with two types of loss functions: the first is triplet loss which takes as input two different embedding vectors of same person in two different images and embedding vector of different person. And this minimizes the cosine distance for the same person and maximizes it between two different people. The second type of losses consists in modeling the task as a classification where the label is a unique identifier representing each person individually.

The **third task**, person classification, consists of classifying people by gender and age. This task is solved with cross entropy loss. Despite its continuous nature, "age" can be modeled as a classification problem where the classes are different age ranges. In our case these ranges are "1–2", "3–9", "10–20", "21–25", "26–27", "28–31", "32–36", "37–45", "46–54", "55–65" and "66–116". A disadvantage of this modeling is that mis-predicting "26–27" to "28–31" has the same effect as mis-predicting "26–27" to "66–116". Regression modeling naturally solves this problem. For annotators, it is difficult to determine the exact age of a person from an image. To address this challenge, we added a confidence range of age to the ground truth and we use Gaussian loss where the mean is the age value and the standard deviation is the confidence range.

$$loss = \frac{1}{2}\left(\ln(\max(var, eps)) + \frac{(input - target)^2}{\max(var, eps)}\right)$$

With this multitude of tasks, the data annotation process takes more time. To reduce this time, we can exploit insights from unlabeled data by applying self-supervised and semi-supervised learning methods. Self-supervised training consists of performing a transformation on the input image and predicting a mapping from input image to the new image. Semi-supervised training involves training the model on labeled and unlabeled data. An example of semi-supervised methods, the pseudo-labeling method, which involves training the model on labeled data, generating predictions for unlabeled data, enriching the training dataset with the most confident predictions and to train with new data.

Self-supervised, semi-supervised and multi-task learning are considered regularization techniques and they help reduce overlearning. Moreover, we can exploit the advantages of transfer learning and ensembling techniques to work with less data and regularize our model.

3.3. Dataset

To solve the three interfering tasks: person detection, person re-identification and person classification, we used two families of datasets. The first is a modified version of CrowdHuman (Shao et al., 2018) dataset (Shao et al., 0000). The second is a collection of images capturing a multitude of people with their annotations. Each annotation contains information about the associated person. The information is basically bounding boxes of the head, the visible body and whole body. This dataset only solves person detection task. To extend its usability, we made new annotations: gender and age of some people and conducted training on full bounding box annotation with partial annotation of gender and age. Partial annotation of the CrowdHuman dataset (Shao et al., 0000) is sufficient in terms of cost, performance and results.

The second family of datasets contains those published openly in Milan et al. (0000) namely MOT 2015 (Leal-Taixé, Milan, Reid, Roth, & Schindler, 2015), MOT 2016 (Milan, Leal-Taixé, Reid, Roth, & Schindler, 2016), MOT 2019 (Dendorfer et al., 2019), and MOT 2020 (Dendorfer et al., 2020). Each is a collection of videos featuring a multitude of people in natural scenes. Each person in each video has an associated annotation: bounding box indicating their location with a unique identifier. Redundant information in consecutive frames leads any blind person detection training to model over-fitting. At the same time, this form is very effective for people re-identification task.

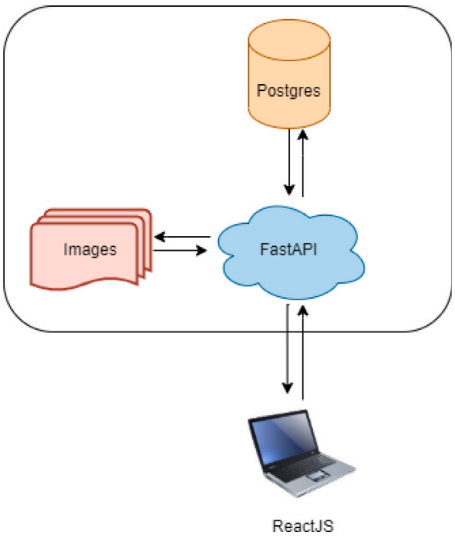

Fig. 2. Annotation toolkit architecture.

After training our Deep learning model on those datasets, we benchmarked our work on a test dataset. Our test dataset is a collection of videos containing people doing real-life activities like walking in street, shopping in malls or stores, etc. Our training detection dataset has 15 000 images. Our validation detection dataset has 4370. The training classification dataset has more than 7600 images with more than 34 000 persons: half men and half women while the validation classification dataset counts has than 2200 images with more than 8000 persons: half men and half women.

Our data annotations are published online at https://github.com/jasseur2017/people_gender_age.

3.4. Annotation toolkit

There are multitude of annotation tools on the Internet. Each of them has its pros and cons in terms of price, formatting, efficiency, work sharing, manual work, etc. To meet our very specific needs, we have implemented a dedicated annotation web toolkit consisting of front-end, back-end, and database as shown in Fig. 2.

The front-end is implemented with React JS. It offers a simple interface for end users to draw boxes and apply class annotations. As shown in Fig. 3, this interface is made up of four components: The first is a list of filenames that we will annotate with a button to load new filenames when the current list is complete. When we click on a filename in the list, its color changes from beige to purple indicating that it is already selected if annotation process is paused. The second component is an image, loaded from the images folder, containing annotations loaded from the database. This image appears with its name and number of boxes above after clicking on a filename of the first component. The bounding boxes have different colors representing the classes. In our case blue is for males, red for females and yellow for unknown gender. Each bounding box is clickable. When clicked, its ground truth annotation classes appear below in radio buttons representing the third component. And its prediction of gender and age appear at right. The prediction division helps the end user to correct erroneous annotations.

The back-end is implemented with FastAPI, a Python3 Rest API framework. It offers some services like querying image names, the image with its associated annotations, saving new annotations or corrections. FastAPI acts as a server. It can handle many queries from different clients at the same time, so multiple users can share the annotation of the same dataset. The database is in PostgreSQL. It contains 3 tables: annotations, predictions, and filenames:

Fig. 3. Annotation toolkit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- annotations: is the main table because it contains ground truth annotations. Each row represents an annotation for a single person in a specified image. This annotation contains image name, box id, head box coordinates, visible body box coordinates, full body box coordinates, gender and age. Gender and age are nullable columns to support empty classification of hard cases.
- predictions: additionally contains model predictions for each box. These predictions are calculated offline and then stored in this table. This table will be suppressed when predicting online from a direct machine learning model.
- filenames: contains all filenames. It helps end user to track annotated and unannotated images. And this allows the back-end to distribute the dataset among end users.

The above-described architecture of the annotation tool has several advantages:

- New users can easily join the annotation process at any time without any code modification or extra configuration.
- Every user can pause the annotation process at any time and easily get back.
- Our system offers radio buttons for classes annotation which is more practical than menu select style.
- Our system offers model classes predictions of every person in each image.
- Our system separates tasks from each other for example it defines object classification given object detection.
- Our system offers an extensible platform for future requirements such as active learning.

In our future plan, we aim to making the annotation toolkit generic and offer it as an open source/free toolkit to the community. We also plan to add more enhancements in terms of functionality, look and feel.

The code of our annotation toolkit is published online at https://github.com/jasseur2017/people_annotator.

3.5. Embedded deep learning

Real-time Video stream processing is a big challenge as it requires a lot of resources to run without time lag. One possible solution is to allocate a dedicated stream channel from camera system to a processing server. But this solution requires highly available connectivity with a reasonable throughput which cannot always be provided by end-users. Another alternative strategy is to process the video in local on-board cards connected directly to the camera. However, embedded boards are known to have limited resources to run complex Deep learning models. That is why we chose FairMOT (Zhang, Wang et al., 2021), a lightweight model, and Nvidia Jetson, a GPU based hardware, to address this use case. Nvidia also offers a multitude of software solutions to run Deep learning based solutions in an optimized way. We can list:

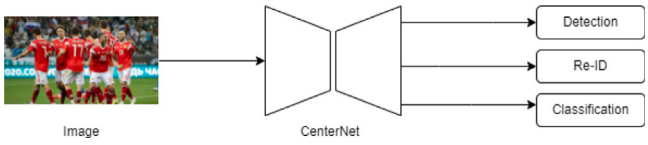

Fig. 4. Architecture of the model.

- TensorRT: a machine learning framework with an optimized API in C++ and Python to run inference models
- DeepStream: an end-to-end platform for intelligent videos processing. It is similar to GStreamer and FFmpeg but most of its code run on GPU instead of pure CPU.
- Triton Inference Server: a server platform for running and scaling trained Deep learning models from any framework. Its advantages such as concurrent model execution and dynamic batch processing are very beneficial for our case of multi-camera tracking and distributed camera solutions.

The system has been designed and implemented to be deployed at the edge or in the cloud depending on the needs of the end-user and the availability or not of always-on connectivity.

4. Experiments and results

We used CenterNet (Duan et al., 0000) as a backbone model. CenterNet is one-stage object detection and it removes the RoI extraction process and directly classifies and regress the candidate anchor boxes. We created a decoupled head for each task: person detection, person re-identification and gender & age classification as illustrated in Fig. 4. Each head of them is a sequence of two convolutional layers with ReLU activation in-between.

Our dataset contains bounding boxes annotations for all people with partial annotations for classes. This is due to the lack of clarity and the small size of some people in the images. Adding extra class "unknown" in the classification task is not a good option as it will add more confusion to the model for uncommon/border cases. So, we modified the classification loss function to backward only when the class exists.

Before starting a complete multitask training and in order to get a baseline for our results we trained and tested our model on each task separately by activating the specified neural network head and freezing the other heads.

- For person detection task, we first trained our model with detection head for 5 epochs with Adam optimizer, learning rate 1e−4 and confidence threshold 0.3 then for 15 epochs with learning rate 1e−5 and confidence threshold 0.5. Model training only takes 20 epochs to reach stable results. The chosen validation metric was MAP (Mean Average Precision).
- For person gender and age classification task, we first trained our model with gender and age classification heads for 20 epochs with Adam optimizer, learning rate 1e−4 and weight_decay 1e−5 then for 10 epochs with learning rate 1e−5 and weight_decay 0. Additionally, we used cross entropy loss with sum reduction as a training loss function for gender classification and KL divergence loss for age training. We changed the cross entropy loss reduction from mean to sum to speed up our training phase. Model training only takes 30 epochs to reach stable results. Validation metric is "accuracy".
- For person re-identification task, we first trained our model with re-identification head for 20 epochs with Adam optimizer, learning rate 1e−4 then for 20 epochs with learning rate 1e−5. The training loss is same as in FairMOT paper (Zhang, Wang et al., 2021). Here we implemented our own validation metric that measures the capacity of our model to make dissimilar vectors of

Table 1
Evaluation results with multitask training and without.

	Single-task training	Multi-task training
People detection (mAP)	69.2%	70.1%
People re-identification (kNN accuracy)	63.6	64.7%
Gender classification (Accuracy)	94.3	95.6%
Age classification (Accuracy)	65.2	65.4%

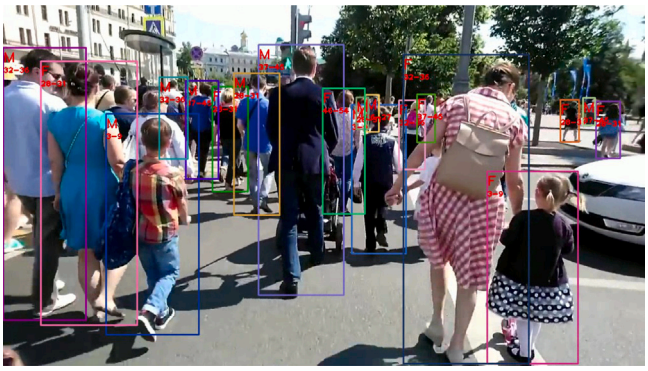

Fig. 5. Example of image processed by our model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

features for different people. So, the metric is simply the accuracy of a k-nearest neighbor algorithm applied on these vectors of features where the ground truth is the people identifiers and the distance between two inputs is the cosine similarity distance.

After making a baseline of results of the separate tasks, we adopted a multitask training strategy as explained in Section 3.2 and we kept the same validation metrics as before. We first trained our model with all heads for 30 epochs with Adam optimizer, learning rate 1e−4, weight_decay 0 and detection confidence threshold 0.3, then for 30 epochs with learning rate 1e−5 and detection confidence threshold 0.5. All results are illustrated in Table 1. Multitask training slightly improves results for gender and age classification and for people re-identification. This is explained by the fact that multi-task training shares features extraction between all tasks and reduces risk of overfitting on one particular task.

After making a baseline of results of the separate tasks, we adopted a multitask training strategy as explained in Section 3.2 and we kept the same validation metrics as before. We first trained our model with all heads for 30 epochs with Adam optimizer, learning rate 1e−4, weight_decay 0 and detection confidence threshold 0.3, then for 30 epochs with learning rate 1e−5 and detection confidence threshold 0.5. All results are illustrated in Table 1. Multitask training slightly improves results for gender and age classification and for people re-identification. This is explained by the fact that multi-task training shares features extraction between all tasks and reduces risk of overfitting on one particular task.

To further improve these results, we had followed a data-centric deep learning approach by annotating more data, especially images containing multiple persons. The information rate obtained from images with several people is greater than images with a single person.

We tested our model on a dozen of videos: some of them are recorded in mall stores and others are downloaded from YouTube. Their lengths varies from 5 min to 30 min. These videos do not have ground truth annotations, so we run our model on them for demo purpose only and not for benchmarking. One example is YouTube video about the 2018 FIFA World Cup Russia. Fig. 5 represents a frame, containing crowd of people, processed with our model. It shows bounding boxes with gender and age classification: M for male and F for female (QMIC, 2023a, 2023b). To deal with the discrepancy between our training dataset and testing videos in term of people shape, we applied in training phase the data augmentation: resizing to small size and padding with black pixels.

5. Conclusion and future works

Video analytics solutions are vital to our daily operations. Numerous industries can profit from this technology, particularly as the complexity of potential applications has increased in recent years. From smart cities to security controls in hospitals and airports to people tracking for retail and shopping malls, video analytics enables operations that are simultaneously more effective, less time-consuming, and less expensive for humans and businesses.

In this paper, we have proposed an end-to-end deep learning based system for people tracking, counting, and classification in crowded/noisy environments. Unlike previously proposed methods for this problem, our system takes advantage of the characteristics of the whole body of the person, and not just the face, to detect gender and age classes. A specific annotation toolkit, which will be made freely available to the community, has been developed to speed-up the process of annotating thousands of images.

As future work, we intend to add more capabilities to the multitask model in order to infer other people-related characteristics, such as whether or not they are wearing glasses/masks, hair color, weight, and height.

These features could be utilized to facilitate tailored marketing campaigns on mall billboards. In addition, we intend to utilize the system during the FIFA World Cup Qatar 2022™, which will be held in Qatar from 21st November to 18th of December 2022, by incorporating the ability to infer the supported Football team/Nation for each detected person/crowd of people, particularly in the multiple fan zones that will be available throughout the country.

CRediT authorship contribution statement

Jasseur Abidi: Writing – original draft, Data curation, Software. **Fethi Filali:** Conceptualization, Supervision, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was made possible by NPRP Grant No.: NPRP12S-0304-190212 from the Qatar National Research Fund (a member of The Qatar Foundation). Open Access funding provided by the Qatar National Library.

References

An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., et al. (2021). Partial FC: Training 10 million identities on a single machine. In *In Proc. of 2021 IEEE/CVF international conference on computer vision workshops (ICCVW)* (pp. 1445–1449). http://dx.doi.org/10.1109/ICCVW54120.2021.00166.

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331. http://dx.doi.org/10.1016/j.patrec.2020.11.008.

Carletti, V., Greco, A., Saggese, A., & Vento, M. (2020). An effective real time gender recognition system for smart cameras. *Journal of Ambient Intelligence and Humanized Computing*, *11*, 2407–2419. http://dx.doi.org/10.1007/s12652-019-01267-5.

Cipolla, R., Gal, Y., & Kendall, A. (2018). A fuzzy variant of the rand index for comparing clustering structures. In *In Proc. 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 7482–7491). Los Alamitos, CA, USA: IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2018.00781.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., et al. (2019). CVPR19 tracking and detection challenge: How crowded can it get? arXiv:1906.04567 [cs] URL: http://arxiv.org/abs/1906.04567.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., et al. (2020). MOT20: A benchmark for multi object tracking in crowded scenes. arXiv:2003.09003 [cs] URL: http://arxiv.org/abs/1906.04567.

Deng, J., Guo, J., Liu, T., Gong, M., & Zafeiriou, S. (2020). Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *In proc. of computer vision – ECCV 2020* (pp. 741–757). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-58621-8_43.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localisation in the wild. In *In proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5202–5211). http://dx.doi.org/10.1109/CVPR42600.2020.00525.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *In Proc. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4685–4694). http://dx.doi.org/10.1109/CVPR.2019.00482.

Deng, J., Roussos, A., Chrysos, G., Evangelos, V., Irene, K., Jie, S., et al. (2019). The menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. *International Journal of Computer Vision*, *127*, 599–624. http://dx.doi.org/10.1007/s11263-018-1134-y.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. CenterNet: Keypoint Triplets for Object Detection.

Garg, D., Jain, P., Kotecha, K., Goel, P., & Varadarajan, V. (2022). An efficient multi-scale anchor box approach to detect partial faces from a video sequence. *BDCC: Big Data and Cognitive Computing*, URL: https://www.mdpi.com/2504-2289/6/1/9.

Guo, J., Deng, J., Lattas, A., & Zafeiriou, S. (2021). Sample and computation redistribution for efficient face detection. *arXiv - CS - Computer Vision and Pattern Recognition*, http://dx.doi.org/10.48550/ARXIV.2105.04714.

Guo, J., Deng, J., Xue, N., & Zafeiriou, S. (2018). Stacked dense U-Nets with dual transformers for robust face alignment. *arXiv - CS - Computer Vision and Pattern Recognition*, http://dx.doi.org/10.48550/arXiv.1812.01936.

Holkara, A., Walambe, R., & Kotecha, K. (2022). Few-shot learning for face recognition in the presence of image discrepancies for limited multi-class datasets. *Image and Vision Computing*, URL: https://www.sciencedirect.com/science/article/abs/pii/S026288562200049X.

InsightFace InsightFace, https://insightface.ai/.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., & Schindler, K. (2015). MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942 [cs] URL: http://arxiv.org/abs/1504.01942.

Lee, J., Chan, Y., Chen, T., & Chen, C. (2018). Joint estimation of age and gender from unconstrained face images using lightweight multi-task CNN for mobile applications. In *In proc. of 2018 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 162–165). http://dx.doi.org/10.1109/MIPR.2018.00036.

Levi, G., & Hassncer, T. (2015). Age and gender classification using convolutional neural networks. In *In proc. of 2015 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 34–42). http://dx.doi.org/10.1109/CVPRW.2015.7301352.

Marathe, A., Walambe, R., & Kotecha, K. (2021). Evaluating the performance of ensemble methods and voting strategies for dense 2D pedestrian detection in the wild. *CVF: Computer Vision Foundation*, URL: https://openaccess.thecvf.com/content/ICCV2021W/ABAW/html/Marathe_Evaluating_the_Performance_of_Ensemble_Methods_and_Voting_Strategies_for_ICCVW_2021_paper.html.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs]URL: http://arxiv.org/abs/1603.00831.

Milan, A., et al. The Multiple Object Tracking (MoT) Benchmark!. https://motchallenge.net/.

QMIC (2023a). Real-time AI-based inference of people gender and age video 1. https://www.youtube.com/watch?v=LqNJ1hD_t0w.

QMIC (2023b). Real-time AI-based inference of people gender and age video 1. https://www.youtube.com/watch?v=TXoM3ssxeBI.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv - CS - Computer Vision and Pattern Recognition*, http://dx.doi.org/10.48550/arXiv.1804.02767.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al. CrowdHuman: A Benchmark for Detecting Human in a Crowd, https://www.crowdhuman.org/.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al. (2018). CrowdHuman: A benchmark for detecting human in a crowd. *arXiv - CS - Computer Vision and Pattern Recognition*, http://dx.doi.org/10.48550/arXiv.1805.00123.

Sharma, N., Sharma, R., & Jindal, N. (2022). Face-based age and gender estimation using improved convolutional neural network approach. *Wireless Personal Communication*, http://dx.doi.org/10.1007/s11277-022-09501-8.

Somaldo, P., & D., C. (2021). Comparison of FairMOT-VGG16 and MCMOT implementation for multi-object tracking and gender detection on mall CCTV. *Journal of Computer Sciences and Information*, *14*, 49–64. http://dx.doi.org/10.21609/jiki.v14i1.958.

Tang, J., Liu, X., Cheng, H., & Robinette, K. M. (2011). Gender recognition using 3-D human body shapes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews*, *41*(6), 898–908. http://dx.doi.org/10.1109/TSMCC.2011.2104950.

Ultralytics (2020). Yolo V5. https://github.com/ultralytics/yolov5.

Walambe, R., Marathe, A., & Kotecha, K. (2022). Multiscale object detection from drone imagery using ensemble transfer learning. *Drones*, URL: https://www.mdpi.com/2504-446X/5/3/66.

Walambe, R., Marathe, A., Kotecha, K., & Ghinea, G. (2021). Lightweight object detection ensemble framework for autonomous vehicles in challenging weather conditions. *Computational Intelligence and Neuroscience*, URL: https://www.hindawi.com/journals/cin/2021/5278820/.

Wojke, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *In proc. of 2017 IEEE international conference on image processing (ICIP)* (pp. 3645–3649). http://dx.doi.org/10.1109/ICIP.2017.8296962.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., et al. (2021). ByteTrack: Multi-object tracking by associating every detection box. *arXiv - CS - Computer Vision and Pattern Recognition (if)*, http://dx.doi.org/10.48550/arXiv.2110.06864.

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, *129*, 3069–3087. http://dx.doi.org/10.1007/s11263-021-01513-4.

Zheng, G., Songtao, L., Feng, W., Zeming, L., & Jian, S. (2021). YOLOX: Exceeding YOLO series in 2021. *arXiv - CS - Computer Vision and Pattern Recognition*, http://dx.doi.org/10.48550/arXiv.2107.08430.

Jasseur Abidi is a Research Associate at Qatar Mobility Innovations Center (QMIC), Doha, Qatar. He has long industrial experience in several domains: computer vision (image, video and 3D data processing), audio analysis and NLP. Jasseur has also good experience in software development including design, debugging, profiling, unit testing, versioning with Git, database, and Docker configuration and deployment.

Dr. Fethi Filali is the Director of Technology and Research at the Qatar Mobility Innovations Center (QMIC). He spearheads the development of innovative and applied technologies within the realms of artificial intelligence, geospatial data analytics, embedded sensing, scalable and distributed algorithms, along with communication systems and protocols.

His inventions have been incorporated into QMIC products across various domains including smart cities, connected and automated mobility, the Internet of Things, intelligent transportation systems, as well as security and defense, culminating in a commercial impact amounting to millions of dollars.

Dr. Filali earned his Ph.D. in Computer Science and Habilitation degrees from the University of Nice Sophia Antipolis, France, in 2002 and 2008, respectively.

Before joining QMIC in 2010, Dr. Filali was a part of the Mobile Communications department at EURECOM, France, serving initially as an Assistant Professor and later as an Associate Professor for a span of 8 years.

He has been the recipient of 25 competitive awards from various funding agencies including the European Commission, The French National Research Agency, and the Qatar National Research Fund. As a Ph.D. Director, he mentored ten Ph.D. students in the fields of intelligent transportation, wireless sensor and mesh networks, vehicular communications, big data analytics, the Internet of Things, and mobility management.

Dr. Filali has co-authored over 145+ research papers for international peer-reviewed conferences and journals, and has (co-)filed more than 10 patent applications.