Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
14/1 (2021), 49-64. DOI: http://dx.doi.org/10.21609/jiki.v14i1.958

Comparison of FairMOT-VGG16 and MCMOT Implementation for Multi-Object Tracking and Gender Detection on Mall CCTV

Pray Somaldo, Dina Chahyati

Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Email:pray.somaldo91@ui.ac.id, dina@cs.ui.ac.id

Abstract

The crowd detection system on CCTV has proven to be useful for retail owners in mall areas where data can be used as a guide by owners to find out the number of visitors who enter at a certain time. However, such information was still insufficient. The necessity for richer data has led to the development of specific person detection which involves gender. Gender detection can provide specific details about men and women visiting a particular location. However, gender detection alone does not provide an identity label for every detection that occurs, so it needs to be combined with a multi-person tracking system. This study compares two methods, method that are tracks person using one-class FairMOT and classify detected person based on gender and method that track multi-class object where the class is male and female at the same time simultaneously called MCMOT which is extended version of FairMOT. The first method produces MOTA, MOTP, IDS, and FPS of 78.56, 79.57, 19, and 24.4, while the second method makes 69.84, 81.94, 147, and 30.5. Besides, evaluation of gender was also carried out where the first method resulted in a gender accuracy of 65% while the second method was 62.35%.

Keywords: *CCTV, Detection, Gender, Tracking*

1. Introduction

Object detection have become common among researchers especially researchers that focused on computer vision. The field also have many real-world applications around industry based on the method. The method used in object detection is always being developed to make it faster and more accurate. This is done so that the method can be used in real scenarios. Object detection also has a role to recognize objects for a long time. However, this still has a problem where objects are often lost from detection. This problem can be solved by involving tracking on object detection. The tracking is useful for filling the gap in between part of lost detections. It is also made object detection more accurate than ever. Tracking have been deployed for computer vision case usually in CCTV. Tracking in video surveillance such as

Closed-Circuit Television (CCTV) has many real-world applications. CCTV can be used for security functions in places that are considered vital so that CCTV is widely used as evidence. Besides, CCTV is also used to record particular case such as incidents or disasters at specific locations to be reported quickly. Furthermore, CCTV can track traffic in some areas and use the information as an announcement to the public. Another function of CCTV that is starting to be widely used is to generate real-time video data. At the same time, CCTV footage can collect data types such as faces, number of vehicles, number of visitors, and types of objects for individual or industrial purposes. plications. Besides, CCTV is also used to record particular case such as incidents or disasters at specific locations to be reported quickly. The data obtained then processed to provide meaningful insights that will impact future decision-making and offer material benefits to the company or industry. Industry engaged in the automotive sector, for example, can use data on the types of vehicles used on the road to see which vehicles are passing so that the industry can produce the type of car needed. Besides, the retail sector can also use data on the number of visitors to find out specific times when visitors are crowded to arrange strategies for selling its products. Thus, a system capable of retrieving data from the CCTV video is needed. Therefore, a person object detection using CCTV can help the industry to retrieve the data in real-time.

However, person data is not informative enough compared to gender. A gender data can benefit many industries, especially in the retail industry, where they can get an insight about which gender is interested in their product. This research aims to tackle that problem using gender detection and tracking specifically for the Indonesian surveillance system attached to a shopping center or retail with a unique gender appearance compared to other countries.

Hence, we want to experiment with real-time multi-object tracking that suitable for CCTV and can do tracking accurately on male and female class using Indonesian gender data so that it can be deployed on Indonesian mall and shopping center.

2. Related Works

In this section, we briefly review related method that utilize gender detection and multi-object tracking. Gender detection method from previous research with different human cues and their limitation is described. Multi-class multi-object tracking is discussed since this research focusing on multi-class object tracking on gender.

2.1. Gender Detection

Gender detection research on the image has become widely known since the rise of deep learning. Some researcher has done gender detection using various human body aspect such as the face, body image, head-shoulder, head, clothes, and 3-D body shape. These aspects have some advantage and disadvantage in specific condition.

Gender detection using face features proposed by [1] with Histogram of Oriented Gradients (HOG) feature extraction and Support Vector Machine (SVM) as face classification. Then, the author tracks the person based on Majority Voting Classification. The system can recognize gender with overall more than 90% accuracy between male and female. This method also achieves real-time performance such that it can run on smart cameras. A similar approach also proposed by [2] using decision mechanism consists of neighboring face detection, context-regions enhancement, and confidence-based weighting assignment. Besides, the method needs the face to appear in the camera, and some will work on specific scenarios. There is also head-shoulder-based method to infer gender proposed by [3] since it includes cues such as hair-style, face, and neckline style. Partial Least Squares (PLS) is used for learn low dimensional discriminative subspace and SVM for gender category classification. This method can classify gender from back of the body. Unfortunately, the system requires much annotated head shoulder data which is hard to find. Moreover, the method which based the looks on the person's clothes using mask segmentation is conducted by [4]. The segmented clothes was cropped into regions and for every region they estimate features that describe color, texture, and shape. They use Mahalanobis distance to measure the feature and obtain final prediction for gender. The disadvantage of this system was the case when man or woman wore similar clothes and also the presence of sunglasses my cause the system to fail.

Furthermore, more complex cues such as 3-D human body shape was used in [5] to retrieve 3-D human data by laser scanning and then applied SVM as classifier. Fourier Descriptor (FD) method was applied on location of breast regions and shown to be robust for 3-D imaging applications. Although this method used robust feature, the constraints of the method was limited data and its difficulty to implement on camera. Mostly, surveillance camera used 2-D images and installed on different angles. Hence, a method to recognize gender from 2-D images from any angles may be preferred. Since using only body image, the implementations is expected to run on real-time inference. Geelen and Dubbelman [6] proposed gender classification in surveillance video using Random Forest SVM where they achieve accuracy around 89.9%. Khryashchev [7] also using similar method where he used Radial Basis Function (RBF) and Local Binary Pattern (LBP) feature instead Random Forest.

Zeni and Jung [8] achieved state-of-the-art result on gender detection with body images using deep neural networks. The author re-annotate PASCAL

VOC 2007 and CelebA data to include gender in the label. The architecture of YoloV2 was used in their method so that the system can work in real-time with 24 Frame Per Second (FPS). They experimented various combination of PASCAL VOC and CelebA which obtain 97.3% accuracy. Nguyen and Park [9] use visible light camera and thermal cameras data with reducing dimension feature Principal Component Analysis (PCA) method and SVM classification. Gender detection with body images also done by [10] using Indonesian body image data. They used Faster-RCNN with VGG16 as convolution layer to extract features. The author achieves 80% average precision between man and woman. This achievement is great considering the amount of data is only 181 images. This result encourages our research to build a gender tracking method on Indonesian gender data.

2.2. FairMOT

The detection and tracking generally uses two stages, namely the person detection and embedding model for association with re-identification. These two stages are carried out sequentially where each stage requires intensive computation so that the system cannot run in real-time.

Referring to this problem, Zhou et al. [11] proposed a method to combine person detection and re-identification so that low-level features can be distributed at both stages simultaneously. This method is proven to be able to run in real-time with competitive accuracy. However, this method has a weakness where the anchor used in object detection refers to multiple identities so that the extracted re-identification feature can come from other objects. This triggers FairMOT's use of anchorless object detection.

FairMOT was state-of-the-art of multi-object tracking with high speed [12]. In motchallenge.net, it achieve 8th rank with MOTA 74.9 (first is 77.6 MOTA) in MOT16 dataset under ”private” detections where the current state-of-the-art on private detector achieved MOTA 77.6 by Lenovo Research company. Even though, the highest Hz (Hertz) or FPS between the first rank until eight rank achieved by FairMOT with 25.4 Hz. Furthermore, FairMOT achieved 8th rank in MOT17 dataset (first is 77.1 MOTA also by Lenovo Research) under ”private” detections where the highest FPS is FairMOT compared with other highest seven method.

CenterNet is used by FairMOT because this method predicts the bounding box length, bounding box width, and keypoint located at the center of the object which then produces a bounding box from that center point. This method is then combined with re-identification on one network in order to save computing [12]. The output of the network is in the form of re-identification detection and embedding. The embedding will be used as a feature to associate detection between frames.

After that, the Hungarian and Kalman Filter algorithms are used to link these features and predict the object's location for the next frame so that each detection has a unique id.

2.3. Multi-Class Multi-Object Tracking

Multi-class multi-object tracking was an extended case for multi-object tracking where the category is more than one. Each object is tracked independently and the id number is generated consecutively based on that object. The number of object class can be expanded to a large number. Moreover, Lee et al. [13] conducted research with unlimited object classes using Changing Point Detection (CPD) algorithm. An ensemble of neural network detector and Lucas-Kanede Tracker (KLT) motion detector was utilized to compute likelihoods of foreground regions as responses of different object classes. This method acquire state-of-the-art results on ImageNet VID and MOT benchmark 2016.

A more significant performance method is suggested by [14] which runs in real-time. The author apply YoloV2 tracker, Kalman Filter and Hungarian algorithm on TUD-Crossing dataset and ETH-Crossing dataset that employ multi-object pedestrian tracking algorithm based on multi-class multi-object tracking. The method achieves competitive accuracy with inference at 17 FPS. These previous research show us that multi-object tracking with multi-class can be done in real-time and the accuracy is still acceptable.

3. Methods

This section describes the dataset used in this paper and the methods used to detect person with/without gender and track them in the subsequent frames. After that, we implement Kalman Filter to predict the position of the target and Hungarian algorithm to assign object in each class. Then, we compare modified method like FairMOT with gender classification and extended version of FairMOT called MCMOT. We also give the detail on evaluation metrics we used.

Table 1. Dataset description

	Scene Type	
Attributes	Scene 1	Scene 2
Images	‘	
Resolution (pixel)	2944 x 1656	1920 x 1080
Number of training images	1025	1015
Number of testing images	300	300
Gender ratio in training data	Man:Woman = 11:6	Man:Woman = 8:9
Gender ratio in testing data	Man:Woman = 8:9	Man:Woman = 13:6

3.1. Dataset

The dataset constructed from videos retrieved from Indonesian shopping center CCTV. The location of this shopping center is kept secret for public use. Video is processed using ffmpeg tools from Ubuntu operation system to extract continuous images from video. After that, we split image for training and testing data where the timestamp is different. We extract the images from video where the number of frame per second is 15 and the number of training datasets is limited, which is around 2000 frames. The reason we use 15 FPS is that due to the limited amount of data, the video duration can runs longer but still maintains continuity between frames consecutively. This is also done so that the objects (people) in the video are more varied in terms of appearance and movement so that the model can recognize various kinds of people and their movement in the image. The result of the processed images is divided into two scene where the details of the images can be seen in Table 1.

3.2. Multi-Class FairMOT

Multi-class multi object tracking method was show that the class of multi-object tracking can be extended and have good result. Hence, this research using extended of FairMOT called MCMOT (Multi-Class Multi Object Tracking). MCMOT will allow us to detect man and woman class at once and track each class with assigned identity. The flow of this method can be seen in Figure 1. After each class is detected, it will used the re-identification embedding of each class to associate the detections based on that class instead associate it in a whole class.

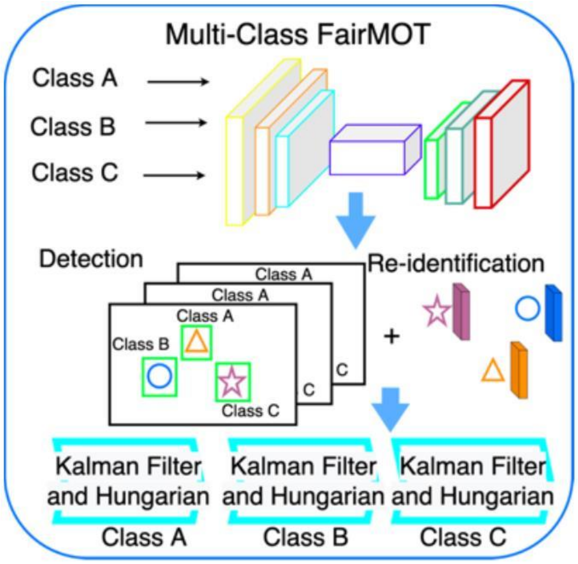

Multi-Class FairMOT
Class A
Class B
Class C
Detection
Re-identification
Class A
Class A
Class A
Class B
Class C
C
C
+
Kalman Filter and Hungarian Class A
Kalman Filter and Hungarian Class B
Kalman Filter and Hungarian Class C

Figure 1. Multi-class FairMOT pipeline

3.2.1. Kalman Filter.

Kalman Filter is used to predict the position of objects for the current frame based on the position of the previous frame. Kalman Filter also functions to associate detection for better results. The position of an object can be expressed as with respect to discrete time on the set of real numbers. Kalman Filter [15] assumes that the position at time t is influenced by the previous position at t−1. Kalman filter consists of two stage, first stage and the second stage. The first stage is to predict the position of target which involve two equations, there are

$$\hat{x}_{t|t-1} = F_t \hat{x}_{t-1|t-1} + B_t u_t \quad (1)$$

$$\hat{P}_{t|t-1} = F_t P_{t-1|t-1} F_t{}^T + Q_t \quad (2)$$

where $x_{t|t-1}$ is projection of state ahead and $P_{t|t-1}$ is projected covariance ahead. The second stage is updated the position based on measuring position between predicted and corrected position. This stage involves two equations specifically

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(z_t - H_t \hat{x}_{t|t-1}) \quad (3)$$

$$\hat{P}_{t|t} = \hat{P}_{t|t-1} - K_t H_t P_{t|t-1} \quad (4)$$

where $\hat{x}_{t|t}$ is updated estimate of state based on measurement and $P_{t|t}$ updated error covariance. The predicted, correct, measurement step can be seen in Figure 2.

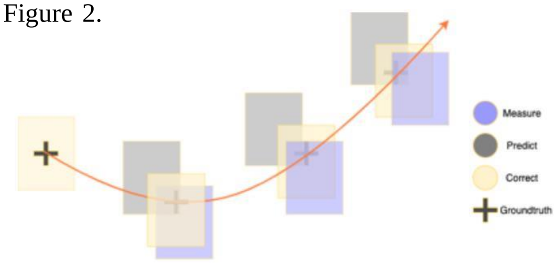


Figure 2. Step of Kalman Filter

Kalman Filter is used because of its simplicity which used linear functions and only requires small computation power so that it can be used for real world applications.

3.2.2. Hungarian Algorithm. Hungarian Algorithm is an algorithm to find the optimal value on an assignment problem. This algorithm can be implemented in tracking people problem by using a cost matrix against the distance values between objects. An example of using this algorithm in association will be explained. For example in frame t − 1 there are detection of people in which each identity namely A, B, and C. After that, in the current frame, namely frame t, there are other detections with different identities, namely K, L, and M. An example of an illustration of detection association can be seen in Figure 3.

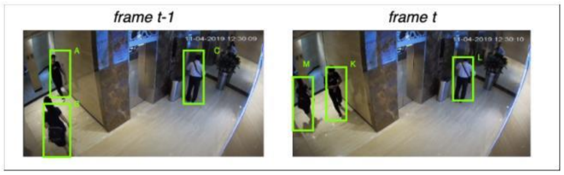


Figure 3. Illustration of detection association

Each distance value for every bounding box will be computed and the distance will be matched using Hungarian algorithm and thus the association of bounding box would be A with K, B with M, and C with L. We used this algorithm because it is the solution for assignment problem which is effective to associate objects given the cost matrix of the object distance. The computation power when reducing matrix also relatively small.

3.3. FairMOT with Gender Classification and MCMOT Comparison

In this section, we want to detect and track gender in real-time for CCTV use case but most multi-object tracking research only available for one-class tracking for person. One-class FairMOT can track person in real-time and accurate but can't detect more than one class. On the other hand, MCMOT which is an extended version of FairMOT can detect more than one class and track it simultaneously. Thus, we combine FairMOT with gender classification stage and compare it with MCMOT that detects and track gender simultaneously so that we can evaluate both method based on accuracy and speed. One class FairMOT will produce detections and labels, then we crop these detections to use it as training data for gender classification. The trained gender classification models will be used to classify crop detections from one class FairMOT. Hence, we get a gender label and draw it on the frame with bounding box and id. On the other hand, MCMOT just use one model to detect object, track it, and predict its gender at the same time. Both methods will be evaluated based on tracking metrics and speed. We also compare the ability to recognize gender and each class such as male and female. This whole method can be seen in Figure 4.

3.4. Evaluation Metrics

The evaluation used to measure multi-object tracking performance by default is put forward by https://www.motchallenge.net page. One of the most frequently used metrics is the Multi-Object Tracking Accuracy (MOTA) which measures the

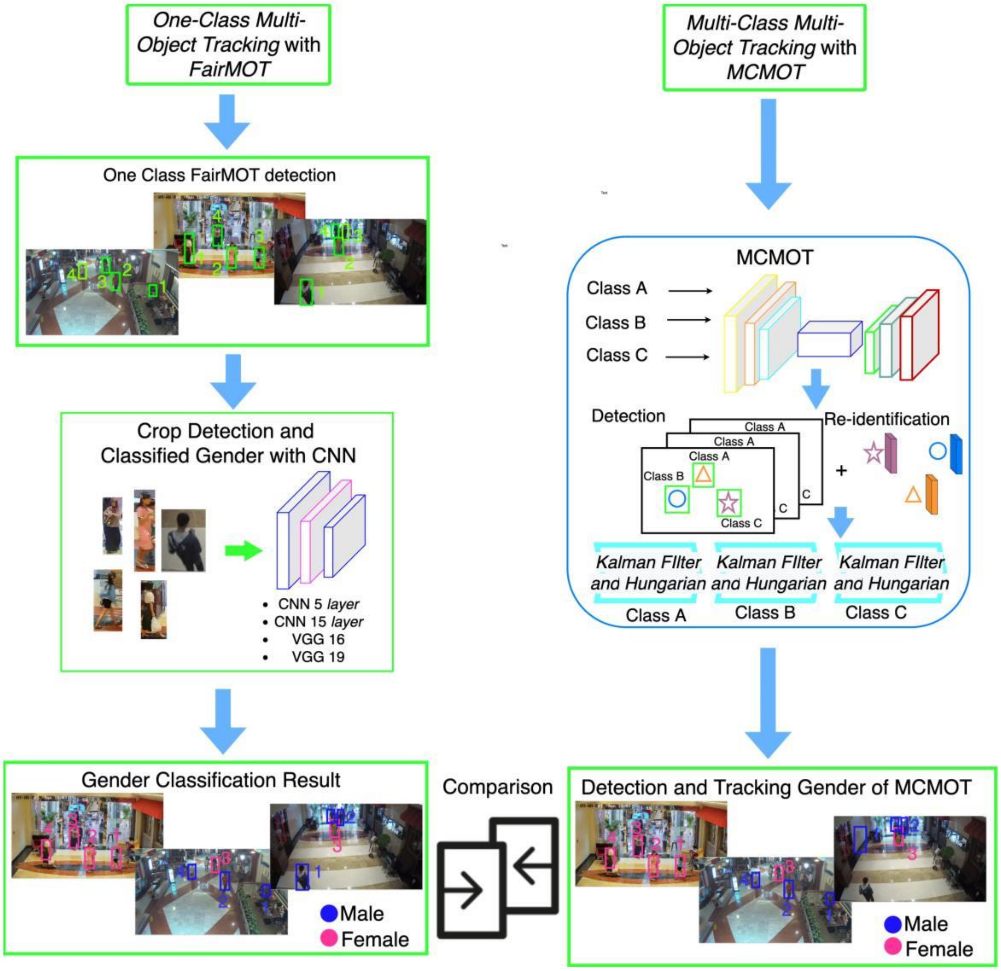

One-Class Multi-Object Tracking with FairMOT
One Class FairMOT detection
Crop Detection and Classified Gender with CNN
• CNN 5 layer
• CNN 15 layer
• VGG 16
• VGG 19
Gender Classification Result
● Male
● Female
Comparison
Multi-Class Multi-Object Tracking with MCMOT
MCMOT
Class A
Class B
Class C
Detection
Re-identification
Kalman FIlter and Hungarian
Class A
Class B
Class C
Detection and Tracking Gender of MCMOT
● Male
● Female

Figure 4. Comparison Scheme Between One Class FairMOT with Gender Classification and MCMOT

tracking performance of the entire image. MOTA can be calculated through the following equation.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t - IDSW_t)}{\sum GT_t} \quad (5)$$

where FN_t, FP_t, and $IDSW_t$ or IDS are the number of false negatives, false positives, and identity switches on frame t while GT_t is the ground truth of tracking. Apart from MOTA, there are other metrics used to measure tracking such as Multi-Object Tracking Precision (MOTP), IDF1, Recall, and Precision which are shown in Equation 6, 7, 8, 9.

$$MOTP = 1 - \frac{\sum_{t,i} d_{t,i}}{\sum_{t,i} c_{t,i}} \quad (6)$$

where $d_{t,i}$ is the overlap bounding box of target i with the ground truth object that has been given and c_t is the number of detections that corresponds to the ground truth.

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (7)$$

where $IDTP$ is all true detected identity or ID, $IDFP$ is the number of false positive ID, and IDFN is the number of false negative ID.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives.

The use of TP, FP, FN, and TN variables has its own meaning in detection in tracking. TP can be interpreted as detected object that is true as a target, FP is an object that is detected but not a target, FN means that there is a target that is not detected, while TN is that there is no detection on a frame that does not contain a target.

Apart from the tracking method, evaluation will also be carried out on the classification method. The metrics used in this method have the same formula for finding recall and precision show in Equation 8 and 9. In addition, there is also an accuracy value to see the model's ability to classify all data correctly. This equation can be written as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

The evaluation metric in the classification method is known as confusion matrix to see the result of prediction of the model. The confusion matrix is arranged based on the TP, TN, FP, and FN values. The positions in order are generally the upper left corner, the lower right corner, the upper right corner and the lower left corner. If there is ground-truth data where the number of classes A and B are 100 each, then the FP and FN will be worth 0 and the confusion matrix looks like in Equation below.

$$ConfusionMatrix = \begin{matrix} [100 & 0] \\ [0 & 100] \end{matrix} \tag{11}$$

4. Experiment and Result

In this section we describe our training schema and pretrain model that we use for both FairMOT one class and MCMOT. Then, we compare the result between one class FairMOT with gender classification and MCMOT multi-class on gender. The detection generated by FairMOT will be cropped and classified into male and female. In contrast, MCMOT will detect and predict gender at the same time. After that, the result between the two is compared.

4.1. Training Scheme of One Class FairMOT and Gender Classification

In this method, there are two training schemes, namely a training scheme with one person class and a training scheme with image classification. The pretrain models used include all dla34, fairmot dla34, and hrnet18. The pretrain model has been trained with several combinations of existing datasets such as CrowdHuman, Caltech Pedestrian, CityPersons, MOT16, MOT 17 and others. All dla34 and fairmot dla34 models use a Deep Layer Aggregation (DLA). backbone with 34 layers while hrnet18 uses an hourglass network backbone with 18 layers. The training scheme for one class can be observed in Table 2.

Table 2. Training scheme for one class FairMOT

Model	Epoch	Batch size	Re-id Dim
alldla34	50	4	128
alldla34	100	4	128
alldla34	150	4	128
alldla34	50	4	256
alldla34	100	4	256
alldla34	150	4	256
fairmot dla34	50	4	128
fairmot dla34	100	4	128
hrnet18	50	4	256
hrnet18	100	4	256
hrnet18	150	4	256

The results obtained are in the form of a tracking model with one class. The model will be selected based on the highest MOTA, then the results of the detection of the training data will be used as training data for image classification.

The scheme used in this method is more varied, such as the use of pretrain models, image size, batch size, differences in the number of layers, and the optimizer. In addition, finetuning was also carried out on the last 2 and 10 layers in the pretrain model. The pretrain model has been provided by the deep learning framework Tensorflow through the Keras library. Apart from using the pretrain model, the author also builds a simple architecture from CNN with several layers for comparison. In this method, there are two training schemes, namely a training scheme with one class namely person and a training scheme with image classification. The training scheme for image classification for gender can be seen in Table 5.

In this scheme, a classification model is built by changing several parameters such as image size, batch-size, and optimizer. The pretrain model used is only VGG16 and VGG19 because the other pretrain models are biased in one class or the results are not as good as the models in the table. In addition, finetune is also done by freezing several layers.

4.1.1. CNN Architecture for Gender Image Classification. CNN is often used in the case of image classification [16]. CNN is also the foundation of object detection where classification is performed on a window or part of the image. This research will use 5 convolution layers and 15 convolutions layers as one of the classification models. The model was built on the grounds that the amount of image classification data is still minimal variation because between frames there are still many of the same people found for a certain duration. In addition, the target classes, namely men and women, have a similar shape. Table 3 and Table 4 is a summary of the CNN architecture that will be used which consists of 5 layers and 15 convolution layers. Table 3 and Table 4 show the detail of CNN used in this research.

Table 3. Architecture of 5 layers convolution

Layer Type	Output Shape	Param #
Conv2D	(None, 100, 100, 64)	1792
Maxpooling 2D	(None, 50, 50, 64)	0
Dropout	(None, 50, 50, 64)	0
Conv2d_1	(None, 50, 50, 32)	18464
Conv2d_2	(None, 50, 50, 32)	9248
Maxpooling2d_1	(None, 25, 25, 32)	0
Conv2d_3	(None, 25, 25, 16)	4624
Conv2d_4	(None, 25, 25, 16)	2320
Maxpooling2d_2	(None, 12, 12, 16)	0
Droput_1	(None, 12, 12, 16)	0
Flatten	(None, 2304)	0
Dense	(None, 128)	295040
Dense_1	(None, 2)	258

Table 4. Architecture of 15 layers convolution

Layer Type	Output Shape	Param #
conv2d	(None, 100, 100, 32)	896
nax_pooling2d	(None, 50, 50, 32)	0
conv2d_1	(None, 50, 50, 64)	18496
conv2d_2	(None, 50, 50, 64)	36928
max_pooling2d_ 1	(None, 25, 25, 64)	0
conv2d_3	(None, 25, 25, 128)	73856
conv2d_4	(None, 25, 25, 128)	147584
conv2d_5	(None, 25, 25, 128)	147584
max_pooling2d_ 2	(None, 12, 12, 128)	0
dropout	(None, 12, 12, 128)	0
conv2d_6	(None, 12, 12, 256)	295168
conv2d_ 7	(None, 12, 12, 256)	590080
conv2d_8	(None, 12, 12, 256)	590080
max_pooling2d_3	(None, 6, 6, 256)	0
conv2d_9	(None, 6, 6, 256)	590080
conv2d_10	(None, 6, 6, 256)	590080
conv2d_11	(None, 6, 6, 256)	598080
max_pooling2d_4	(None, 3, 3, 256)	0
conv2d_12	(None, 3, 3, 512)	1180160
conv2d_13	(None, 3, 3, 512)	2359808
conv2d_ 14	(None, 3, 3, 512)	2359808
max_pooling2d_5	(None, 1, 1, 512)	0
flatten	(None, 512)	0
dense	(None, 1024)	525312
dense_1	(None, 2)	2050

These CNNs architecture is varied on image size between 250 and 100 pixels to make the weight stay small and faster to train so that the model can run faster. The optimizer we used is adam since it is well-known on fast to reach global minimum and better accuracy when training. The details can be seen in Table 5.

4.1.2. Variations of VGG16 and VGG19. Visual Geometry Group or VGG is the CNN architecture proposed by [17]. This model was used in the 2014 Large Scale Visual Recognition Challenge (ILSVRC2014) competition and obtained 92% on the accuracy test with the ImageNet dataset. ImageNet itself consists of 14 million images which have been annotated manually. During its development, the most popular models using this architecture are VGG16 and VGG19. The VGG16 model has 16 layers with 138 million parameters while VGG 19 has 19 layers with 143 million parameters. The VGG16 model makes improvements over AlexNet by replacing large kernel filters (15 layers and 5 layers in the first and second convolutional layers, respectively) with several 3 x 3 kernel sized filters one after another. The VGG19 model has a deeper layer with 3 more convolution layers. The pretrain of these two models has been trained with 1000 classes and is available in the Keras framework.

This research using the VGG16 and VGG19 pretrain model for classification because of its popularity and accessibility on Keras framework. We freeze the layer of the pretrained network on the last two layers and 10 layers on VGG16 because the data is small and contains many same people because of the tracking task and then put a classification network for gender on the last layer. We also varied the epochs where we use small number of epoch on small network like CNN 5 layer and 15 layer. On the other hand, we use big number of epoch on bigger network like VGG16 and VGG19 since we want to see if we can avoid overfitting the model and make better generalizations.

The training scheme of these model and CNN is given in Table 5.

Table 5. Training scheme for gender classification

Model	Epoch	Batch size	Image size	Optimizer
5 layers CNN	30	128	100x100	adam
15 layers CNN	30	32	250x100	adam
VGG16 (finetune 2 layer)	30	128	224x224	adam
VGG16 (finetune 10 layer)	30	128	1224x224	adam
VGG16 dense 1024	60	64	224x224	adam
VGG16 (freeze layer)	30	32	224x224	adam
VGG16 (freeze layer)	50	64	224x224	adam
VGG16 (freeze layer)	70	32	224x224	adam
VGG19 (freeze layer)	200	32	224x224	adam

In this scheme, a classification model is built by changing several parameters such as image size, batch-size, and optimizer. Initially, some models were created the same epoch, namely 30 epochs, after that the epoch variations were carried out on VGG16 and VGG19 because these models had a deeper layer so that more epoch was needed. Other models such as 5 convolution layers and 15 convolution layers are still 30 epochs because there are not so many layers and also to avoid overfitting which can be caused by long training so that the model loses its generalize ability.

In addition, the batch-size is made to vary with values that are multiplies of two, which affects the batch division process in the dataset. The batch-size model with a layer that is slightly larger is made because the GPU is still able to process it. On the other hand, in the larger model the batch-size is made smaller to prevent system crashes. The image size used is the default of the pretrain model whereas the adam optimizer is used because adam combines the best features of AdaGrad and RMSProp to provide an optimization algorithm that can handle sparse gradients on noisy data problems.

4.2. Training Scheme of MCMOT

The scheme used in this method refers more to the use of a different pretrain model downloaded from [18] and [19] where the model has been trained with data contains persons in its target. The pretrain model is then finetuned based on target with male and female classes with data from mall in Indonesia so that the model can adapt to the characteristic and features of mall data in Indonesia. The training scheme is listed in Table 6.

Table 6. Training scheme for MCMOT

Model	Epoch	Batch size	Image size	Re-id Dim
alldla34	50	4	1068x608	128
alldla34	100	4	1068x608	128
hrnet18	50	4	1068x608	128
hrnet18	100	4	1068x608	256

4.3. Result of FairMOT with Gender Classfication

The one class tracking model is then reviewed against predefined evaluation metrics. The results of the one class tracking model on the test data can be seen in Table 7 and Figure 5.

Based on Table 7, the best results are obtained by hrnet18 model. The hrnet18 model has the highest MOTA as well as the lowest IDS. Apart from excelling at both metrics, other metrics such as precision and recall only have small differences with the highest. These results indicate that the hrnet18 model is capable of tracking objects with one class using data malls in Indonesia.

Overall, the model did well on tracking. However, some obstacles were still encountered in tracking one class so that the results were not optimal. These constraints are in the form of ID exchange that affect the MOTA value. This often occurs in multiple frames where the person being detected is standing side by side or walking past. This case can be seen in Figure 6.

Other obstacles that affect precision and recall are false positive and false negative. False positives on tracking of one class occur when an object other than the target is detected, such as a background in an image or an object that resembles a person. Conversely, false negatives occur when there

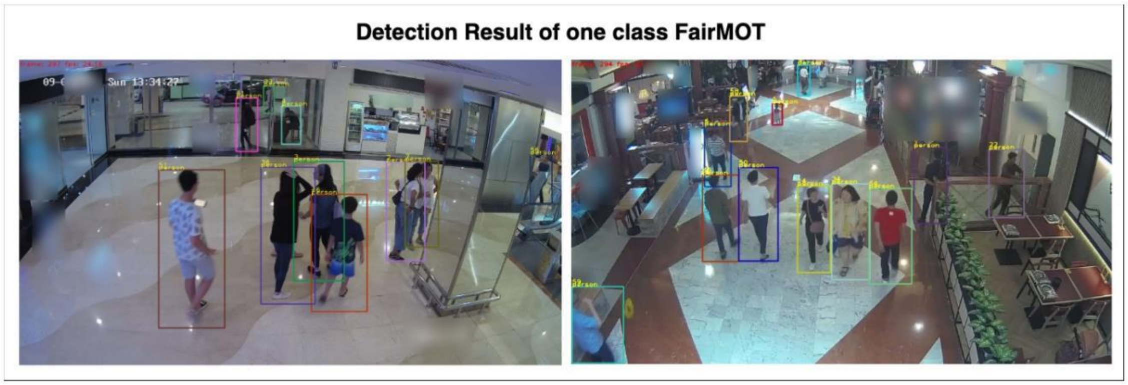


Figure 5. Detection result of one class FairMOT

Table 7. Tracking evaluation result of one class FairMOT

Model	Epoch	Re-id Dim	Batch size	MOTA ↑	MOTP ↑	IDS ↓	IDF1 ↑	Recall ↑	Precision ↑	FPS
person all dla34	50	128	4	76.33	81.17	28	**85.03**	88.24	88.54	29.5 - 32.6
person all dla34	100			76.93	**81.24**	44	80.02	86.85	90.47	29.5 - 32.6
person all dla34	150			76.81	80.71	33	82.91	86.52	90.46	29.5 - 32.6
person all dla34	50	256	4	77.67	81.1	25	78.44	87.36	90.42	29.5 - 32.6
person all dla34	100			78.08	80.9	28	76.12	87.59	90.66	29.5 - 32.6
person all dla34	150			77.92	81.09	24	81.06	86.24	**91.62**	29.5 - 32.6
person fairmot dla34	50	128	4	77.67	81.12	40	83.61	**88.67**	89.6	29.5 - 32.6
person fairmot dla34	100			73.91	80.61	38	82.57	88.46	86.44	29.5 - 31.5
person hrnet18	50	256	4	75.93	79.24	63	71.89	86.44	90.18	23.8 - 24.4
person hrnet18	100			78.38	79.5	33	81.7	89.24	90.48	23.8 - 24.4
person hrnet18	150			**78.56**	79.57	**19**	80.1	86.9	91.56	23.8 - 24.4

are objects that are target but not detected. The examplecan be observe in Figure 7 and Figure 8.

After the one class model is obtained, the detection results of the model will be forwarded to the gender classification model. The gender classification model that was trained was then evaluated with test data. The test data used previously has been cropped and grouped by class. The results of the evaluation on this test data will show the accuracy of the model in classifying classes between men and women. Evaluation can be seen in Table 8.

The results in Table 8 shows that the accuracy, precision, and recall showed quite low. This is also seen by the number of male data crop that is more than that of women. The impact of bigger epoch on VGG16 (freeze layer) did not always improve accuracy. Freeze layer tend to give better result than other model because it was trained on data that contains people in Imagenet. Some models produce detection in only one of the classes as

Table 8. Gender evaluation result of one class FairMOT

Model	Epoch	Accuracy	Precision	Recall	Confusion Matrix	FPS
5 layers CNN	30	0.61	0.61	0.61	[[1554 1448] [790 1875]]	187.33
15 layers CNN	30	0.54	0.53	0.53	[[1997 1005] [1618 1047]]	399.07
VGG16 (finetune 2 layer)	30	0.47	0.24	0.5	[[0 3002] [0 2665]]	189.72
VGG16 (finetune 10 layer)	30	0.53	0.26	0.5	[[3002 0] [2665 0]]	38.16
VGG16 dense 1024	60	0.6	0.59	0.59	[[1991 1011] [1265 1400]]	166.19
VGG16 (freeze layer)	30	0.62	0.63	0.63	[[1640 1362] [788 1877]]	189.12
VGG16 (freeze layer)	50	0.54	0.54	0.54	[[1974 1028] [1533 1132]]	150.11
VGG16 (freeze layer)	70	**0.65**	**0.59**	**0.59**	**[[1964 1038] [924 1741]]**	143.81
VGG16 (freeze layer)	200	0.61	0.61	0.58	[[2552 450] [1863 802]]	143.81

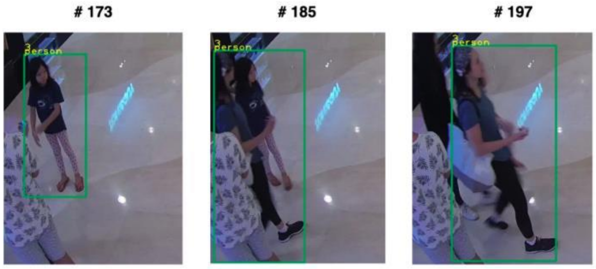


Figure 6. ID exchange between person

given by the VGG16 model with finetune for 2 layers and 10 layers where in the configuration matrix the true positive value is 0 in the finetune against 2 layers and true negative is 0 on the finetune for 10 layers. In total, the best gender classification results were obtained by the VGG16 model with 70 epochs. The accuracy, precision, and recall value of the model is 0.65.

4.4. Result of MCMOT

After training was done, the multi-class model will be used to evaluate the test data to analyze the accuracy of the model in tracking. The evaluation result of the test data is in Table 9 and an example of the MCMOT tracking results can be seen in Figure 9.

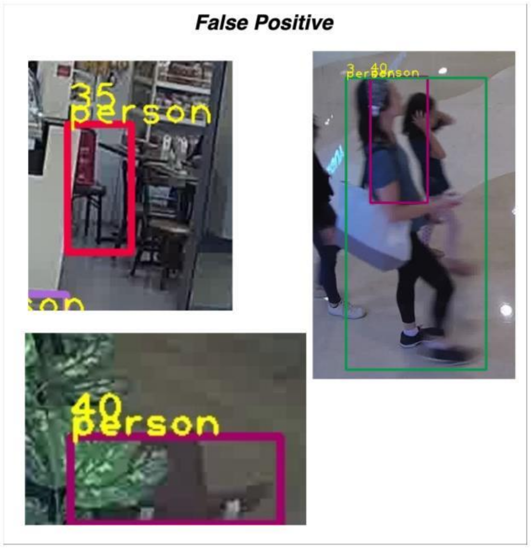
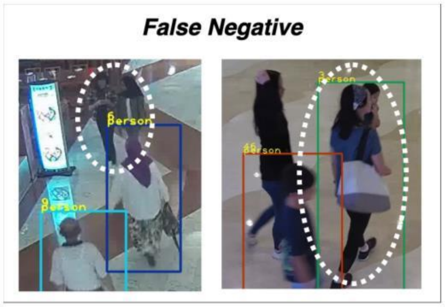


Figure 7. False positive of one class FairMOT

Table 9 show s that the two models namely all dla34 and hrnet18 were trained on 100 epochs. Each pretrain model has its own advantages across several metrics. Based on observations on the pretrain model

Table 9. Tracking result of MCMOT on test data

Model	Epoch	MOTA ↑	MOTP ↑	IDS ↓	IDF1 ↑	Recall ↑	Precision ↑	FPS
all dla34	50	69.59	81.91	169	65.42	84.93	87.3	28.9 - 30.5
all dla34	100	69.84	81.94	147	60.36	84.93	87.17	28.9 - 30.5
hrnet18	50	68.36	79.77	238	56.88	83.58	88.34	28.9 - 30.5
hrnet18	100	65.34	79.58	261	52.70	83.85	85.77	28.9 - 30.5



Figgure 8. False negative of one class FairMOT

all dla34, MOTA, MOTP and IDS increased from epoch 50 to 100 but not too significant. On the other hand, MOTA, MOTP and IDS actually experienced a decline using the hrnet18 pretrain model. The overall metric results show that all dla34 excels in gender tracking on four metrics where the other two metrics namely IDF1 and precision have only a small difference with the highest scores. Table 9 also shows the high IDS value for each model. This means that IDS or object ID changes are often found. Some examples of ID changes can be seen in Figure 10.

In addition, there are false positives and false negatives that cause the MOTA value to be not maximal. False positives occur when there is a bounding box on an object that should not be detected while false negative occurs when an object is targeted but does not have a bounding box. A large number of false positives and false negatives will result in a smaller MOTA value. An example of an image that shows a false positive and a false negative of MCMOT result can be seen in Figure 11 and Figure 12.

After that, the model was also evaluated based on its ability to detect male and female classes. The measurement used is mAP because this method is an object detection problem that focuses on the accuracy of the bounding box position and the appropriate class. The mAP per class is also given in Table 10 to provide a more precise description of each class.

Based on the Table 10, gender accuracy, gender precision, and gender recall give results that have small different from each model. These three metrics are calculated based on the number of male and female detections of each model.

The highest gender accuracy and precision is obtained by hrnet18 with 50 epochs, while gender recall is obtained by all dla34 model with 100 epochs. These results indicate a large number of epochs has no effect on hrnet18. In addition, the highest mAP values for men and mAP for overall were obtained by all dla34 where hrnet18 with 50 epochs had the highest mAP for female. Overall, the best evaluated pretrain model obtained hrnet18 with 50 epochs.

Based on mAP model, it can be seen that the mAP value for men tends to be greater than the mAP value for women. This result is quite reasonable because the ratio of the number of men is greater than women in the training data in Table 1 so that the model is better able to recognize the color and pixel features of the male. However, the average mAP value obtained from the four models is not optimal. This is because men and women have similar appearance and features so that models tend to have difficulty distinguishing between them. Some case can be seen in Figure 13.

4.5. Comparison Result between FairMOT with Gender Classification and MCMOT

After getting the results of each method, the advantages of each method will be shown. These advantages can be seen in Table 11.

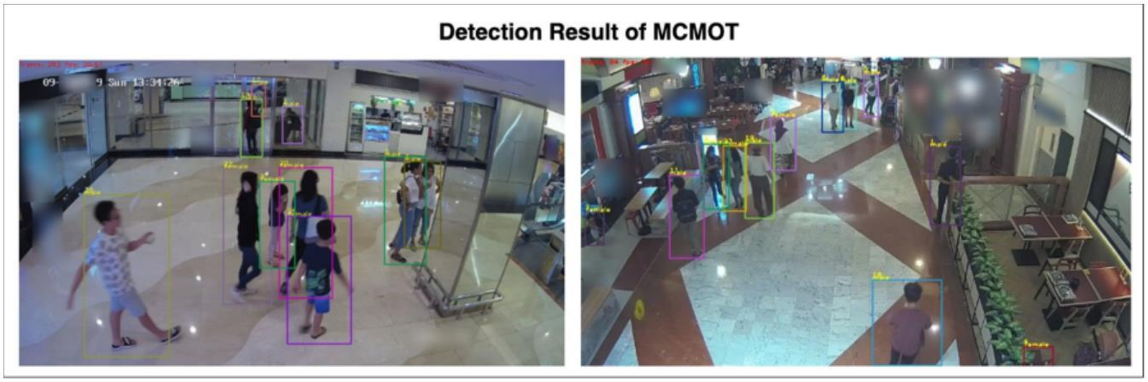


Figure 9. Detection result of MCMOT

Table 10. Gender evaluation result of MCMOT on test data

Model	Epoch	Gender Accuracy	Gender Precision	Gender Recall	mAP Male	mAP Female	mAP
all dla34	50	60.08	62.67	**70**	56.21	28.16	42.19
all dla34	100	61.49	64.38	69.72	**57.04**	31.12	**44.08**
hrnet18	50	**62.35**	**66.21**	63.75	50.09	**34.64**	42.37
hrnet18	100	60.42	64.57	62	51.57	32.57	42.07

Table 11. Tracking comparison between one class FairMOT and MCMOT

Method	Model	MOTA ↑	MOTP ↑	IDS ↓	IDF1 ↑	Recall ↑	Precision ↑	FPS
One Class FairMOT with classification	hrnet18	**78.56**	79.57	**19**	**80.1**	**86.9**	**91.56**	20.42 - 20.86
MCMOT	all dla34	69.84	**81.94**	147	60.36	84.93	87.17	**28.9 - 30.5**

The best model of each method is compared with the respective FPS. Based on Table 11, FairMOT with classification gender excels in many metrics when it comes to tracking except MOTP metric. In contrast, MCMOT excels in MOTP and speed (FPS). The highest MOTA in FairMOT with classification is due to the difference in the number of classes being tracked. Method one class FairMOT tracks only one class whereas MCMOT tracks two classes. This is quite difficult to do in MCMOT where more objects are tracked but the number of frames in the dataset remains the same. Another obstacle is that male and female objects have similar shapes and characteristics.

In addition, the IDF1 value in FairMOT with classification of 80.1 has a significant difference from MCMOT which has a value of 60.36. This happens because IDF1 is affected by a large number of IDFPs (false positives of IDs) and IDFNs (false negatives of IDs). IDFP shows the number of IDs contained in objects that are not targets while IDFN shows that there are IDs on the undetected target. This difference is also caused because method MCMOT tracks two similar classes, namely men

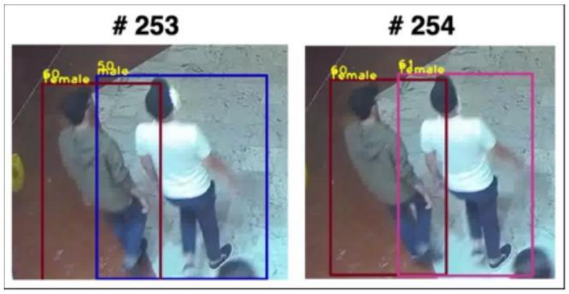


Figure 10. IDS changes in MCMOT

and women, while method one class FairMOT with classification only has one class. This also affects the high IDS value in MCMOT where many IDs are swapped or switched between objects. The recall and precision values also lower but not significant compared to method B due to the addition of classes while the amount of data used between one class FairMOT with classification and MCMOT is the same.

Another metric that can be compared is FPS. In general, the FPS value for standard video is 30. This means that MCMOT can process standard video of 10 seconds in length 289 - 305 seconds. Meanwhile, the combination of FairMOT and gender classification can process the video in 238 - 244 seconds so that many frames will not be processed.

Another evaluation that needs to be considered is the ability of the model to predict gender. The results of the comparison of gender evaluations can be seen in Table 12.

In Table 12., we compare the ability to determine gender between MCMOT and gender classification. MCMOT still needs further development in differentiating sexes because it shows a low predictive rate for both men and women. This is natural because MCMOT performs two tasks at once, namely detection and tracking. In addition, detection is carried out on objects that have similar characters and shapes so that the task is more suitable for gender classification. The gender classification in one class FairMOT is better than MCMOT but further research is still needed to increase the prediction rate because the accuracy is still around 65%.

Overall, FairMOT with classification excels in tracking because only one object is being tracked. Even though MCMOT is an extension of FairMOT, tracking metrics are not superior because the accuracy drops when the class is split in half. In addition, the male and female object classes in MCMOT have similar shapes and features, making it difficult to distinguish them. However, method MCMOT is still superior in its speed of processing frames quickly.

6. Conclusion

The problem of multi-object tracking is a topic that is still being researched today because of its wide application. One of the state-of-the-art multi-object tracking methods recently is FairMOT which provides accurate results but can run in real-time. The existing model gives quite good results in several scenes but is still not good enough when using mall data in Indonesia. This research was then extended to use two classes, namely men and women to enrich information. MCMOT which is an extension of FairMOT by using more than one class is used. Then, the two methods were compared between MCMOT on two classes with one class FairMOT with gender classification.

The advantages of MCMOT are that it can run faster and only use one model so that it is more efficient but MOTA, MOTP, and IDS are smaller than one class FairMOT. In contrast, the FairMOT method with one class and gender classification excels on many tracking metrics but does not run in real-time. This can be a consideration if it will be used in real applications in the real world. If speed is needed and tracking accuracy is tolerable, then MCMOT is the right choice. Otherwise, if tracking accuracy is important then FairMOT with gender classification can be implemented.

In addition, gender classification in one class FairMOT provides a balanced accuracy of about 65.32 and 65.42 for males and females respectively where the accuracy of method MCMOT is only about 50.09 for males and 34.64 for females. Implementation of these results in malls in Jakarta can be done by looking at the largest number of genders in an ID and associating that gender with the corresponding ID.

The implementation of these two methods can be used to obtain information from CCTV in the form of the number of visitors based on gender. This can be done by using a digital line placed on the frame and calculating each center point of the bounding box that intersects the line. At the same time, the class in the form of gender and id from the bounding box can be retrieved.

The suggestion from this research is that it is necessary to add data from various scenes in several malls in Indonesia. This will increase the variety of data and increase the gender data so that the model is more accurate. In addition, further research is needed to find a more appropriate method to differentiate the image character of the whole body between men and women because the characters of the images are

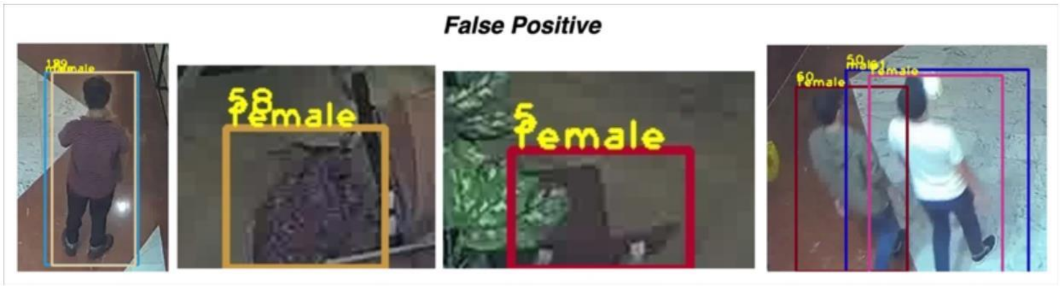


Figure 11. False positive of MCMOT

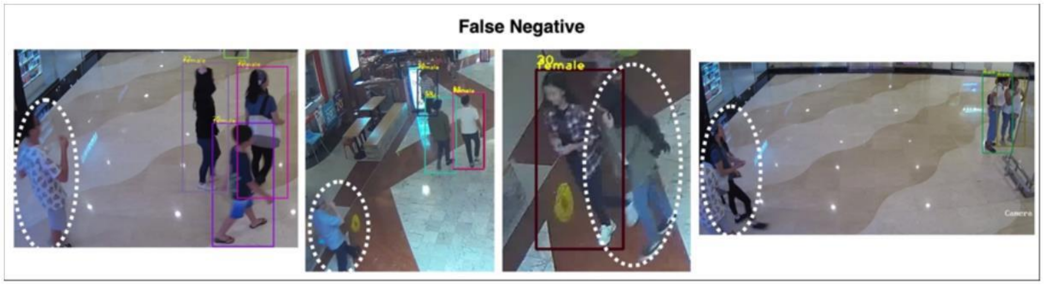


Figure 12. False negative of MCMOT

Table 12. Gender comparison between one class FairMOT and MCMOT

Method	Model	Gender Accuracy	Gender Precision	Gender Recall	Male	Female
Gender Classification One Class FairMOT	VGG16 (freeze layer)	**65**	65	**65**	**65.32**	**65.42**
MCMOT	hrnet18	62.35	**66.21**	63.75	50.09	34.64

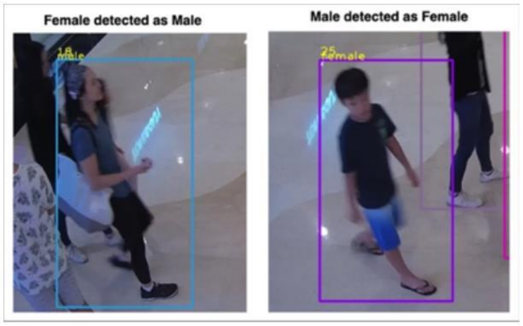


Figure 13. False predict of Gender

so similar that the existing models are not able to distinguish them.

Acknowledgement

This work is supported by Publikasi Terindeks Internasional (PUTI) Prosiding Grant 2020 from Universitas Indonesia.

References

[1] V. Carletti, A. Greco, and M. Saggese, Alessia Vento, “An effective real time gender recognition system for smart cameras,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 6, pp. 2407–2409, 2020.

[2] D.-Y. Chen and K.-Y. Lin, “Robust gender recognition for real-time surveillance system,” in IEEE International Conference on Multimedia and Expo, 2010, pp. 191–196.

[3] M. Li, S. Bao, W. Dong, Y. Wang, and Z. Su, “Head-shoulder based gender recognition,” in 2013 IEEE International Conference on Image Processing, 2013, pp. 2753–2756.

[4] S. Cai, J. Wang, and L. Quan, “How fashion talks: clothing-region-based gender recognition,” in Iberoamerican Congress on Pattern Recognition, 2014, pp. 515–523.

[5] J. Tang, X. Liu, H. Cheng, and K. M. Robinette, “Gender recognition using 3-d human body shapes,” vol. 41, no. 6, pp. 898–908, 2011.

[6] C.D. Geelen, R. GJ Wijnhoven, and G. Dubbelman. “Gender classification low-resolution surveillance video: in-depth comparison of random forests and SVMs.” *Video Surveillance and Transportation Imaging Applications 2015*, vol. 9407, p. 94070M. International Society for Optics and Photonics, 2015.

[7] V. Khryashchev, A. Priorov, and A. Ganin. "Gender and age recognition for video analytics solution." In *2014 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1-6. IEEE, 2014.

[8] d. A. Zeni, L. Felipe, and C. R. Jung, “Real-time gender detection in the wild using deep neural networks,” in 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018, pp. 118–125.

[9] D. T. Nguyen and K. R. Park, “Body-based gender recognition using images from visible and thermal cameras,” vol. 16, no. 2, p. 156, 2016.

[10] D. Chahyati, M. I. Fanany, and A. M. Arymurthy, “Man woman detection in surveillance images,” in 2017 5th International Conference on Information and Communication Technology (ICoIC7), 2017, pp. 1–4.

[11] X. Zhou, D. Wang, and P. Kra¨henbu¨hl, “Objects as points,” arXiv preprint arXiv:1904.07850, 2019.

[12] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, “Fairmot: On the fairness of detection and re-identification in multiple object tracking,” arXiv e-prints, pp.arXiv-2004, 2020.

[13] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, “Multi-class multi-object tracking using changing point detection,” in European Conference on Computer Vision.

[14] K. Jo, J. Im, J. Kim, and D.-S. Kim, “A real-time multi-class multi-object tracker using yolov2,” in 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017, pp. 507–511.

[15] R. Faragher, “Understanding the basis of the kalman filter via a simple and intuitive derivation [lecture notes],” vol. 29, no. 5, pp. 128–132, 2012.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in Proceedings of the IEEE, vol. 86, no. 11, 1998, pp. 2278–2324.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[18] Y. Zhang. (2020, Nov.) Fairmot. [Online]. Available: http://www.github.com/ifzhang/FairMOT

[19] C. Even. (2020, Nov.) Mcmot. [Online]. Available: http://www.github.com/CaptainEven/MCMOT