Deep Learning for People Counting in Videos by Age and Gender

Andrei de Souza Inácio, Rafael Hora Ramos, Heitor Silvério Lopes
Bioinformatics and Computational Intelligence Laboratory – LABIC/CPGEI
Federal University of Technology Parana (UTFPR)
Emails: {andrei.inacioo, ramosrafh, hslopes}@gmail.com
hslopes@utfpr.edu.br

***Abstract*—Currently, many companies or even cities use surveillance cameras all the time, and due to the COVID-19 pandemic, many places have to limit the number of people in attendance. This paper proposes a method for people counting by gender and age in videos using deep learning techniques. The proposed method is based on a face detection and tracking approach combined with an alignment process to minimize the negative effect of the background information, considering occlusions and avoiding duplicate counting. Then, specialized Deep Neural Networks based on the EfficientNet architecture are employed for age and gender classification. Experimental results show that our method achieves satisfactory performance on people counting by age and gender, demonstrating the effectiveness of the present method.**

***Index Terms*—People Detection, Deep Learning, Neural Networks**

I. Introduction

Surveillance cameras have become popular nowadays, and they are used for several purposes, including monitoring human behavior, detecting abnormal events, and, in some cases, helping to identify criminals.

The use of cameras for counting and tracking people in public and private areas has also emerged as an important issue as the global population increases, specially in the large cities around the world [1]. In particular, counting people can be an important means to gather information for the management of restaurants, supermarkets and retail shops. Such information can help managers to analyze customer behaviour, enabling better physical organization and efficient target marketing [2]. In addition to business purposes, people counting systems can also be useful for health purposes. For example, they allow to count and limit the number of people inside public or private areas. This is an indirect way for combating the spread of the COVID-19 disease [3].

Although somewhat similar to crowd density estimation, the people counting task has different methods, and its peculiarities make it is more complicated [4]. For instance, indoor scenes may have moving people that can appear more than once in a scene. Also, occlusions from objects and other people can make the counting task even more challenging.

To perform people counting in low-to-medium density crowds, considering occluded objects and different camera orientations, [1] proposed a seven-module framework by using computer vision and digital image processing techniques, such as binarization, multivalued threshold, Lucas-Kanade, and Self-Organized Map based clustering method.

Focusing on low-cost real-time people counting, [2] proposed a method to estimate the number of people entering and exiting a specific area. Firstly, the foreground was extracted out from the background. Then, a motion detection method was applied, and a multivariate linear regression method was used to count people.

Deep Neural Networks (DNN) have achieved excellent results in solving several complex computer vision problems, for instance: object detection [5], object segmentation [6], anomaly detection in videos [7], people reidentification [8], gender and age recognition [9]. Following this idea, [10] proposed a deep learning model based on convolutional layers followed by LSTM layers to extract spatial and temporal features from videos. Then, the proposed architecture learned a non-linear regression model to predict people count in videos.

Most of the previous approaches addressed for the people counting task focused only on identifying the number of people in the video scene, not taking into account if occlusions happen or a person moves fast through the scene resulting in double-counting. Of course, to devise if a person has appeared in a previous video frame, it is necessary, some minimal identification of people. To avoid the complexity of face re-identification, reporting gender and age information may be qualitative way to count people in many real-world scenarios.

This work presents a method for counting people in videos taking into account gender and age information. It uses a specialized DNN to extract gender and age information from detected faces in videos. A novel alignment method is also proposed for partial person re-identification (constrained to the current video scene), considering occlusions and avoiding duplicate counting.

This paper is structured as follows. Section II describes the proposed methods. Section III shows the experimental results and discussion. Section IV presents the conclusion and points out future research directions.

II. Methods

A. Dataset

For people counting in videos, we create a new dataset containing 70 videos. In these videos, 98 different people are found, performing several different daily actions, including

walking, talking, cooking, playing, taking off clothes, and reading a book. All videos were collected from three publicly available datasets: MSVD [11], MSR-vtt [12], and Charades [13]. The length of the videos range from 4 to 38 seconds, with an average of 15 seconds. Each video was manually annotated with gender and age group information for all people detected. The dataset will be put on public domain so as to foster further research in this area, and allow future comparisons with other approaches. Figure 1 presents the gender and age group distribution of the proposed dataset.

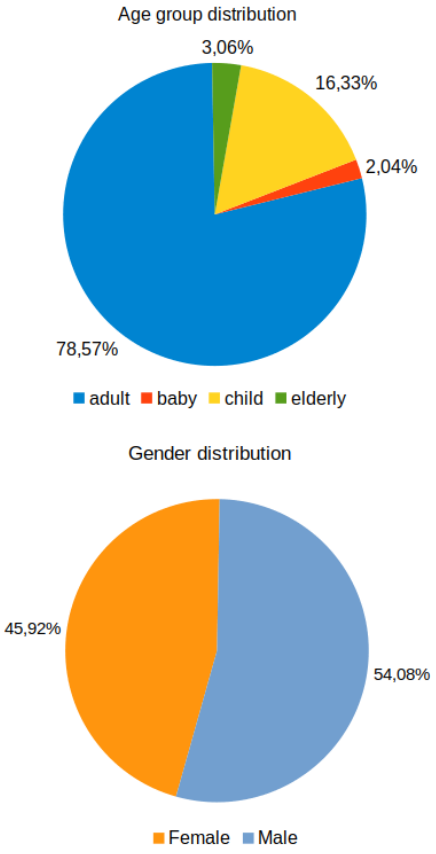

Figure 1. Gender and age group distributions of the proposed dataset.

B. The proposed method

The proposed method, presented in Figure 2, consists of four main blocks: face detection, people identification, age group and gender classification, and people counting. Each block is explained in the following sections.

1) Face Detection: This module aims at detecting people's faces in video frames. That is, given an input frame, it detects all faces in the frame and outputs the cropped faces, as shown in Figure 2.

Face detection can be considered a particular type of object detection task in computer vision [14]. Thus, we use the Tiny-Yolo-v2 [15] DNN for this purpose. It is a simpler version of the well-known YOLO DNN, and uses 9 convolution layers and 6 pooling layers, instead of the 24 convolution layers in the traditional YOLO DNN. The model was pre-trained on the FDDB dataset [16]. The pre-trained weights were obtained from the YoloKerasFaceDetection project [1].

2) People Identification: Counting people in videos is quite different from counting them in static images. The former is much more challenging because people are often moving. Therefore, it is necessary to identify and label every person at all frames. This fact rises the problem of avoiding multiple detections of the same person. A simple Intersection Over Union (IOU) method may not achieve satisfactory results due to the speed of the movement variation and the occlusions of the object of interest. Figure 2 shows that the people identification step actually comprises two procedures: Alignment and Similarity.

Given two cropped image faces, detected at subsequent time frames, f_t and f_{t-1}, the alignment method is used to minimize the negative effect of the background information in the comparison of the faces. Everything in the cropped image that is not face, is the background, and it is considered noise, since it can affect the face similarity checking. For tackling this issue, a HOG (Histogram of Oriented Gradients) feature extractor along with a SVM (Support-Vector Machines) classifier is applied on the face image to obtain an improved bounding box. Then, the face landmarks are detected using the Dlib library [2] landmark predictor, followed by an Affine Transformation taking as reference the eyes and bottom lips to preserve those landmark positions on every image. The final aligned face is returned in a 96×96 pixels image. Later, features from the aligned faces are extracted using the nn4.small2.v1 OpenFace model [3], pre-trained with a combination of the two datasets FaceScrub [17] and CASIA-WebFace [18], resulting in a 128-dimensional embedding vector for each face image.

Finally, the similarity between these two embedding vectors representing two faces in consecutive frames is calculated according to Equation 1:

$$sim(vet1, vet2) = \frac{vet1 \cdot vet2}{\sqrt{vet1 \cdot vet1}\sqrt{vet2 \cdot vet2}}, \quad (1)$$

If the similarity calculated is higher than a predefined threshold θ, both face images are considered from the same person. In this work, we empirically set the θ to 0.85. On the other hand, if the similarity is smaller than θ, the Euclidean distance between the centroids is considered to capture sudden angle changes or other changes that drastically impact the similarity calculation. If the centroids distance is less than a Δ, both face images are also considered from the same person. Otherwise, the images are considered belonging to different people.

Algorithm 1 presents the pseudocode for the people identification process.

3) Age group and gender Classification: After detecting people in the video, gender and age group classification

[1]https://github.com/abars/YoloKerasFaceDetection
[2]https://pypi.org/project/dlib/
[3]https://cmusatyalab.github.io/openface/models-and-accuracies/

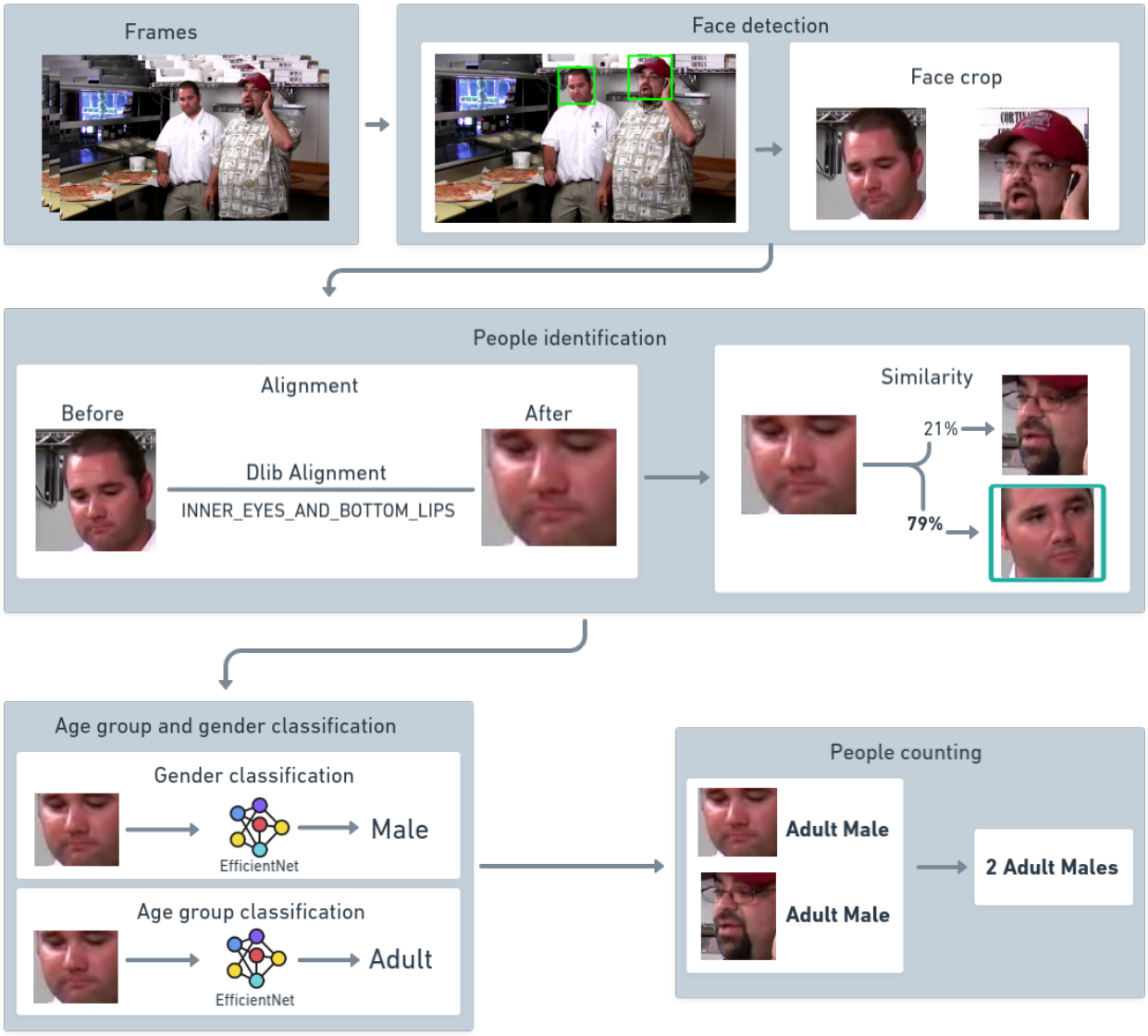

Figure 2. Flowchart for the proposed method for people counting.

classification are performed. In this step, we used a Transfer Learning (TL) approach. TL is the process of using a DNN, trained for a specific problem, into another problem of similar. This is accomplished by transferring the features learned [19]. Such a procedure is less time-consuming than training a new DNN from scratch. The trained DNN is used for extracting features and, next, a new classifier is trained for the specific problem in hand. This approach was proved to be very useful in many computer vision applications [7], [20].

In this work, six pre-trained DNN models were tested: EfficientNet [21], Inception-v3, VGG16, SqueezeNet, SqueezeNet2, and MobileNet. All of them were trained on the ImageNet dataset and these pre-trained models are natively provided by the Keras framework [4]. Each model contains a feature learning component composed of many convolutional layers and a classification component with fully connected layers and softmax outputs.

To improve the robustness of the classifier, online data augmentation was used. Data augmentation consists in creating new images for training the classifier, by means of applying transformations to the original images. The following transformations in the original images were accomplished: shearing, zooming, horizontal flip, translation, rotation, and Random Erasing [22]. The objetive of using data augmentation is to avoid overfitting during the training of the classifier.

To evaluate the performance of the proposed approach, experiments were done using the Adience dataset [23]. The Adience dataset contains 26,580 images of human faces. It was created from Flickr albums to assist the study of gender and age classification. This dataset contains annotations for gender

[4]https://keras.io/

Algorithm 1 People identification algorithm.

Input: current detected face ($current\_face$) and a list of faces already detected, aligned, and vectorized ($detected\_faces$)

Output: Updated list of aligned and vectorized detected faces ($detected\_faces$)

new_face = vectorize(align($current\_face$))
foreach $saved\_face$ *in* $detected\_faces$ **do**
  if *sim*(new_face,$saved\_face$) $< \theta$ **then**
    return True
  if $|$*centroid*(new_face)$-$*centroid*($saved\_face$)$|$ $<=$ Δ **then**
    return True
append($new\_face, detected\_faces$)
return $detected\_faces$

(male and female) and age groups (0-2, 4-6, 8-12, 15-20, 25-32, 38-43, 48-53, 60+). However, such spacing between age groups leads to many categories that are not commonly used in the human behaviors analysis and surveillance context [24], [25]. Therefore, the age group annotations were reorganized into four groups: baby (ages between 0 and 2), child (ages between 4 and 13), adult (ages between 17 and 53), and elderly (ages starting at 60 years).

It is important to notice that the Adience dataset was chosen for this work because it has a large variety of images of people taken in real world scenarios. Also, there are people using accessories, such as hats, glasses, and earphones, in different resolutions and directions, which are commonly captured by real-world surveillance cameras. Figure 3 shows a sample of images found in the Adience dataset.

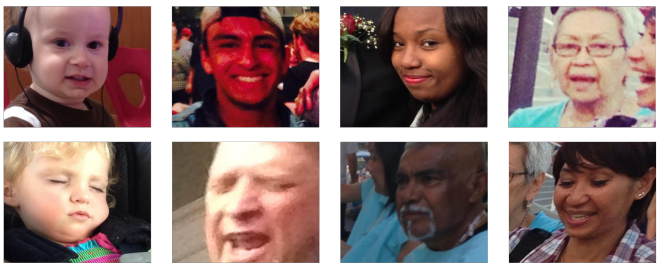

Figure 3. Samples of Adience dataset images.

4) People counting: The last module of the pipeline shown in Figure 2 consists of counting people based on the face attributes extracted in the previous steps, i.e., gender and age group. The output of the module consists of four main pieces of information: total people, total people by gender, total people by age, and total people grouped by gender and age.

III. Experimental Results and Discussions

This Section presents experimental results obtained by the people counting method described in the previous Section. It is also presented some experimental results on the age group and gender classification task by comparing six different DNN used for feature extraction. The 5-fold cross-validation method in a stratified manner was used to assess the generalization performance of the trained models over the Adience dataset. For a fair comparison, all models were trained using Adam algorithm with categorial cross-entropy as loss function. The initial learning rate was set as 0.001. If the validation loss not improved for 2 epochs, the learning rate was reduced by a factor of 5% until achieved 0.0001. The training process is stopped until it reached 100 epochs or the evaluation metric does not improve on the validation set at a patience of 10. Online data augmentation, such as shear, zoom, rotation, and horizontal flip were applied to the original training dataset to enhance the classification performance. All experiments were performed using a workstation with IntelCore-i7 8700 processor, 32GBytes RAM, and an Nvidia Titan Xp GPU. The Keras library using the TensorFlow backend was used to train and test the models.

A. Age group and Gender classification performance

Table I presents the gender classification performance comparison using different DNN models for feature extraction as described in Section II-B3.

The EfficientNet achieved the best result with an accuracy of 97.4%, and the SqueezeNet2 architecture achieved the worst result reaching 91.7%. Notice that we report the performance using accuracy as the gender distribution in the Adience dataset is balanced.

Table I
Comparison of the gender classification performance with the Adience dataset, using different feature extractors.

Model	Accuracy(%)
EfficientNet	**97.4**
InceptionV3	97.1
MobileNet	96.8
VGG16	94.4
SqueezeNet	94.9
SqueezeNet2	91.7

Similar to gender classification, we performed a comparison study to verify the influence of different feature extractors for the age group classification. Results are shown in Table II.

EfficientNet also achieved the best results in age group classification, achieving 96% accuracy, and SqueezeNet2 achieved the worst results with 82.1% accuracy. Notice that we report the age group performance using the F1-score as its distribution is imbalanced in the Adience dataset.

B. People counting experiments

In this section, we present an evaluation of the outputs predicted by the proposed approach in the dataset as described in Section II-A, for people counting in videos by age group and gender.

Table III presents the overall people counting performance provided by the proposed method. Notice that the proposed

Table II
COMPARISON OF THE AGE GROUP CLASSIFICATION PERFORMANCE WITH THE ADIENCE DATASET, USING DIFFERENT FEATURE EXTRACTORS.

Model	F1-score(%)
EfficientNet	**96.0**
InceptionV3	94.3
MobileNet	89.3
SqueezeNet	88.4
VGG16	87.6
SqueezeNet2	82.1

method counted correctly the number of people in 72.8% of the videos and achieved an average F1-score of 79%.

In the one hand, the proposed method detected more people than the correct number of people in 10 videos. Most of these videos were captured by a moving camera, making it difficult to capture faces at different angles, many of them with occlusions. This issue makes difficult to track people correctly.

On the other hand, the proposed method detected fewer people than the correct number in 9 videos. We observed that some videos contain faces at complex angles and occlusions not recognized by the proposed method. This suggests that the proposed method is very dependent on a good face detector.

Table III
PEOPLE COUNTING PERFORMANCE.

Detection	N. videos	N. People	Detected People	Precision	Recall	F1-Score
Exact	51	65	65	100%	100%	100%
Fewer	9	22	13	100%	59%	74%
More	10	10	22	45%	100%	62%
Total	70	97	100	82%	86%	79%

Table IV shows the people counting performance by gender. The proposed approach reached an average F1-score of 79%. In 51 videos where the method correctly predicted the number of people, the proposed approach reached 95% in terms of F1-score. This value is similar to the score obtained in the gender classifier score, shown in Table I, indicating a good generalization capability of the trained gender classifier.

On the other hand, in videos where the proposed approach did not recognize well the several people of the videos, the detected genders were still correctly detected by reaching a precision of 100% in the videos with fewer people detected and recall of 100% in the videos with more people incorrectly detected.

Table V shows the people counting performance by age group. We can notice that the proposed approach also achieves a reasonably good performance, reaching an average F1-score of 71%. Despite being trained over an unbalanced dataset, the age group classifier was able to satisfactorily detect and classify the elderly and child people.

Table VI presents the running time of the proposed method. We vary the number of frames between 25 and 150. The results show that the proposed approach operates at about 6.8 frames per second (FPS) with 150 frames as input. It demonstrates the feasibility of the proposed approach to be used in a quasi-real-time manner. Note that FPS can also be improved by processing the age and gender classification stages in parallel, as indicated in Figure 2. To ensure more reliable results, these experiments were executed 5 times and the average time was reported.

IV. CONCLUSIONS

People counting in videos is an important task that can be employed in a wide range of applications, including video captioning and retrieval, surveillance, and for business purposes.

This paper presented a new method for people counting by gender and age group in moving cameras. It combines a face detection and tracking approach with specialized Deep Neural Networks for gender and age group classification. The proposed method achieved satisfactory results on the proposed dataset by counting people in videos by gender and age group, despite the low quality of some videos and some people at tricky angles. The most challenging part of this work was dealing with occlusions and recognizing faces in angles while tracking faces.

The results reached so far encourage future work towards some improvements in the proposed method to deal better with partially occluded faces or at difficult angles. We intend, also, to explore new methods for face detection, since this is a critical element of the proposed method. In addition, the dataset created will be extended by including new videos so as to reduce the current class imbalance.

ACKNOWLEDGMENT

A. S. Inácio thanks UNIEDU/FUMDES - PÓS GRADUAÇÃO for the scholarship and Federal Institute of Santa Catarina for the support; R. H. Ramos thanks CNPQ for the scholarship; H.S. Lopes thanks CNPq for the research grants no. 311778/2016-0 and 423872/2016-8, and Fundação Araucária for the research grant PRONEX 042/2018. Special thanks to NVIDIA Corp. for the donation of the Titan-Xp GPU boards used in this work.

REFERENCES

[1] M. Pervaiz, Y. Y. Ghadi, M. Gochoo, A. Jalal, S. Kamal, and D.-S. Kim, "A smart surveillance system for people counting and tracking using particle flow and modified som," *Sustainability*, vol. 13, no. 10, p. 5367, 2021.

[2] S. I. Cho and S.-J. Kang, "Real-time people counting system for customer movement analysis," *IEEE Access*, vol. 6, pp. 55 264–55 272, 2018.

[3] E. M. L. Aquino, I. H. Silveira, J. M. Pescarini *et al.*, "Social distancing measures to control the COVID-19 pandemic: potential impacts and challenges in Brazil," *Ciência & Saúde Coletiva*, vol. 25, no. Supl.1, pp. 2423–2446, 2020.

[4] V. Nogueira, H. Oliveira, J. A. Silva, T. Vieira, and K. Oliveira, "Retailnet: A deep learning approach for people counting and hot spots detection in retail stores," in *Proc. of the 32nd Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2019, pp. 155–162.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[6] A. S. Inácio and H. S. Lopes, "Epynet: Efficient pyramidal network for clothing segmentation," *IEEE Access*, vol. 8, pp. 187 882–187 892, 2020.

Table IV
PEOPLE COUNTING PERFORMANCE BY GENDER. M AND F STANDS FOR MALE AND FEMALE, RESPECTIVELY.

Detection	N. videos	Ground Truth		Predicted		Precision		Recall		F1-Score
		M	F	M	F	M	F	M	F	
Exact	51	35	30	38	27	92%	100%	100%	90%	95%
Fewer	9	10	12	8	5	100%	100%	80%	42%	74%
More	10	8	3	17	5	47%	60%	100%	100%	69%
Total	70	53	45	63	37	64%	80%	90%	77%	79%

Table V
PEOPLE COUNTING PERFORMANCE BY AGE. B DENOTES BABY; C DENOTES CHILD; A DENOTES ADULT; E DENOTES ELDERLY.

	N. videos	Ground Truth				Prediction				Precision				Recall				F1-Score
		B	C	A	E	B	C	A	E	B	C	A	E	B	C	A	E	
Exact	51	1	3	59	2	1	2	58	4	100%	100%	100%	50%	100%	67%	98%	100%	86%
Fewer	9	0	11	11	0	0	8	5	0	-	100%	100%	-	-	73%	45%	-	73%
More	10	1	2	6	1	0	4	17	1	0%	50%	35%	100%	0%	100%	100%	100%	55%
Total	70	2	16	76	3	1	14	80	5	50%	66%	78%	50%	50%	69%	81%	67%	71%

Table VI
RUNNING TIME RESULTS.

N. Frames	Running time (seconds) Face Detection	People Identification	Age group classification	Gender classification	People counting	Total time	FPS
25	2.39	2.50	2.29	2.17	0.0001	9.36	2.67
50	3.55	3.71	2.22	2.38	0.0002	11.87	4.21
100	5.75	6.15	2.52	2.65	0.0008	17.08	5.86
150	7.66	8.50	2.78	2.89	0.0006	21.83	6.87

[7] M. Gutoski, M. Ribeiro, L. T. Hattori, M. Romero, A. E. Lazzaretti, and H. S. Lopes, "A comparative study of transfer learning approaches for video anomaly detection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 4, p. 2152003, 2021.

[8] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. early access, 2021.

[9] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzalez, "Age and gender recognition in the wild with deep attention," *Pattern Recognition*, vol. 72, pp. 563–571, 2017.

[10] L. Massa, A. Barbosa, K. Oliveira, and T. Vieira, "LRCN-RetailNet: A recurrent neural network architecture for accurate people counting," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5517–5537, 2021.

[11] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc.of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2011, pp. 190–200.

[12] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5288–5296.

[13] G. A. Sigurdsson, G. Varol, X. Wang, I. Laptev, A. Farhadi, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *ArXiv preprint*, vol. 1604.01753, 2016.

[14] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[15] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[17] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 343–347.

[18] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint*, vol. 1411.7923, 2014.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[20] M. Romero, M. Gutoski, L. Hattori, M. Ribeiro, and H. Lopes, "A study of the influence of data complexity and similarity on soft biometrics classification performance in a transfer learning scenario," *Learning and Nonlinear Models*, vol. 18, no. 2, pp. 56–65, 2020.

[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint*, vol. 1905.11946, 2019.

[22] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[23] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[24] K. R, "Deep learning for age group classification system," *INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES*, vol. 4, no. 2, pp. 16–22, Dec. 2018.

[25] R. Madigan, S. Nordhoff, C. Fox, R. Ezzati Amini, T. Louw, M. Wilbrink, A. Schieben, and N. Merat, "Understanding interactions between automated road transport systems and other road users: A video analysis," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 196–213, 2019.