Deep Learning based approach to detect Customer Age, Gender and Expression in Surveillance Video

Dr. Earnest Paul Ijjina*
Assistant Professor, Department of Computer Science and Engineering
National Institute of Technology Warangal, India-506004
Email : *iep@nitw.ac.in
Goutham Kanahasabai†, Aniruddha Srinivas Joshi ‡
B.Tech., Department of Computer Science and Engineering
National Institute of Technology Warangal, India-506004
Email : † gauthamkanags@gmail.com, ‡ aniruddha980@gmail.com

***Abstract*—In the current information era, customer analytics play a key role in the success of any business. Since customer demographics primarily dictate their preferences, identification and utilization of age & gender information of customers in sales forecasting, may maximize retail sales. In this work, we propose a computer vision based approach to age and gender prediction in surveillance video. The proposed approach leverage the effectiveness of Wide Residual Networks and Xception deep learning models to predict age and gender demographics of the consumers. The proposed approach is designed to work with raw video captured in a typical CCTV video surveillance system. The effectiveness of the proposed approach is evaluated on real-life garment store surveillance video, which is captured by low resolution camera, under non-uniform illumination, with occlusions due to crowding, and environmental noise. The system can also detect customer facial expressions during purchase in addition to demographics, that can be utilized to devise effective marketing strategies for their customer base, to maximize sales.**

***Index Terms*—Age estimation, Gender prediction, Expression recognition , Deep Learning**

I. Introduction

In developing counties like India, where retail sales is major means of marketing, there is a huge potential in utilizing customer behaviour analysis to optimize sales. The customer demographics such as age and gender, in addition to the sentiment a customer experiences towards particular products plays a significant role in the sales and operations of retailers and small scale vendors. The analysis of the customer base is crucial for retailers to stay on top in the business marketing and maintaining a healthy group of consumers to make profits. As a retailer, anticipating and understanding the wants and interests of the existing customers can prove to be an effective route to tackle the battle of knowing what to stock up for the future. Such analysis can help retailers in ensuring that customers find what they are looking for, and can thus ensure a growing consumer base towards the business. Apprehending the aforementioned traits not only helps small scale vendors, but also advocates retail mall's such as Walmart and Target in better engaging customers [1], [2]. The knowledge of the average shopper in addition to the expressions one portrays while shopping, plays a paramount role.

In the consumer base, we find that each individual customer exhibit inclination towards a said commodity. So categorizing customers by age and gender is essential for better marketing. The studies [3] suggest that around 20% of new businesses survive their first year of operation and half of the small businesses close down in the first 5 years. Although several factors such as on-line competition, low profits, unavailable resources etc., may lead to the decline of a business. Analysing the consumer base over a period and understanding their sentiments over the product line-up can aid retailers in securing better profit margins ensuring survival and growth in the long run. Targeting consumers by segmenting the market ensures a healthy increase in consumers over time, which is a crucial factor in the functioning of current businesses.

Facial expression recognition has gained significant attention in the field of marketing, aiding enterprises to comprehend opinions expressed by customers while purchasing products. Traditionally, majority of enterprises used conventional methods of marketing such as advertisements, customer satisfaction surveys etc. These methods often prove to be rather time consuming and are more often than not an expensive venture. With the advent of technology, customer analytics can be used to deliver fitting products to the consumers. Sentiment is the emotion behind engaging consumers, and capturing customer sentiment helps us understand the following metrics that can aid in devising upcoming marketing strategies. Some of the common customer sentiments are:

- **Overall satisfaction:** this helps us understand whether the experience at the location proved fruitful for the customer
- **Loyalty:** whether existing customers would recommend the said retailer to others
- **Future Engagement:** whether the customers would engage with the retailer in the future

Analysing the aforementioned metrics gives valuable insight to potent marketing strategies that are more engaging and beneficial to the customer.

In this work, we propose a deep learning based framework to identify the age, gender and expressions of the customers

from surveillance videos. Video surveillance through Closed Circuit TeleVision (CCTV) cameras is the most commonly used set-up to monitor and track humans in retail stores [4].

In this paper, we propose an approach to detect age, gender and expressions of the customers effectively in CCTV footage obtained from a typical video surveillance system. The surveillance videos from CCTV cameras in stores is given as input to the proposed framework to detect faces and estimate age, gender and expression of the person. The proposed framework is designed to operate under real world environmental conditions like poor illumination, over exposure of background, challenging angles of view etc., which makes the recognition/estimation task more challenging.

II. Related Works and Literature

In this work, we primarily focus on two pivotal demographics of the consumer base: age and gender, which are the crucial factors in helping retailers and marketers come up with effective business strategies. In addition, we also focus on extracting the facial expressions that the customers exhibit during their engagement in the stores, for reasons mentioned in section I. A typical computer vision approach for age detection either assigns a numerical age of the estimated age (or) a categorical age-group like youth, old etc., to each subject. The typical gender recognition system outputs a binary label for each subject indicating the gender as male (or) female.

The earliest work on Age classification involved a method proposed by Kwon *et al.* that extracted the primary facial features such as eyes, chin, mouth, nose etc. and these features were therein used to compute ratios that set apart seniors from toddlers [5]. The computed ratios aid a wrinkle index which is used to categorise the people into three different age groups. However, this approach has the constraint that the person in consideration needs to be looking directly at the camera and it does not provide us with age as a physical number. Moreover, this approach requires an accurate localisation of facial features, which is a challenging feat on its own and thus proves inadequate for real world environments such as shops. A similar approach to model the progression of age attributes was proposed by Ramanathan *et al.* [6] but it too suffers from the aforementioned limitations.

An aproach based on local features representing images of faces is proposed by Yan *et al.*. This methodology involved the utilization of Gaussian Mixture Models (GMM) [7] to symbolise the distribution of facial patches [8]. In a similar work [9], GMMs were used for representing the distribution of facial measurements where more robust descriptors were used instead of the traditional pixel patches. Super vectors were used for representing facial patch distributions in [10].

The methods described above prove to be effective on datasets containing frontal images of faces and are not suitable for surveillance videos that always tend to have images of faces captured at different angles.

Coming to gender classification, early works include a method proposed by Moghaddem *et al.* [11] that involved the utilisation of Support Vector Machine(SVM) classifiers, wherein they were applied to input intensity values of images for classifying gender. Another work that involved image intensities was by Baluja *et al.* [12], however they used Adaboost.

For the purpose of gender classification, most of the works tend to utilise a well versed dataset known as the FERET benchmark [13]. The FERET dataset is a composition of images that were obtained from well illuminated, controlled environments and are thus not by any means close to images obtained from the frames of CCTV videos, which involves daunting factors such as noise, occlusions in between frames etc. As a result, we tend to focus more on CCTV videos placed in stores which tend to capture the environment in its natural element, posing credible challenges to work with.

Expression recognition from facial features has gain attention in the marketing space as it is a natural indicator of a customers emotions. In the work of Ekman *et al.*, Facial Action Coding System(FACS) [14], action parameters were designated to each of the facial expressions which were therein used to classify human emotions.

The layout of the rest of the paper is as follows: Section III delineates the proposed framework of the paper. Section IV contains the analysis of our experimental results. Section V gives the conclusions and future work. Section VI contains the acknowledgements.

III. Proposed Approach

In this section, we elaborate our proposed framework, which is illustrated in Figure 1. We describe how the framework is realized in order to detect the customer related information. Our predominant goal is to identify the age, gender and expressions of customers in surveillance video, so that it can be further utilized in consumer analytics to anticipate sales.

The proposed methodology involves the following tasks:

T1: Face Detection
T2: Age and Gender Estimation
T3: Expression recognition

The details of each of these tasks will be exemplified in the remainder of this section.

A. Face Detection

In this section, we describe the method used to detect the region of the face in the video frames. For this purpose, we use the Haar Cascade object detection model [15], which is a machine learning based approach wherein a cascade function is trained on a positive and negative images, i.e., images with faces and images without faces respectively, thereby detecting faces like objects in given images.

The details of this method is as follows. The algorithm consists of four major steps, namely: Haar Feature Selection, Creating Integral Images, Adaboost Training and finally Cascading Classifiers. The first step is to collect all the Haar features, which operate on rectangular regions in the detection window, calculate the intensity of the pixels in each of these rectangular regions, enumerate the difference between these sums. Integral images are used to speed up this process.

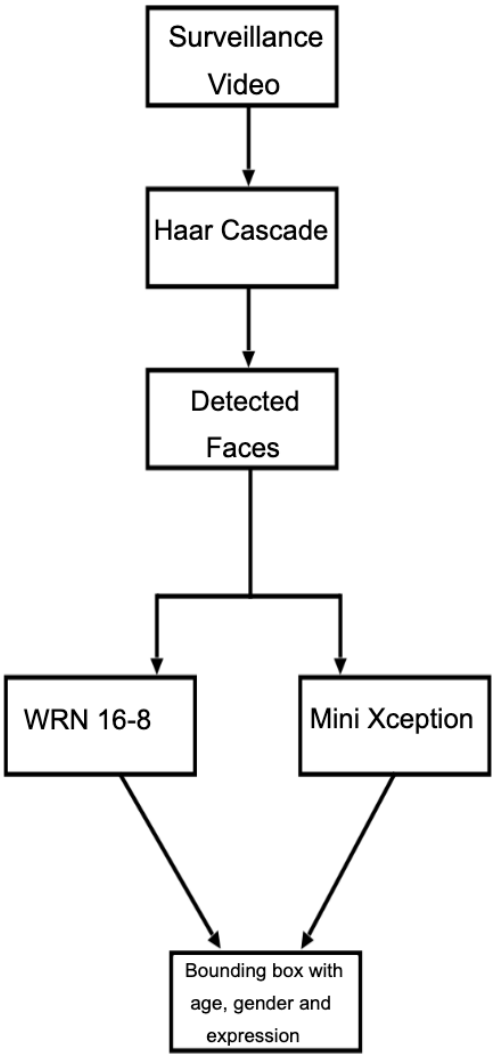

Fig. 1: Workflow of the proposed Framework

Among all the features extracted, the best and the most relevant features are selected using Adaboost, which performs the aforementioned task in addition to training the classifiers that use them. Adaboost builds a strong classifier as a linear weighted combination of weak classifiers. Each Haar feature acts like a weak classifier, and a significant amount of Haar features characterizing the object are cascaded to form a strong classifier.

The cascade classifier is therefore an agglomeration of stages, wherein each stage is a unity of weak learners. Stages are trained to high degrees of accuracy by taking into picture, a weighed average of the decisions undertaken by the weak learners. Each stage of the classifier labels the current region as either positive or negative, indicating that an object was found or not found respectively. The output of this stage is the coordinates of the bounding boxes enclosing the face regions in the input. This information is utilized in remaining tasks i.e, T2 and T3.

B. Age and Gender Estimation

In this section, we describe the methodology used to extract the age and gender corresponding to the face region obtained from task T1. In this task, we use the Wide ResNet 16-8 (WRN-16-8), a Wide Residual Network [16] architecture to estimate age and gender. The architecture considered in this work gave preference to width over depth as effective training of deeper models is a complex task. We chose a Wide Residual Network architecture with more convolution layer filters to improve its effectiveness with less number of layers. The other key advantage of increasing width instead of depth is more computational efficiency. This Wide ResNet model when compared to the traditional ResNet [16], has less number of layers and performing twice as fast as the former. The model considered in this work, the Wide ResNet $16-8$, i.e WRN-$16-8$ with 16 convolutional layers and a widening factor of 8, which is the number of feature maps per layer is shown in Figure 2. The input to this network is a 64×64 RGB image. This 16 layer neural network has the same accuracy as a deep network which has 1000-thin layers, and is also several times faster to train. This suggests that residual blocks are crucial factor to the effectiveness of a deep residual network. The output of this phase is the estimated age, which is numeric value and the gender i.e., male or female, of the subject.

C. Expression recognition

In this section, we explain the methodology used for facial expression recognition. We utilize the output of task T1, which is the region of the face in the frame, enclosed by the bounding box. The model chosen for recognizing the facial expression is the mini Xception model [17] which is inspired by the well versed Xception architecture [18]. The mini Xception model doesn't use fully connected (FC) layers in the network architecture, but utilizes the residual modules [19] and depth-wise separable convolutions [20]. The use of depth-wise separable convolutions reduce the computation in comparison to regular convolutions. Residual models serve the role of altering the desired mapping between two successive layers in the network, causing the difference between the original and desired featured map to be learnt.

Figure 3 illustrates the mini Xception architecture considered in this work. It is a Fully Convolutional Network (FCN) with 4 residual depth wise separable convolutions, with each convolution operation being followed by a batch normalisation and ReLU activation. Global average pooling and a softmax activation function is applied in the last layer to predict the expression.

The output of this phase, is the expression portrayed by the person, which belongs to one of *happy*, *sad*, *anger*, *neutral*, *surprise*, *fear*, *disgust*. This behavioural information can be used to understand the customer preferences of merchandise.

IV. Experimental Evaluation

In this section, we discuss the results of our proposed approach on the surveillance video dataset. All the experiments

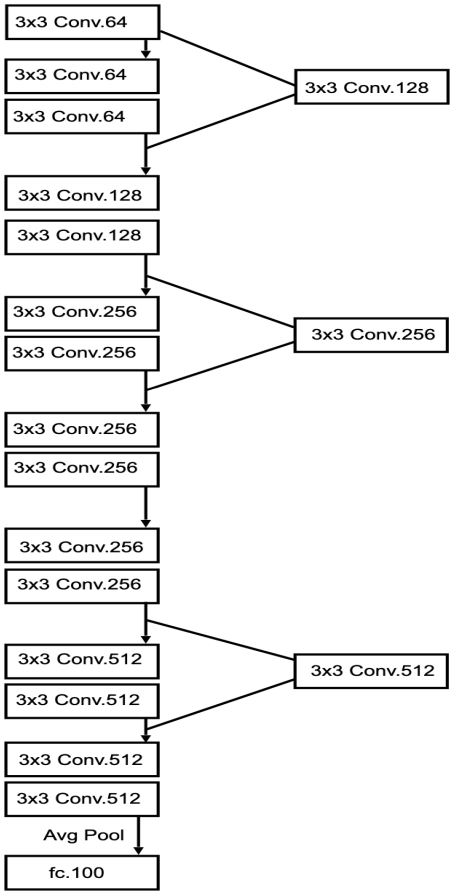

Fig. 2: WRN-16-8 model architecture

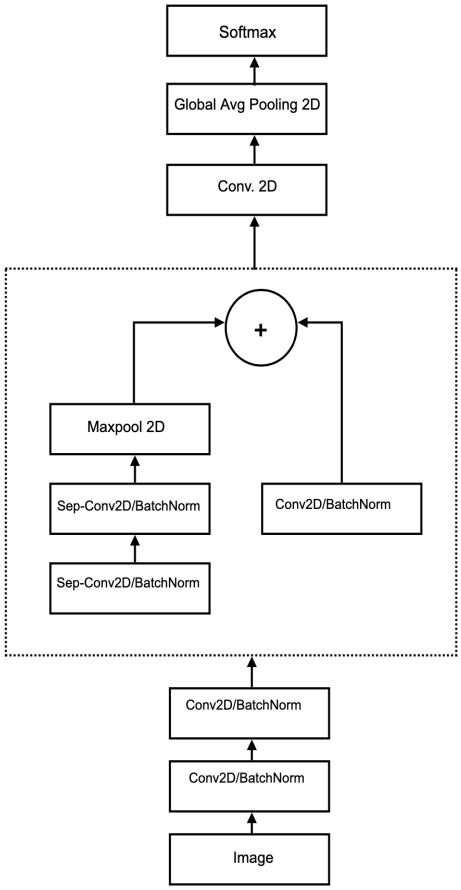

Fig. 3: mini Xception model architecture

in this work were conducted on Google CoLab [21] in an environment with Intel(R) Xeon(R) CPU @ 2.00GHz with NVIDIA Tesla T4 GPU, 16 GB GDDR6 VRAM and 13GB RAM. All programs were written in Python - 3.6 and utilized Keras - 2.3.1 and OpenCV - 4.2.0.

A. Dataset Used

The dataset used in this study is a collection of surveillance videos of a garment store, capturing sales information. The videos contain the interaction between the salesmen and customers while selecting a garment for purchase. The videos are obtained from 2 different CCTV cameras in the store's infrastructure, thus enabling us to work on real-life surveillance videos taken from multiple angles of view and practical illumination conditions. There are a total of 15 video samples in the dataset with an average duration of 5 minutes. The resolution of the videos is 944 $\times$ 576 pixels. The raw videos without any preprocessing like enhancement (or) denoising is used in this work to evaluate the suitability of the model for practical use. The typical sales videos in the dataset are shown in Figure 5.

B. Results and Statistics

This section presents the results obtained by evaluating the proposed approach on the video dataset. Figure 4 illustrates the output of the proposed approach on few instances of the dataset. The bounding box around the detected faces as described in Section III-A for task T1, which is then utilized to obtain age, gender and facial expression as described in Section III-B for task T2 and Section III-C for task T3 respectively.

The numerical results of the proposed approach are given in Table I. On the surveillance videos dataset, the model achieved a Gender classification accuracy of 82.9%. The age estimation is evaluated to assign the subject to the right age group, in {0-9, 10-19,....,70-79} ranges. For this metric, the model obtained an accuracy of 70.8%. These results were obtained on real-world in-door video with significant background noise.

Proposed approach	Gender	Age
	82.926	70.804

TABLE I: Performance of the proposed model in percentage

The model achieved better accuracy when the subjects is facing the camera. However, its accuracy reduces when the customer is not facing the camera directly. The framework still managed to achieve good estimate of age in both scenarios.

The model had difficulties in detecting the side facial profile of customers not facing the camera, thereby affecting its

Fig. 4: Visualization of results for video

accuracy. However, the framework still managed to achieve good estimate of age in such scenarios.

The predicted expressions were manually verified, as the labels for of the expressions of the subjects in the surveillance video is unavailable i.e., unlabelled video. The manual verification suggests that the expressions were reasonably accurate. Similar works in the field of face and gender detection do not cater specifically to identifying subjects of Indian origin, as they were largely trained on established datasets comprising of facial images of people with non-Indian origin. This model was optimized for detecting the age and gender traits of customers belonging to Indian origin.

V. Conclusion and Future Works

In this work, an approach to detect the age, the gender and the expressions of the consumers in surveillance video is proposed. The proposed approach is able to recognize faces of consumers in below average video resolution and is also able to identify their age and gender to a fair degree of accuracy, achieving a 82.9% accuracy rate for gender and 70.8% accuracy for age-range. The accuracy of the predicted expressions is verified manually, and is found to be reasonably accurate. The future work aims to extend this work to more challenging data.

VI. Acknowledgement

This work was done by Mr. Goutham Kanahasabai and Mr. Aniruddha Srinivas Joshi (final year B.Tech students) under the guidance of Dr. Earnest Paul Ijjina (Assistant Professor), in Department of Computer Science and Engineering, National Institute of Technology Warangal, as a part of their final year project. The authors express their gratitude to the CSE department and the Institute, NIT Warangal for their efforts in developing the research environment, where this study was conducted. We also thank Google Colab for providing access to computational resources used to run these algorithms.

References

[1] "Walmart customer demographics," https://snapshot.numerator.com/retailer/walmart.
[2] "The average target shopper," https://extension.psu.edu/understanding-your-customers-how-demographics-and-psychographics-can-help.
[3] https://www.shopkeep.com/blog/why-small-businesses-fail.
[4] T. Kanade, R. Collins, A. Lipton, H. Fujiyoshi, and D. Duggins, "A system for video surveillance and monitoring cmu vsam final report," p. 135, 11 1999.
[5] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images."
[6] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, June 2006, pp. 387–394.
[7] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, July 2008.
[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition (2nd Ed.)*. USA: Academic Press Professional, Inc., 1990.
[9] Shuicheng Yan, Xi Zhou, Ming Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from patch-kernel," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.
[10] S. Yan, M. Liu, and T. Huang, "Extracting age information from local spatially flexible patches," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ser. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 9 2008, pp. 737–740.
[11] B. Moghaddam and Ming-Hsuan Yang, "Learning gender with support faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 707–711, May 2002.
[12] S. Baluja and H. Rowley, "Boosting sex identification performance," *International Journal of Computer Vision*, vol. 71, pp. 111–119, 06 2007.
[13] P. J. Phillips, Hyeonjoon Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 22, no. 10, pp. 1090–1104, Oct 2000.
[14] P. Ekman and W. V. Friesen, "Facial action coding system: Manual," 1978.
[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Dec 2001, pp. I–I.
[16] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016.
[17] O. Arriaga, M. Valdenegro, and P. Plger, "Real-time convolutional neural networks for emotion and gender classification," 10 2017.
[18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.
[21] "Google colab," https://colab.research.google.com/.

Fig. 5: Sample video in the dataset