Customer and Target Individual Face Analysis for Retail Analytics

Nabil Tahmidul Karim[1], Sanjana Jain[1], Jednipat Moonrinta[2], Matthew N. Dailey[2], Mongkol Ekpanyapong[2]
Asian Institute of Technology (AIT), Pathum Thani 12120, Thailand
nabil@ait.asia, sanjana@ait.asia, st117746@ait.asia, mdailey@ait.asia, mongkol@ait.asia

Note: The first two authors contributed equally to this work.

***Abstract*— Statistics about customer satisfaction are extremely valuable for retail stores. We present an automated approach to obtaining this information using image processing and deep learning. Our system combines face detection and tracking, best view estimation, repeat customer identification, blacklisted customer warnings, and facial sentiment classification. A series of experiments shows that each of the modules in the combined systems achieves satisfactory results.**

Keywords—Deep Learning; Convolutional Neural Networks; Face Detection; Best-view Analysis; Age Classification; Gender Classification; Emotion Classification; Face Verification; Repeat-customer Analysis; Target Customer Classification

I. INTRODUCTION

With rapid changes in technology, small retailers face difficulties keeping pace with new development. Customer experience, customer interest, and an exclusive, engaging surrounding are factors influencing the development of any retail store. Hence, acquiring information about and optimizing consumer experience will enhance business development. Businesses that actively respond to customer sentiment and provide a safe and secure environment will tend to be more competitive than those that do not. Applications for identifying and tracking target individuals in stores are finding their way onto retail businesses' agendas. These include applications for identifying blacklisted customers, identifying members of loyalty schemes, and giving rewards for special customers. Already, most retailers offer membership card programs that offer benefits in return for useful information on customer preferences. This source of data can be inconvenient for customers and imprecise for retailers, as customer sentiment cannot be drawn from transaction data alone. We therefore propose more active collection of customer feedback information through the use of video analytics, deep learning, and image processing. In the remainder of the paper, we describe a prototype system design and an experimental evaluation of the prototype. We find that the customer identity, gender, and emotion classification modules perform fairly well, but more improvement is needed for age estimation.

II. VIDEO ANALYTICS FOR RETAIL

A schematic overview of our system is shown in Fig. 1. Face detection is performed to extract face images of individuals arriving at the store. Faces are tracked as long as

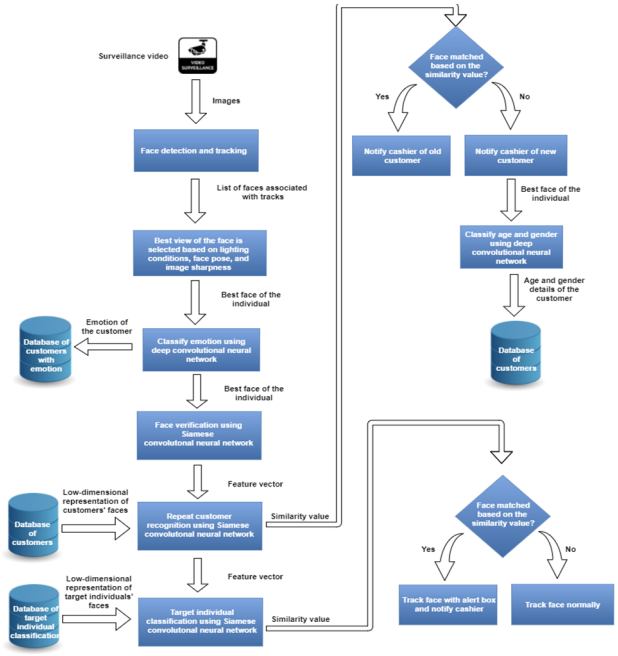

Fig. 1. Data flow for the proposed system.

they are visible. A variety of heuristics, such as the quality of the lighting conditions, image sharpness, and face pose are used to obtain the best view of a given face over the period of the customer's visit. The face image best matching the pose criteria is passed periodically through a convolutional neural network (CNN) to determine the individual's sentiment. A face verification CNN is used to determine whether the individual is a new or old customer and to raise alerts for blacklisted customers. For new customers, we use separate CNNs to determine the individual's age and gender.

As discussed earlier, face detection is the first phase of the proposed system. It is performed on each frame of the live feed acquired from the video stream using a custom Viola and Jones AdaBoost detection cascade classifier using Haar-like features. The quality of the face image is determined using three metrics to assess the quality of a face image: sharpness, illumination, and pose of the face. The DLIB fiducial point detector is used to determine the approximate boundaries of the face region in the image. A gradient-based approach is performed with a Gaussian weighted envelope centered on the face region to determine the sharpness of the face image.

Entropy-based analysis is performed using a frequency histogram of gray-levels to determine the best-illuminated face

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

Input: $\mathcal{T}$ = empty trajectory list
for each *frame* F *of the live feed* **do**
  Detect faces in frame F
  Extract face portion f from the frame
  for each *extracted face image* f *in* F *meeting size requirement* **do**
    t = Person_Tracking($\mathcal{T}$,f){Associates tracks with faces}
    Update $\mathcal{T}$ to associate f with t
  end for
  for each *track* t *in* $\mathcal{T}$ **do**
    if a face is detected for t **then**
      $updated = false$
      Retrieve face image f for t
      Perform sharpness analysis, entropy based analysis, and pose estimation for f
      Integrate the above measures to obtain a quality measure q for image f
      if $q > t.q$ **then**
        $t.f = f$
        $t.q = q$
        $updated = true$
      end if
      $t.e$ = Emotion_Classification($t.f$) {Determines sentiment of $t.f$ and stores with t}
      if $updated$ **then**
        new = Customer_Verification($t.f$) {Old/new customer. Associates t with customer db}
        if new **then**
          $t.age$ = Age_Classification($t.f$) {Determines age of $t.f$ and stores t}
          $t.gender$ = Gender_Classification($t.f$) {Determines and stores gender of $t.f$}
          Store $t.age$, $t.gender$ in the database of customers.
        end if
      end if
    end if
  end for
end for

Algorithm 1: Overall algorithm for proposed system.

images to find fiducial points, with a Gaussian weighting envelope centered on the face region. In the case where DLIB fails, we use a Gaussian weighting envelope over the entire detected face region.

The 3D pose of the face aids in determining the quality of a face image, wherein a frontal pose image represents better quality as compared to a non-frontal pose. The convolutional neural network proposed by Qiexing [1] for predicting 3D head pose is used. All of these quality measures are then integrated to output a single metric that aids in determining the highest quality image.

Face images that pass the pose requirements are forwarded to the emotion convolutional neural network model to classify the sentiment of the individual. This will help to determine whether the customer is feeling positive or negative regarding the retail store's service.

We aim to notify the cashier or other customer service staff-person about the arrival of an old customer or a new customer. We also aim to alert staff or security guards about the presence of a blacklisted person. Therefore, after obtaining the best-view face of an individual, the face image is passed to a Siamese convolutional neural network model for verification purposes. For customer analysis, a comparison is performed between the face image and the gallery images of each customer in the database. For blacklisted customer analysis, a comparison is performed between the face image and the gallery images of each blacklisted individual in the database. The Siamese CNN produces a feature vector for the probe and target face. Afterwards, the cosine similarity between the faces' feature vectors is calculated. Based on the similarity and a threshold, the cashier is notified as to whether a new or an old customer has come as well as if a blacklisted individual has entered. In case of the arrival of a new customer, his/her best view face is stored in the database of customers.

After verifying whether an old or a new customer has come, in case of a new customer, the best face image of that individual is passed to two convolutional neural network models to classify the age and gender of that individual. Lastly, the classified age and gender are stored in the customer database for that particular customer.

III. EXPERIMENTAL VALIDATION

To evaluate the system, we acquired training and testing video feeds from cameras at the Hom Krun Coffee shop at the Asian Institute of Technology (AIT).

A. Face Detection:

OpenCV's face detection cascade with integrated frontal and profile view cascades was initially run on a test video from the same view. However, many false positives, most belonging to the fixed environment of the shop, were obtained.

Hence, we trained a custom face detection cascade. A series of experiments was performed, and the positive and negative images obtained from these experiments were further used as training images for the following experiments. Open-source datasets were also used to improve the results and to prevent the detector from being too dependent on the Hom Krun-specific data. Positive images were extracted from the Hom Krun video data (7,960 images) and from the testing images of the CBCL dataset [2] (472 images), giving a total of 8,432 faces. Negative images were extracted from the non-faces detected as faces by the OpenCV detector and previously trained Hom Krun based detectors when run on the test videos (1,300 images), from non-face selection after extracting 50 × 50 images with a stride of 25 on five Hom Krun video frames (15,550 images), from the CBCL training dataset (4,548 images), and from the Ali and Dailey head dataset [3] (428 images), giving a total of 21,176 non-faces. The model was trained for 22 stages. On testing, the best suitable parameters for the cascade were determined, and convincing results were obtained. The custom Haar cascade obtained fewer false positives as compared to the original OpenCV face detector. The final face detector was run on the training and testing videos. The extracted face images were then manually classified in terms of identity, age (0-7, 8-20, 21-35, 36-59, 60+ years), gender, and emotion (positive, negative, or neutral) these labels are used by the following modules.

B. Face Recognition:

In order to perform face recognition, the model provided by Wen et al. [4] was trained with the CASIA Webface dataset from scratch for 112,000 iterations and a batch size of 64. Each training image was resized to 100×100 and subtracted from the mean image before being passed to the model. The model was then tested on the Hom Krun test dataset. A feature vector is generated for each image being passed to the model. The cosine similarity between feature vectors of the pairs is then computed. If the cosine similarity is more than the estimated threshold, then it is considered a positive pair. Otherwise it is considered a negative pair. We achieved a test accuracy of 70.8%. Afterwards, the Hom Krun training dataset was used to fine-tune the network. The dataset consists of 212 subjects with approximately 70 images per subject in different poses, expressions, and lighting conditions, resulting in a total of 10,678 training and 1,060 validation images. The model was

fine-tuned for 80,000 iterations. As before, the images were resized and the mean image was subtracted. On testing on the Hom Krun test dataset and following the previously-stated steps, an accuracy of 87.02% was obtained.

For comparison purposes, the GoogleNet model constructed by Szegedy et al. [5] was used to perform the same experiments. The model was trained on the unaligned CASIA-WebFace dataset from scratch for 276,000 iterations with a batch size of 32, due to memory restrictions. The images were resized to 256×256 to meet the requirements of the GoogleNet model. The model was then fine-tuned using the Hom Krun dataset for 66,800 iterations with a batch size of 32. On testing on the Hom Krun test dataset and following the previously stated steps, an accuracy of 89.04% was obtained.

C. Gender Classification

In order to classify customers' gender, the CNN model proposed by Levi and Hassncer [6] was used. It consists of three convolutional layers and two fully-connected layers. The images are non-isotropically rescaled to a size of 227 × 227 and fed to the network. The last fully connected layer, the output layer, has 2 units fully connected to the output of the previous fully connected layer of size 1×512, yielding un-normalized class scores for male and female. This score is further passed to a soft-max layer, which generates a normalized probability score for each gender. The class with the highest probability is chosen as the prediction for the test input image. Initially, the network was trained on the Adience dataset provided by Eidinger, Enbar, and Hassner [7], consisting of 16,263 training images that are taken in different poses and illumination. Out of 16,263 images, 7,617 are males and 8,646 are females. Afterwards, the Caffe model was fine-tuned using the Hom Krun dataset. The dataset comprises 7,481 images, out of which 3,053 are males and 4,428 are females. The dataset was divided with a 80:20 ratio into training and validation sets, which resulted in 5,985 training and 1,496 validation images. For both datasets, the training images were resized to 256×256, and later a crop of 227×227 was taken during training, which was conducted for 50,000 iterations. Afterwards, the model was tested on the Hom Krun test dataset, which consisted of 384 images. The model successfully predicted 375 images correctly, giving an accuracy of 97.65%.

D. Age Classification:

In order to classify the customers' age, the same CNN model proposed by Levi and Hassncer [6] was used as before. Initially, we aimed to classify images into two age classes, i.e., 21-32 and 33-59 years. This model was trained on the Adience dataset, which comprises 8,700 images, out of which 3,800 are from the range of 21-32 years old and 4,900 are from the range of 33-59 years old. Afterwards, the trained model was fine-tuned using the Hom Krun dataset, consisting of 10,812 images. The training dataset was divided in an 80:20 ratio to obtain 8,805 training and 2,007 validation images. For both datasets, training was conducted for 50,000 iterations, Afterwards, the model was tested on 378 test images unseen during training, out of which 175 are subjects of age 21-32 years and 203 are subjects of age 33-59 years. The model obtained an accuracy of 72.49%, predicting 274 images correctly.

As the network did not provide strong enough results to be applicable in real-life scenarios, the same experiments were performed with a smaller version of the CNN model. Originally, the network contained three convolutional layers and three fully-connected layers. The network was cut short by removing one convolutional layer and one fully connected layer. The network was again trained with Adience initially and then with the Hom Krun dataset like before. Afterwards, this model was tested again on the same dataset and achieved an accuracy of 74.44%, showing improvement from the previous network.

Since classifying the age of an individual into two classes is itself not useful in many real-life scenarios, the network was restructured again by changing the number of neurons in the output layer to 5, which enables us to classify age into five classes, i.e., 0-7, 8-20, 21-35, 36-59, and 60+ years. The network was trained as before with Adience and afterwards with the Hom Krun dataset. Then the network was tested on the same test dataset, which gave an accuracy of 67.7%.

As the network was unable to accurately predict age with five classes, the above experiments were repeated using the Googlenet CNN model proposed by Szegedy et al. [5]. Initially, the network was trained with the Adience and Hom Krun datasets combined together, which resulted in 22,508 training images and 2,138 validation images. Afterwards, the network was fine-tuned using "The Images of Groups" dataset built by Gallagher and Chen [8]. This dataset consists of 23,953 images. For both cases, training was conducted again for 100,000 iterations. Then, the model was evaluated on the Hom Krun test dataset, and it obtained an accuracy of 79.9% with 100% 1-off accuracy, surpassing the previous best reported result of 96.6% 1-off accuracy by Rothe, Timofte, and Van Gool [9]. We should note however that those authors used 7 classes to classify age, whereas we use only 5.

E. Emotion Classification:

In order to classify the sentiment of customers, the same CNN model proposed by Levi and Hassncer [6] was used as before. To classify emotions into three classes, i.e, positive, neutral, and negative, the number of neurons in the output layer of the network was changed to 3. Initially, the network was trained on the extended Cohn-Kanade dataset constructed by Lucey et al. [10], containing 593 sequences from 123 subjects. Each sequence is constructed from a neutral expression in the first frame changing to the last frame containing a peak expression. Since the subjects portray seven different expressions, i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise, we took the frames containing the peak anger, contempt, disgust, and sadness expressions as negative images, and the frames containing the peak happy expression as positive images. The initial frames containing the pure neutral expressions were considered neutral images. Therefore, the network was trained with 8,568 training images, out of which 1,646 are positive, 3,695 are neutral, and 3,227 are negative. Afterwards, the network was fine-tuned with a combined dataset of the California Facial Expressions of Emotion (CAFE) created by Dailey et al. [11] and the Japanese Female Facial Expressions (JAFFE) designed by Lyons and Akamatsu [12], containing 288 images, out of which 82 are positive, 97 are

neutral, and 109 are negative. Afterwards, the network was further fine-tuned using the Hom Krun dataset comprising 5,546 images, out of which 1,941 are positive, 3,036 are neutral, and 549 are negative. For all datasets, training was conducted for 100,000 iterations for each case. The model was evaluated on the Hom Krun test dataset consisting of 377 test images, 111 of which are positive, 244 are neutral, and 22 are negative. The network obtained an accuracy of 75.06% on the test dataset.

By performing data augmentation on the Hom Krun dataset, the number of negative images was increased to make a more balanced dataset, which resulted in 19,176 training images, 6,488 of which are positive, 6,567 are neutral, and 6,121 are negative. This dataset was used again to fine-tune the network acquired after training with CAFE and JAFFE. This resulting model was tested again on the same test dataset and gave a significant improvement in accuracy, obtaining 87.36% accuracy on the test dataset.

F. Integrated System:

All of the experiments were performed on an Intel Core i7-6700 with 8GB RAM and GeForce GTX 1070 GPU. The modules were integrated into a single application with a basic tracking algorithm. Test video was recorded at the Hom Krun coffee shop at 15 FPS. The integrated system processed, classified, and displayed the output in real time. A sample of results obtained from running the entire system on test video is shown in Fig. 2.

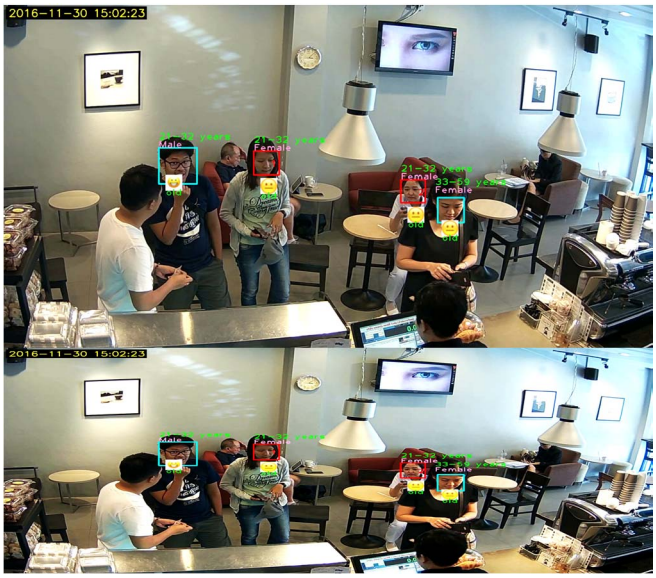

Fig. 2. Experimental results on a test video.

IV. Discussion and Conclusion

The experiments reported on in this paper demonstrate the feasibility of real-time customer identification and sentiment analysis for developing retail enterprises. Due to space limitations, we are unable to provide all details, but the customer identity, gender, and emotion modules perform well enough to be used in real world environments; however, the age module needs some improvement. The system can serve as a baseline for many retail stores to improve customer service. The hardware installation is particularly easy. Further steps can be taken to enhance the performance of each module. After combining with appropriate human interfaces, we expect the prototype to help improve customer satisfaction in practice. As one caveat, we note that placement of cameras in public places raises concern for invasion of privacy and rights of individuals.

Acknowledgement

The authors are grateful for the support of the owners of Hom Krun Coffee shop at AIT for allowing us to install cameras and prototype the system. We are grateful to Faisal Islam for help with ground truth annotation of video data. Nabil Karim and Sanjana Jain were supported by scholarships and fellowships from AIT and the Royal Thai Government.

References

[1] Qiexing. (2016). 3D Pose Estimation. https://github.com/qiexing/face-landmark-localization.

[2] MIT(2017) Face dataset, CBCL. http://poggio-lab.mit.edu/

[3] Ali, I., & Dailey, M. N. (2012, December). Multiple human tracking in high-density crowds. Image and Vision Computing, 30(12), 966–977.

[4] Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). "A discriminative feature learning approach for deep face recognition," In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), 14th European Conference on Computer Vision, amsterdam, the netherlands, october 11–14, 2016, 89 proceedings, part vii (pp. 499–515). Cham: Springer International Publishing.

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., et al. (2014). Going deeper with convolutions. Computing Research Repository

[6] Levi, G., & Hassncer, T. (2015). "Age and gender classification using convolutional neural networks," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 34–42.

[7] Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 9(12), 2170–2179.

[8] Gallagher, A., & Chen, T. (2009). "Understanding images of groups of people," IEEE Conference on Computer Vision and Pattern Recognition

[9] Rothe, R., Timofte, R. & Van Gool, L. (2016) International Journal of Computer Vision.

[10] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): "A complete facial expression dataset for action unit and emotion-specied expression," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 94–101.

[11] Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J., et al. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. Emotion, 10(6), 874–893.

[12] Lyons, M., & Akamatsu, S. (1998). "Coding facial expressions with GaborWavelets," third IEEE Conference on Automatic Face and Gesture Recognition, 200–205.