



Deep learning-based face analysis system for monitoring customer interest

Gozde Yolcu^{1,2} · Ismail Oztel^{1,2} · Serap Kazan¹ · Cemil Oz¹ · Filiz Bunyak²

Received: 1 August 2018 / Accepted: 28 April 2019 / Published online: 3 May 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

In marketing research, one of the most exciting, innovative, and promising trends is quantification of customer interest. This paper presents a deep learning-based system for monitoring customer behavior specifically for detection of interest. The proposed system first measures customer attention through head pose estimation. For those customers whose heads are oriented toward the advertisement or the product of interest, the system further analyzes the facial expressions and reports customers' interest. The proposed system starts by detecting frontal face poses; facial components important for facial expression recognition are then segmented and an iconized face image is generated; finally, facial expressions are analyzed using the confidence values of obtained iconized face image combined with the raw facial images. This approach fuses local part-based features with holistic facial information for robust facial expression recognition. With the proposed processing pipeline, using a basic imaging device, such as a webcam, head pose estimation, and facial expression recognition is possible. The proposed pipeline can be used to monitor emotional response of focus groups to various ideas, pictures, sounds, words, and other stimuli.

Keywords Facial expression recognition · Head pose estimation · Facial analysis · Customer monitoring · Convolutional neural network

1 Introduction

Quantification of customer interest and measuring relevant advertisement and product likability are important for advertising (Barretto 2017). The customer satisfaction survey, which is a traditional approach to quantify customer interest, has come to be considered as an invasive method in recent years (Barretto 2017). According to Barretto (2017), survey response rates have decreased because of the limited collection of data from specific demographic groups. The ongoing challenge is to find new ways to monitor customer interest (Barretto 2017).

Recording customer interest by a salesperson who observes customers' behavior during the advertisement watching or shopping phase is another approach. However,

this task requires specific skills for every salesperson, and each observer may interpret customer behaviors differently. In this regard, successful salesperson-customer interactions happen for only a few exceptionally tactful and talented salespersons (Menon and Dubé 2000). According to Misaglia et al. (2017), subjective emotional perceptions-based methods may not always work correctly to capture personal emotional state. In contrast, automatic measurements give a more accurate and reliable output. Consequently, there is a critical need to develop non-invasive, objective, and quantitative tools for monitoring customer interest.

In the marketing literature, a recent topic has been understanding human choices using different tools such as brain images (Langleben et al. 2009), EEG (Ohme et al. 2009; Cook et al. 2011), eye tracking (Wedel and Pieters 2000; Ungureanu et al. 2017), heart rate registration (Micu and Plummer 2010), and other strategies. Also, some researches have involved customer behavior classification (Popa et al. 2010; Liu et al. 2017, 2018) and customer facial analysis studies (Kasiran and Yahya 2007; Teixeira et al. 2012).

Estimating customer visual focus of attention is one of the approaches for monitoring customer interest. In the

✉ Filiz Bunyak
bunyak@missouri.edu

¹ Department of Computer Engineering, Sakarya University, 54050 Sakarya, Turkey

² Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

literature, several studies on visual focus of attention are based on head pose estimation (Ba and Odobez 2011; Riener and Sippl 2014; Sheikhi and Odobez 2015). Also, detecting customer emotions for marketing purposes is a growing and challenging research field (Karu-salo 2013). There is a strong effect of personal mood on the intuitive decision-making process (Karu-salo 2013). People in a positive mood think that things are going well, and the environment is safe. But in a negative mood, they feel that things are not going well, and there may be a threat of an event for which vigilance is required (Kahneman 2011). Consumer emotions and moods are important for marketers (Sahney 2011). An accurate understanding of these psychological states can guide a marketer to design a stimulus that leads to positive states (Sahney 2011). When a consumer is in a positive mood, he would be more receptive to the product, service, and the brand offering (Sahney 2011). Facial expression recognition can be useful in order to understand customers' feelings. Thus, making the right decisions that help to increase sales can be possible. According to Ekman and Friesen (1975), anger, fear, sadness, and disgust are categorized as negative emotions, whereas happiness is a positive emotion. Surprise is excluded because it is impossible to classify as a positive or negative emotion. According to Laros and Steenkamp (2005), happy consumers feel optimistic, hopeful, enthusiastic, encouraged, pleased, joyful, relieved, and thrilled.

This study presents a deep learning-based system that focuses on head-pose orientation and facial expression recognition for monitoring customer interest. The proposed system first detects the human face, then estimates the head pose orientation for detecting visual focus of attention. It is assumed that the camera is on the relevant advertisement or product so the frontal faces show visual focus of attention. If the detected face is oriented toward the advertisement

or product of interest, then the system starts to recognize the facial expression. The system collects customer facial expressions during a specific time that can be determined by experts. According to the collected facial expressions, the system decides whether the customer's mood is positive or negative. Figure 1 shows the proposed system pipeline. With the integration of a camera on the advertisement display or for online advertisement using a basic imaging device, such as a webcam, a head pose estimation, and facial expression recognition system is possible. Thus, a non-invasive, quantitative, and low-cost customer interest monitoring system can be used. Proposed system can be useful for determining some factors such as marketing campaigns, and other business strategies that can be interesting for customers. Salespeople can also change their marketing advertisement according to the feedback of customer behaviors. This platform can also allow effective responses from salespersons to customer emotions, facilitating greater satisfaction.

Main contributions of proposed study can be summarized in four points:

1. A non-invasive, objective and quantitative system for monitoring customer interest is proposed.
2. The proposed system does not save the customer face images thus it protects personal privacy. The system only processes current customer images obtained at the current moment and does not need to save them for processing. If there is a requirement to archive customer facial expression information, de-identified iconized face images can be archived. Iconized face images include facial expression data, while still protecting personal privacy.
3. A multi-task and 3-cascade CNN (Convolutional Neural Network) model is proposed. The third CNN uses

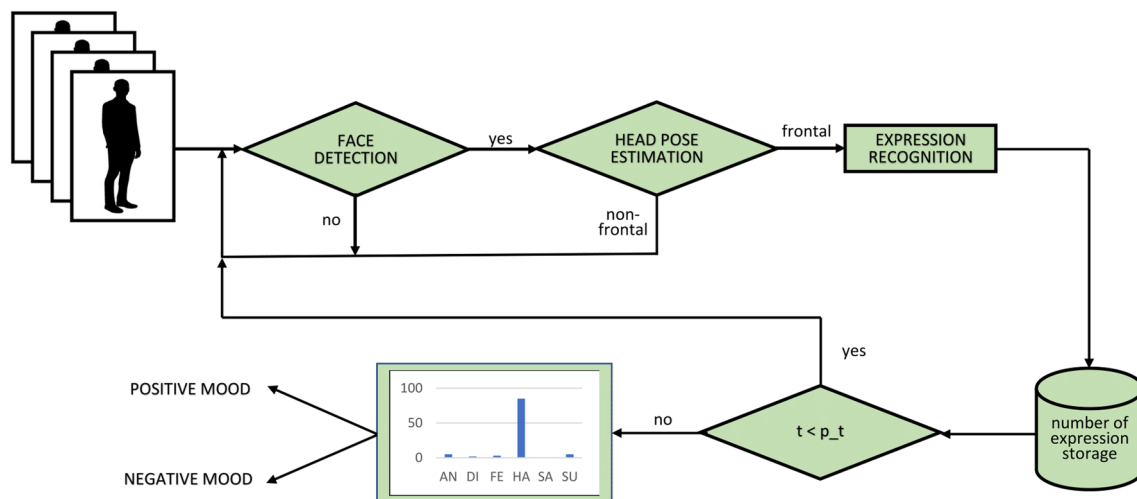


Fig. 1 Proposed processing pipeline (p_t: predefined time)

confidence values of iconized image combined with the raw facial images. Confidence values of iconized image include part-based information while raw facial images include holistic information. The fusion of part-based and holistic information allows improved facial expression recognition.

4. In the facial component segmentation stage, the system detects and localizes the facial components that are known to be important for facial expression recognition and produces an iconized image. In the facial expression recognition stage, using the confidence values of that iconized image as the input, the CNN allows guided training by forcing the earlier layers of the architecture to learn to detect and localize the important facial components.

2 Related works

Numerous papers in the literature report methods for solving real-world problems using images or videos. Runtime is important for video processing and various frameworks are being developed to increase the speed (Yan et al. 2014a, b; Bayrakdar et al. 2017). Object recognition (Yan et al. 2019), text detection (Yan et al. 2018a, b), facial expression recognition (Oztel et al. 2018b), head position estimation (Riener and Sippl 2014), etc. can be given as the example of real-world problems. In this section, the two main subjects of this study which are facial expression recognition and head pose estimation are examined.

2.1 Facial expression recognition

Automated facial expression recognition is an active research area that has growing applications such as avatar animation (Zalewski and Gong 2005), medical (Yolcu et al. 2017), robotic (De Carolis et al. 2017), traffic (Shaykha et al. 2015; Zhang and Hua 2015), smart environments (Jung-Bae Kim et al. 2013; Takahashi et al. 2015; Teyeb et al. 2015; Lozano-Monazor et al. 2017), and human–computer interaction (Terzis et al. 2013; Zhao et al. 2013; Niforatos and Karapanos 2015; Baldassarri et al. 2015; Benmohamed et al. 2015; Arigbabu et al. 2016; Samara et al. 2017). Automated facial expression recognition studies are usually based on the six universal facial expressions that are defined in the early works by Ekman and Friesen, namely: disgust, happiness, fear, anger, sadness, and surprise (Ekman and Friesen 1971).

Facial expression recognition studies can be classified as geometric-based and appearance-based methods (Pons and Masip 2017). Geometric-based methods focus on the features obtained from positional relationships between facial components (Pons and Masip 2017). Appearance-based features define the face texture which caused by expression

(Lee et al. 2012; Ghimire and Lee 2013; Perumal Ramalingam and Chandra Mouli 2018). The appearance-based methods that have been employed for facial expression recognition include local binary pattern (LBP) operator (Zhao and Pietikainen 2007), histogram of orientation gradients (HOG) (Ghimire and Lee 2012), principal component analysis (PCA) (Sobia et al. 2014), etc.

Recent trend in facial expression analysis is deep learning method. Lopes et al. (Lopes et al. 2017) used a CNN network with some image pre-processing techniques such as image rotation, face cropping, and intensity normalization in order to extract only specific features for expressions and studied the six basic expressions. Pitaloka et al. (Pitaloka et al. 2017) also worked on six basic expressions using a CNN. They applied pre-processing techniques like resizing, face detection, cropping, and data normalization, consisting of local normalization, global contrast normalization, and histogram equalization. Matsugu et al. (Matsugu et al. 2003) proposed a rule-based algorithm for smiling detection and face detection with CNNs.

2.2 Head pose estimation

Head pose estimation has been investigated for different purposes such as visual surveillance (Cheng Chen and Odobez 2012; Kang et al. 2014), driver attention (Murphy-Chutorian et al. 2007; Alioua et al. 2016), visual focus of attention (Ba and Odobez 2009; Riener and Sippl 2014), and robotic (Gaschler et al. 2012).

Head pose estimation systems are developed using different methods such as appearance-based, model-based, manifold embedding, and nonlinear regression methods (Patacchiola and Cangelosi 2017). Appearance-based methods are based on the comparison of a new head image to a set of head pose templates in order to find the most similar view. One disadvantage of appearance-based methods is that it is limited to estimating only discrete pose locations (Murphy-Chutorian and Trivedi 2009). Also, some templates require computationally expensive image comparisons (Murphy-Chutorian and Trivedi 2009). In model-based methods, geometric information or facial landmark locations are used for head pose estimation (Murphy-Chutorian and Trivedi 2009). The accuracy of model-based methods depends on the quantity and quality of the geometric cues extrapolated from the image. Manifold embedding methods are based on dimensionality reduction techniques, such as PCA (Murphy-Chutorian and Trivedi 2009). Head pose estimation is applied by projecting an image into a PCA subspace, and the results are compared to a set of embedded templates (McKenna and Gong 1998). The weakness of manifold embedding methods is that appearance variation can be affected by factors other than pose and identity, such as lighting (Murphy-Chutorian and Trivedi 2009). In nonlinear regression

methods, a labeled training set is used to create a nonlinear mapping from images to poses (Patacchiola and Cangelosi 2017). Nonlinear regression methods require a consistent dataset to train the parameters (Patacchiola and Cangelosi 2017).

CNNs are a nonlinear regression method (Patacchiola and Cangelosi 2017). CNNs show high-accuracy results for head pose orientation problems. Tran and Kim (2017) proposed a CNN method for head pose estimation. Their network consists of three CNNs, corresponding to three head pose types that are yaw, pitch, and roll. Mukherjee and Robertson (2015) combined the CNN-based classification and regression models to estimate the head pose. Xu and Kakadiaris (2017) used global and local features obtained from a CNN to estimate head pose and localize landmarks together. Patacchiola and Cangelosi (2017) used CNN and adaptive gradient methods for head pose estimation.

3 Methodology

3.1 Face detection

Each face image is cropped to reduce background information and retain only expression and head pose-specific features using Viola and Jones Algorithm (Viola and Jones 2004) that has shown outstanding performance in the literature. The Viola and Jones algorithm is based on three main ideas. The first idea is an integral image that allows for fast feature evaluations and decreases the complexity of feature detection for each frame (Viola and Jones 2004). The integral image at the location x, y position includes the sum of the pixels above and to the left of x, y with the inclusion of this coordinate (Viola and Jones 2004), which is shown in Eq. 1.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'), \quad (1)$$

where $ii(x, y)$ is an integral image and $i(x, y)$ is a raw image. The integral image can be calculated in one pass over the raw image using Eqs. 2 and 3 (Viola and Jones 2004).

$$s(x, y) = s(x, y - 1) + i(x, y) \quad (2)$$

$$ii(x, y) = ii(x - 1, y) + s(x, y), \quad (3)$$

where $s(x, y)$ is the cumulative row sum, $s(x, -1) = 0$, and $ii(-1, y) = 0$.

The second idea is creating a classifier by selecting a few important features using the AdaBoost algorithm (Viola and Jones 2004). Feature selection is applied with the AdaBoost learning algorithm. Each weak classifier is constrained to depend on only a single feature (Viola and Jones 2004). The weak learning algorithm is designed to select the single

rectangle feature that ideally separates the negative and positive examples (Viola and Jones 2004). The weak learners determine the optimum threshold classification function to misclassify a minimum number of examples for each feature (Viola and Jones 2004). A weak classifier is defined as follows (Viola and Jones 2004):

$$h_j(x) = \begin{cases} 1, & \text{if } (p_j x f_j(x) < p_j x \theta_j) \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where $h_j(x)$ is a weak classifier, f_j is feature, θ_j is a threshold, and p_j is parity indicating the direction of the inequality sign.

The last technique is a method of combining classifiers in a Cascade structure (Viola and Jones 2004). The combined structure is shown in Fig. 2. The cascade classifier improves the performance in terms of speed by focusing attention on important image regions (Viola and Jones 2004).

3.2 Convolutional neural network

Deep learning allows the efficient learning of complex non-linear relationships between raw input and output data through a training process (Choi et al. 2017). Recently, deep learning methods have shown outstanding performance in the pattern recognition and computer vision areas. CNN is one of the most popular type of deep learning in image analysis (Pitaloka et al. 2017; Al-Milaji et al. 2017; Oztel et al. 2017, 2018a; Yan et al. 2018c).

CNNs are based on three ideas: shared weights, local receptive fields, and spatial subsampling (Lawrence et al. 1997). They are built using convolution layers, pooling layers, dropout layers, activation layers, and fully connected layers. Convolutional layers convolve their inputs with a set of learned filters and generate feature maps. Convolutions are often used over more than one axis at a time. For example, if a two-dimensional image is used as an input, a two-dimensional kernel is often also used (Goodfellow et al. 2016). Because of flipping the kernel relative to the input, the commutative property of convolution arises (Goodfellow et al. 2016). Flipping adds some complexity to the CNN; the flipping operation occurs

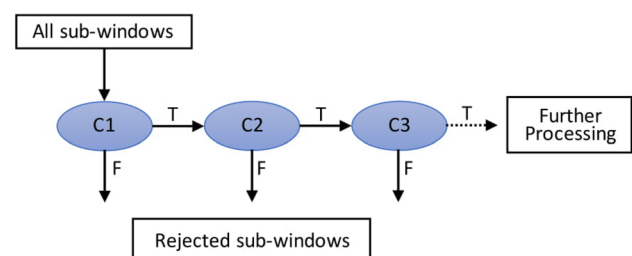


Fig. 2 Schematic description of the cascade structure (Viola and Jones 2004)

during inference and backpropagation of errors (Goodfellow et al. 2016). Many neural network libraries implement a cross-correlation function that is similar to convolution, without flipping the kernel but call it convolution (Goodfellow et al. 2016):

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n), \quad (5)$$

where I is an input image, K is a kernel, and S is the convolution result. Typically, the convolution operation is indicated with an asterisk.

The polling layer reduces the activation area size. Owing to a smaller activation size, the network requires fewer parameters to be learned in the next layers (Srinivas et al. 2016). According to Nair and Hinton (2010), CNNs using a non-linear function such as rectified linear unit (ReLU), tanh, or sigmoid non-linearity have been proven to train faster. In proposed study, ReLU was used as the activation function for all of the CNN structures.

$$ReLU(x) = \max(0, x), \quad (6)$$

where x is input of the activation function.

In order to reduce overfitting, dropout layer has been proven to be very effective (Hinton et al. 2012). Dropout can prevent the network from being too dependent on any neuron or any small neuron group. Owing to dropout, the network can be forced to be accurate even in the absence of certain information (Gu et al. 2017). Batch normalization can also reduce generalization error. In order to stabilize learning, batch normalization standardizes only the mean and variance of each unit, but it allows the relationships between units and the nonlinear statistics of a single unit to change (Goodfellow et al. 2016).

The fully connected layer connects to all activations in the previous layer. Softmax changes the predictions into non-negative values. Also, the softmax normalizes the values to get a probability distribution over classes (Goodfellow et al. 2016; Gu et al. 2017).

$$\text{softmax}(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)}, \quad (7)$$

where x_i is the input obtained from the fully connected layer. Those probabilistic predictions are used to compute the softmax loss (Goodfellow et al. 2016). Softmax loss is a commonly used loss function in CNNs (Gu et al. 2017).

In a feedforward neural network, using the input x and produced output y , information runs forward through the network (Goodfellow et al. 2016). The input x provides the initial information and propagates up to the next layers, and finally produces y (Goodfellow et al. 2016). During training, it continues until producing a scalar cost (Goodfellow et al. 2016).

Backpropagation algorithm works based on the chain rule (Goodfellow et al. 2016). In the backpropagation algorithm, the information obtained from cost flows backward through the network in order to compute the gradient (Goodfellow et al. 2016). As for a complete neural network framework with a loss L , the backpropagation computes the gradient of the parameter matrix W and the input x as follows (Wei et al. 2017):

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial W} \quad (8)$$

$$\frac{\partial L}{\partial x} = \frac{\partial y}{\partial x} \frac{\partial L}{\partial y} \quad (9)$$

Proposed system for customer interest monitoring performs the three learning stages using CNNs. The system starts by detecting frontal faces using CNN-1. Then, using CNN-2, the frontal images are segmented in order to keep facial components that include very important and critical facial expression features (Lin Zhong et al. 2012). Finally, in CNN-3 the system classifies facial expressions, using fully connected layer confidence values of CNN-2 combined with the raw facial images. Figure 3 shows the proposed CNN architecture.

3.2.1 CNN for head pose orientation

Determining whether the customer is looking at the relevant advertisement or product is the first step in the proposed system. The coarsest level head pose estimation that classifies a head in frontal versus non-frontal profile is sufficient for proposed system. Frontal faces are transferred to CNN-2 for facial component segmentation stage and non-frontal faces are ignored.

The CNN-1 is trained to estimate head pose angle (0° , 45° , 90° , 135° , 180°) in terms of yaw movements; which are shown in Fig. 4, given the raw facial image. The CNN-1 structure operates on a resized whole face image ($64 \times 64 \times 3$ pixel). The CNN-1 structure includes ten layers: five convolutional layers (one layer with $64 \ 5 \times 5 \times 3$, two layers with $32 \ 5 \times 5$, one layer with $64 \ 5 \times 5$, and one layer with $64 \ 4 \times 4$ filters), four pooling layers, and one fully connected layer.

3.2.2 CNN for facial component segmentation

The purpose of the CNN-2 is segmenting facial components that are mouth, eye, and eyebrow regions from the rest of the image. Facial component segmentation is formulated as a binary classification problem of facial component versus background. Before the training stage, the original raw images and corresponding training masks are partitioned into 16×16 non-overlapping blocks. Blocks having more

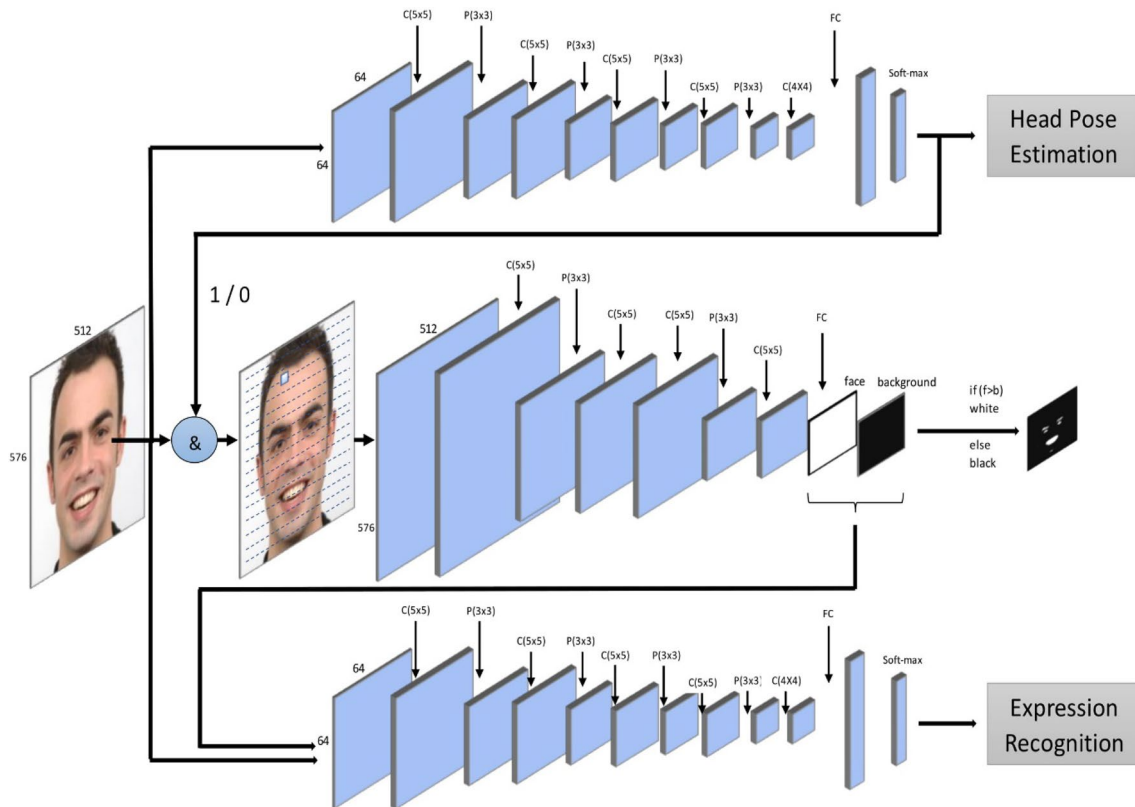


Fig. 3 Proposed CNN system



Fig. 4 Head pose samples in terms of yaw movement

than 80% of either facial component or background pixels are assigned as the majority class. The remaining mixed class blocks are ignored during training. The threshold value 80% was determined according to our experiments. Figure 5 shows the building block steps for an image. The output of fully connected layer are two channels and one of them contains facial components confidence values and other channel contains background confidence values. Iconized face images are obtained according to the higher components confidence value in fully connected layer. Also, the confidence values of two channels are transferred to input of the CNN-3 for performing guided image training and stronger facial expression recognition. Testing is performed using a sliding window described in (Shelhamer et al. 2016).

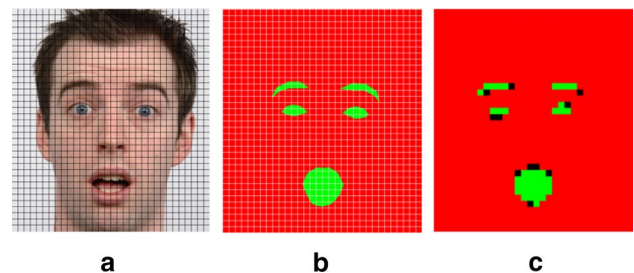


Fig. 5 Building non-overlapping blocks steps **a** divided raw image, **b** divided ground truth, **c** built and labeled blocks (green: facial component, red: background, black: ignored class)

The CNN-2 structure consists of seven layers: four convolutional layers (one layer with $16 \times 5 \times 5 \times 3$, one layer with $16 \times 5 \times 5$, one layer with $32 \times 5 \times 5$, and one layer with $32 \times 4 \times 4$ filters), two pooling layers, and one fully connected layer.

3.2.3 CNN for facial expression classification

CNN-3 is trained to classify facial expression using the raw facial image (3-channel) combined with the confidence values of a corresponding iconized image (2-channel) that is obtained from the CNN-2. Using the confidence values of CNN-2 allows forcing the network to learn to detect facial components that are important for facial expression recognition, thus guided image training and stronger facial expression recognition are provided. While the confidence values of iconized images include part based information, raw facial images include holistic information. Using holistic and part based information together also improves facial expression recognition accuracy. The CNN-3 structure operates on a resized whole face image, similar to CNN-1.

The CNN-3 structure includes ten layers: five convolutional layers (one layer with $64 \times 5 \times 5 \times 3$, two layers with $32 \times 5 \times 5$, one layer with $64 \times 5 \times 5$, and one layer with $64 \times 4 \times 4$ filters), four pooling layers, and one fully connected layer. Table 1 shows the detailed layer information of the proposed CNN structures.

4 Experimental results

Proposed 3-stage CNN architecture was implemented using the MatConvNet toolbox (Vedaldi and Lenc 2015). The network was trained and tested on the Radboud Face Database (RaFD) (Langner et al. 2010).

Table 1 Detailed layer information for the proposed CNN structures. Information on pooling, ReLU, and dropout layers are not listed to simplify the table

Structure	Layer	Kernel	Filter	Output
CNN-1 (Head pose estimation input: 64×64)	conv1	5×5	64	$64 \times 64 \times 64$
	conv2	5×5	32	$32 \times 32 \times 32$
	conv3	5×5	32	$16 \times 16 \times 32$
	conv4	5×5	64	$8 \times 8 \times 64$
	conv5	4×4	64	$1 \times 1 \times 64$
CNN-2 (Facial component segmentation input: 16×16)	conv1	5×5	16	$16 \times 16 \times 16$
	conv2	5×5	16	$8 \times 8 \times 16$
	conv3	5×5	32	$4 \times 4 \times 32$
	conv4	4×4	32	$1 \times 1 \times 32$
CNN-3 (Facial expression recognition input: 64×64)	conv1	5×5	64	$64 \times 64 \times 64$
	conv2	5×5	32	$32 \times 32 \times 32$
	conv3	5×5	32	$16 \times 16 \times 32$
	conv4	5×5	64	$8 \times 8 \times 64$
	conv5	4×4	64	$1 \times 1 \times 64$

Table 2 Confusion matrix for head pose estimation on RaFD

Actual	Predicted				
	0	45	90	135	180
0	99.88	0.012	0	0	0
45	0	100	0	0	0
90	0	0	100	0	0
135	0	0	0	100	0
180	0	0	0	0.037	99.63
Average: 99.90%					

In this study, 8040 images were used for head pose estimation, and 1206 frontal images (for anger, happy, disgust, surprise, sad and fear) were used for facial expression recognition from RaFD. Both the head pose estimation set and the facial expression recognition set were divided into two equal parts for training and testing. Testing and training sets do not include any image belonging to the same person.

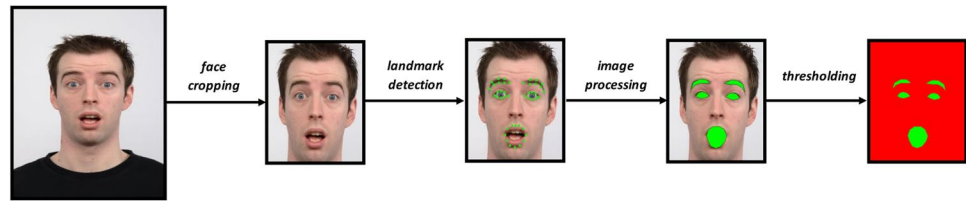
In the head pose estimation section, the system was trained using raw facial images. Although classification of the head pose as frontal versus non-frontal profile is sufficient for proposed system, it was trained to classify five head pose angles in terms of yaw movement. The confusion matrix of head pose estimation for RaFD is shown in Table 2.

In order to evaluate the proposed system, Karolinska Directed Emotional Face (KDEF) database (Lundqvist et al. 1998) was also used. Because KDEF includes both head pose angle and facial expression labels of the face images, it is suitable to evaluate proposed system performance. KDEF includes 4900 images with 5 different viewing angles and 7 facial expression. For head pose estimation 4900 images and for facial expression recognition 840 frontal images were used. These sets were grouped into two subsets for training (90%) and testing (10%) because they contain few images. The confusion matrix of head pose estimation of KDEF is given in Table 3.

In order to obtain training masks for the facial component segmentation pipeline, first, each image was cropped to include facial regions of eyebrows, eyes, and mouth. Then,

Table 3 Confusion matrix (%) for head pose estimation on KDEF

Actual	Predicted				
	0	45	90	135	180
0	97.96	2.04	0	0	0
45	0	100	0	0	0
90	0	0	100	0	0
135	0	0	0	100	0
180	0	0	0	1.02	98.98
Average: 99.39%					

Fig. 6 Mask generation flow**Table 4** Facial expression classification confusion matrix (%) for the proposed 5-channel (raw + confidence values) cascaded CNN architecture and average recognition accuracy (%) of positive and negative expressions on RaFD

Actual \ Predicted						
	Anger	Disgust	Fear	Happy	Sad	Surprised
Anger	97.98	0	0	0	2.02	0
Disgust	1.01	96.97	0	0	0	2.02
Fear	0	0	86.87	0	11.11	2.02
Happy	0	4.04	0	94.95	1.01	0
Sad	4.04	2.02	3.03	0	90.91	0
Surprised	0	0	0	0	0	100
Average: 94.61%						

Positive Exp.	Negative Exp.
94.95	93.18

facial keypoints were detected using the Face++ toolkit (Face++ 2017). The toolkit can detect 83 keypoints on the human face. 45 keypoints were used to generate training masks. Each keypoint was linked to fit a polygon. Training masks were obtained filling these polygons. Finally, thresholding was applied on these images and a final training set was obtained. The generating flow can be seen in Fig. 6.

In the facial expression recognition section, the system was trained using a 5-channel input (combining 3-channel raw facial image with 2-channel iconized face confidence values). Table 4 shows the confusion matrix of facial expression recognition for RaFD. When the accuracy of positive and negative expressions are recalculated from the confusion matrix, the average accuracy of positive expression (happy) is 94.95% and negative expression (anger, disgust, fear and sad) is 93.18%. Also, the confusion matrix of facial expression recognition for KDEF is given in Table 5.

In order to better show the benefits of the recognition system using 5-channel input, the proposed CNN-3 was trained using only raw facial image and only iconized face image as an input and compared the results. Using 5-channel input, facial expression recognition achieved the best accuracy on RaFD (94.61%), while recognition using 1-channel iconized face image outperformed recognition using a 3-channel raw facial image (89.99% vs. 83.33%). The RaFD database has unnatural expressions, so some errors can occur in our system. Figure 7 shows some visual samples of our error cases.

The performance of the proposed system was compared to facial expression recognition studies in the literature. Table 6 shows the comparison results. Facial expression recognition results using 3-channel raw image, 1-channel iconized image, and 5-channel combined image results for RaFD are illustrated in the table.

Time periods for three type of CNN and total execution time are given in Table 7 with approximate values for one

Table 5 Facial expression recognition confusion matrix (%) on KDEF database for 5-channel input data

Actual \ Predicted						
	Anger	Disgust	Fear	Happy	Sad	Surprised
Anger	85.72	7.14	0	0	7.14	0
Disgust	0	92.86	0	0	7.14	0
Fear	0	0	85.72	0	7.14	7.14
Happy	0	0	0	100	0	0
Sad	7.14	0	0	0	92.86	0
Surprised	0	0	0	0	0	100
Average: 92.86%						

Positive Exp.	Negative Exp.
100	89.29











		PREDICTED					
ACTUAL	EXPRESSIONS	ANGER	DISGUST	FEAR	HAPPY	SAD	SURPRISED
	ANGER	✓					
	DISGUST		✓				
	FEAR			✓			
	HAPPY				✓		
	SAD					✓	
	SURPRISED						✓

Fig. 7 Some visual samples of our error cases

image. As can be seen in the table, a large part of the time is taken by the segmentation process. The proposed pipeline is designed for 576×512 resolution images where facial components are distinct. While the inputs of CNN-1 and CNN-3 are 64×64 , CNN-2 runs with 576×512 images. So, CNN-2 needs extra time to handle the input images. If the segmentation process is abandoned to shorten the time, the performance will decrease by $\sim 11\%$ (83.33% for raw image input vs. 94.61% for five channel data).

Figure 8 shows two sample frames during runtime. As can be seen in this figure, if a person lacks a frontal view according to the camera, the system does not generate any expressions. Otherwise, the system is able to detect the expressions.

5 Conclusion

This paper presents, a novel deep learning system for automated head pose estimation and facial expression recognition. This study is the first step toward a noninvasive, objective, quantitative system for monitoring customer interest. The proposed system consists of a cascade three CNN structures. The task of the first CNN is head pose estimation. The second CNN structure was trained to segment facial components. The third CNN was trained to perform facial expression classification. The proposed last two-step (CNN-2, CNN-3) allows guided image classification and integration of part-based and holistic information. Experimental test results achieved 99.90% accuracy for head pose estimation and 94.61% accuracy facial expression recognition on RaFD dataset. When the accuracy of positive and negative

Table 7 Execution times for the CNN structures

CNN structure	Task	Time (s)
CNN-1	Head pose estimation	~ 0.064
CNN-2	Facial component segmentation	~ 0.29
CNN-3	Facial expression recognition	~ 0.084
Total execution time = 0.41 (sec. for one image)		

Table 6 Facial expression recognition accuracies for the proposed system and other studies

Methods	Database	Accuracy (%)
HoG + NNE (Ali et al. 2016)	RaFD, TFEID, JAFFE	93.75
Surf Boosting (Rao et al. 2015)	RaFD	90.64
Gabor F. + GLCM ^a (Li et al. 2015)	RaFD	88.41
Facial Components Detection + Fuzzy (Ilbeygi and Shah-Hosseini 2012)	RaFD	93.96
Facial Components Detection + KNN (Ilbeygi and Shah-Hosseini 2012)	RaFD	75.61
Viola and Jones + AAM + ANN ^a (Bijlstra and Dotsch 2011)	RaFD	89.55
DFD + KNN ^a (Sun et al. 2017)	KDEF	82.24
LSiBP + SVM ^a (Santra and Mukherjee 2016)	KDEF	84.07
HoG + SVM ^a (Liew and Yairi 2015)	KDEF	80.20
HoG + AdaBoost ^a (Liew and Yairi 2015)	KDEF	75.20
HoG + AdaBoost + SVM* (Liew and Yairi 2013)	KDEF	87.20
Proposed method	RaFD (3-channel raw image)	83.33
	RaFD (1-channel iconized)	89.99
	RaFD (5-channel combine)	94.61
	KDEF (5-channel combine)	92.86

^aShows the studies used seven expressions



Fig. 8 Some samples from the system

emotions is recomputed from Table 4, the average accuracy is 93.73%.

The proposed system can facilitate the quantification of customer interest and the measurement of relevant advertisements and product likability. It can also allow identification of business strategies that increase sales. Marketing campaigns can change their strategies according to the customer feedback.

In future work, the proposed system will be extended to track human faces; then the system will follow facial expressions of the indexed person during a period of time and subsequently will determine the level of interest for each indexed person.

Acknowledgements This research is supported by The Scientific and Technological Research Council of Turkey (TUBITAK-BIDEB 2214/A) and Sakarya University Scientific Research Projects Unit (Project Number: 2015-50-02-039).

References

- Ali G, Iqbal MA, Choi T-S (2016) Boosted NNE collections for multi-cultural facial expression recognition. *Pattern Recognit* 55:14–27. <https://doi.org/10.1016/j.patcog.2016.01.032>
- Alioua N, Amine A, Rogozan A et al (2016) Driver head pose estimation using efficient descriptor fusion. *EURASIP J Image Video Process* 2016:2. <https://doi.org/10.1186/s13640-016-0103-z>
- Al-Milaji Z, Ersoy I, Hafiane A et al (2017) Integrating segmentation with deep learning for enhanced classification of epithelial and stromal tissues in H&E images. *Pattern Recognit Lett*. <https://doi.org/10.1016/j.patrec.2017.09.015>
- Arigbabu OA, Mahmood S, Ahmad SMS, Arigbabu AA (2016) Smile detection using hybrid face representation. *J Ambient Intell Humaniz Comput* 7:415–426. <https://doi.org/10.1007/s12652-015-0333-4>
- Ba SO, Odobez J-M (2009) Recognizing visual focus of attention from head pose in natural meetings. *IEEE Trans Syst Man, Cybern Part B* 39:16–33. <https://doi.org/10.1109/TSMCB.2008.927274>
- Ba SO, Odobez J (2011) Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Trans Pattern Anal Mach Intell* 33:101–116. <https://doi.org/10.1109/TPAMI.2010.69>
- Baldassarri S, Hupont I, Abadía D, Cerezo E (2015) Affective-aware tutoring platform for interactive digital television. *Multimed Tools Appl* 74:3183–3206. <https://doi.org/10.1007/s11042-013-1779-z>
- Barretto AM (2017) Application of facial expression studies on the field of marketing. FEELab Science Books, Porto
- Bayrakdar S, Akgün D, Yücedağ İ (2017) An accelerated approach for facial expression analysis on video files. *Pamukkale Univ J Eng Sci* 23:602–613. <https://doi.org/10.5505/pajes.2016.00908>
- Benmohamed A, Neji M, Ramdani M et al (2015) Feast: face and emotion analysis system for smart tablets. *Multimed Tools Appl* 74:9297–9322. <https://doi.org/10.1007/s11042-014-2082-3>
- Bijlstra G, Dotsch R (2011) FaceReader 4 emotion classification performance on images from the Radboud Face Database
- Cheng Chen, Odobez J (2012) We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In: *IEEE conference on computer vision and pattern recognition*, pp 1544–1551
- Choi J-S, Lee W-H, Lee J-H et al (2017) Deep Learning Based NLOS Identification with Commodity WLAN Devices. *IEEE Trans Veh Technol*. <https://doi.org/10.1109/tvt.2017.2780121>
- Cook IA, Warren C, Pajot SK et al (2011) Regional brain activation with advertising images. *J Neurosci Psychol Econ* 4:147–160. <https://doi.org/10.1037/a0024809>
- De Carolis B, Ferilli S, Palestro G (2017) Simulating empathic behavior in a social assistive robot. *Multimed Tools Appl* 76:5073–5094. <https://doi.org/10.1007/s11042-016-3797-0>
- Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. *J Pers Soc Psychol* 17:124–129. <https://doi.org/10.1037/h0030377>
- Ekman P, Friesen WV (1975) *Unmasking the face: a guide to recognising emotions from facial clues*. Consulting Psychologists Press, New Jersey (Prentice Hall [Palo Alto, CA])
- Face ++ SC (2017) Face ++ Cognitive Services. <https://www.faceplusplus.com/>. Accessed 12 Nov 2017
- Gaschler A, Jentzsch S, Giuliani M, et al (2012) Social behavior recognition using body posture and head pose for human-robot interaction. In: *IEEE/RSJ international conference on intelligent robots and systems*, pp 2128–2133
- Ghimire D, Lee J (2012) Histogram of orientation gradient feature-based facial expression classification using bagging with extreme learning machine. *Adv Sci Lett* 17:156–161. <https://doi.org/10.1166/asl.2012.4257>
- Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class Adaboost and support vector machines. *Sensors* 13:7714–7734. <https://doi.org/10.3390/s130607714>

- Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
- Gu J, Wang Z, Kuen J et al (2017) Recent advances in convolutional neural networks. *Pattern Recognit*. <https://doi.org/10.1016/j.patco.2017.10.013>
- Hinton GE, Srivastava N, Krizhevsky A, et al (2012) Improving neural networks by preventing co-adaptation of feature detectors. *arXiv*
- Ilbeygi M, Shah-Hosseini H (2012) A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. *Eng Appl Artif Intell* 25:130–146. <https://doi.org/10.1016/j.engappai.2011.07.004>
- Jung-Bae Kim, Youngkyoo Hwang, Won-Chul Bang, et al (2013) Real-time realistic 3D facial expression cloning for smart TV. In: IEEE international conference on consumer electronics (ICCE), pp 240–241
- Kahneman D (2011) Thinking fast and slow. Allen Lane, Bristol
- Kang S-K, Chung K-Y, Lee J-H (2014) Development of head detection and tracking systems for visual surveillance. *Pers Ubiquitous Comput* 18:515–522. <https://doi.org/10.1007/s00779-013-0668-9>
- Karu-salo I (2013) The effect of universal emotions on customer behaviour. Estonian Business School, Tallinn
- Kasiran Z, Yahya S (2007) Facial expression as an implicit customers' feedback and the challenges. In: Computer graphics, imaging and visualisation, IEEE, pp 377–381
- Langleben DD, Loughhead JW, Ruparel K et al (2009) Reduced prefrontal and temporal processing and recall of high “sensation value” ads. *Neuroimage* 46:219–225. <https://doi.org/10.1016/j.neuroimage.2008.12.062>
- Langner O, Dotsch R, Bijlstra G et al (2010) Presentation and validation of the Radboud Faces Database. *Cogn Emot* 24:1377–1388. <https://doi.org/10.1080/02699930903485076>
- Laros FJM, Steenkamp J-BEM (2005) Emotions in consumer behavior: a hierarchical approach. *J Bus Res* 58:1437–1445. <https://doi.org/10.1016/j.jbusres.2003.09.013>
- Lawrence S, Giles CL, Tsoi Ah Chung, Back AD (1997) Face recognition: a convolutional neural-network approach. *IEEE Trans Neural Netw* 8:98–113. <https://doi.org/10.1109/72.554195>
- Lee C-C, Shih C-Y, Lai W-P, Lin P-C (2012) An improved boosting algorithm and its application to facial emotion recognition. *J Ambient Intell Humaniz Comput* 3:11–17. <https://doi.org/10.1007/s12652-011-0085-8>
- Li R, Liu P, Jia K, Wu Q (2015) Facial Expression Recognition under Partial Occlusion Based on Gabor Filter and Gray-Level Cooccurrence Matrix. In: International conference on computational intelligence and communication networks (CICN). IEEE, pp 347–351
- Liew CF, Yairi T (2013) A comparison study of feature spaces and classification methods for facial expression recognition. In: IEEE international conference on robotics and biomimetics (ROBIO), pp 1294–1299
- Liew CF, Yairi T (2015) Facial expression recognition and analysis: a comparison study of feature descriptors. *IPSI Trans Comput Vis Appl* 7:104–120. <https://doi.org/10.2197/ipsjtcva.7.104>
- Lin Zhong, Qingshan Liu, Peng Yang, et al (2012) Learning active facial patches for expression analysis. In: IEEE conference on computer vision and pattern recognition, pp 2562–2569
- Liu J, Gu Y, Kamijo S (2017) Customer behavior classification using surveillance camera for marketing. *Multimed Tools Appl* 76:6595–6622. <https://doi.org/10.1007/s11042-016-3342-1>
- Liu J, Gu Y, Kamijo S (2018) Integral customer pose estimation using body orientation and visibility mask. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-018-5839-2>
- Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognit* 61:610–628. <https://doi.org/10.1016/j.patcog.2016.07.026>
- Lozano-Monator E, López MT, Vigo-Bustos F, Fernández-Caballero A (2017) Facial expression recognition in ageing adults: from lab to ambient assisted living. *J Ambient Intell Humaniz Comput* 8:567–578. <https://doi.org/10.1007/s12652-017-0464-x>
- Lundqvist D, Flykt A, Öhman A (1998) The Karolinska Directed Emotional Faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, Solna
- Matsugu M, Mori K, Mitari Y, Kaneda Y (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw* 16:555–559. [https://doi.org/10.1016/S0893-6080\(03\)00115-1](https://doi.org/10.1016/S0893-6080(03)00115-1)
- McKenna S, Gong S (1998) Real-time face pose estimation. *Real-Time Imaging* 4:333–347. [https://doi.org/10.1016/S1077-2014\(98\)90003-1](https://doi.org/10.1016/S1077-2014(98)90003-1)
- Menon K, Dubé L (2000) Ensuring greater satisfaction by engineering salesperson response to customer emotions. *J Retail* 76:285–307. [https://doi.org/10.1016/S0022-4359\(00\)00034-8](https://doi.org/10.1016/S0022-4359(00)00034-8)
- Micu AC, Plummer JT (2010) Measurable emotions: how television ads really work. *J Advert Res* 50:137–153. <https://doi.org/10.2501/S0021849910091300>
- Missaglia AL, Oppo A, Mauri M et al (2017) The impact of emotions on recall: an empirical study on social ads. *J Consum Behav* 16:424–433. <https://doi.org/10.1002/cb.1642>
- Mukherjee SS, Robertson NM (2015) Deep head pose: gaze-direction estimation in multimodal video. *IEEE Trans Multimed* 17:2094–2107. <https://doi.org/10.1109/TMM.2015.2482819>
- Murphy-Chutorian E, Trivedi MM (2009) Head pose estimation in computer vision: a survey. *IEEE Trans Pattern Anal Mach Intell* 31:607–626. <https://doi.org/10.1109/TPAMI.2008.106>
- Murphy-Chutorian E, Doshi A, Trivedi MM (2007) Head pose estimation for driver assistance systems: a robust algorithm and experimental evaluation. In: IEEE intelligent transportation systems conference, pp 709–714
- Nair V, Hinton GE (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th international conference on international conference on machine learning. Omnipress, USA, pp 807–814
- Niforatos E, Karapanos E (2015) EmoSnaps: a mobile application for emotion recall from facial expressions. *Pers Ubiquitous Comput* 19:425–444. <https://doi.org/10.1007/s00779-014-0777-0>
- Ohme R, Reykowska D, Wiener D, Choromanska A (2009) Analysis of neurophysiological reactions to advertising stimuli by means of EEG and galvanic skin response measures. *J Neurosci Psychol Econ* 2:21–31. <https://doi.org/10.1037/a0015462>
- Oztel I, Yolcu G, Ersoy I, et al (2017) Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In: IEEE international conference on bioinformatics and biomedicine (BIBM). pp 1195–1200
- Oztel I, Yolcu G, Ersoy I et al (2018a) Deep learning approaches in electron microscopy imaging for mitochondria segmentation. *Int J Data Min Bioinform* 21:91. <https://doi.org/10.1504/IJDMB.2018.096398>
- Oztel I, Yolcu G, Oz C et al (2018b) iFER: facial expression recognition using automatically selected geometric eye and eyebrow features. *J Electron Imaging* 27:1. <https://doi.org/10.1117/1.JEI.27.2.023003>
- Patacchiola M, Cangelosi A (2017) Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods. *Pattern Recognit* 71:132–143. <https://doi.org/10.1016/j.patco.2017.06.009>
- Perumal Ramalingam S, Chandra Mouli PVSSR (2018) Modified dimensionality reduced local directional pattern for facial analysis. *J Ambient Intell Humaniz Comput* 9:725–737. <https://doi.org/10.1007/s12652-017-0473-9>
- Pitaloka DA, Wulandari A, Basaruddin T, Liliana DY (2017) Enhancing CNN with preprocessing stage in automatic

- emotion recognition. *Proc Comput Sci* 116:523–529. <https://doi.org/10.1016/j.procs.2017.10.038>
- Pons G, Masip D (2017) Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Trans Affect Comput*. <https://doi.org/10.1109/taffc.2017.2753235>
- Popa M, Rothkrantz L, Yang Z, et al (2010) Analysis of shopping behavior based on surveillance system. In: *IEEE international conference on systems, man and cybernetics*, pp 2512–2519
- Rao Q, Qu X, Mao Q, Zhan Y (2015) Multi-pose facial expression recognition based on SURF boosting. In: *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, pp 630–635
- Riener A, Sippl A (2014) Head-pose-based attention recognition on large public displays. *IEEE Comput Graph Appl* 34:32–41. <https://doi.org/10.1109/MCG.2014.9>
- Sahney S (2011) Module-6 consumer behavior. Vinod Gupta School of Management, Kharagpur, pp 1–24
- Samara A, Galway L, Bond R, Wang H (2017) Affective state detection via facial expression analysis within a human–computer interaction context. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-017-0636-8>
- Santra B, Mukherjee DP (2016) Local saliency-inspired binary patterns for automatic recognition of multi-view facial expression. In: *IEEE international conference on image processing (ICIP)*. pp 624–628
- Shaykha I, Menkara A, Nahas M, Ghantous M (2015) FEER: Non-intrusive facial expression and emotional recognition for driver's vigilance monitoring. In: *57th international symposium ELMAR (ELMAR)*. IEEE, pp 233–237
- Sheikhi S, Odobez J-M (2015) Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognit Lett* 66:81–90. <https://doi.org/10.1016/j.patrec.2014.10.002>
- Shelhamer E, Long J, Darrell T (2016) Fully convolutional networks for semantic segmentation. *Cogn Emot* 24:1377–1388
- Sobia MC, Brindha V, Abudhahir A (2014) Facial expression recognition using PCA based interface for wheelchair. In: *International conference on electronics and communication systems (ICECS)*. IEEE, pp 1–6
- Srinivas S, Sarvadevabhatla RK, Mopuri KR et al (2016) A taxonomy of deep convolutional neural nets for computer vision. *Front Robot AI*. <https://doi.org/10.3389/frobt.2015.00036>
- Sun Z, Hu Z-P, Wang M, Zhao S-H (2017) Discriminative feature learning-based pixel difference representation for facial expression recognition. *IET Comput Vis* 11:675–682. <https://doi.org/10.1049/iet-cvi.2016.0505>
- Takahashi M, Clippingdale S, Naemura M, Shibata M (2015) Estimation of viewers' ratings of TV programs based on behaviors in home environments. *Multimed Tools Appl* 74:8669–8684. <https://doi.org/10.1007/s11042-014-2352-0>
- Teixeira T, Wedel M, Pieters R (2012) Emotion-induced engagement in internet video advertisements. *J Mark Res* 49:144–159. <https://doi.org/10.1509/jmr.10.0207>
- Terzis V, Moridis CN, Economides AA (2013) Measuring instant emotions based on facial expressions during computer-based assessment. *Pers Ubiquitous Comput* 17:43–52. <https://doi.org/10.1007/s00779-011-0477-y>
- Teyeb I, Jemai O, Zaied M, Ben Amar C (2015) Vigilance measurement system through analysis of visual and emotional driver's signs using wavelet networks. In: *15th international conference on intelligent systems design and applications (ISDA)*. IEEE, pp 140–147
- Tran BH, Kim Y-G (2017) Deep head pose estimation for faces in the wild and its transfer learning. In: *Seventh international conference on information science and technology (ICIST)*. IEEE, pp 187–193
- Ungureanu F, Lupu RG, Cadar A, Prodan A (2017) Neuromarketing and visual attention study using eye tracking techniques. In: *2017 21st international conference on system theory, control and computing (ICSTCC)*. pp 553–557
- Vedaldi A, Lenc K (2015) MatConvNet. In: *Proceedings of the 23rd ACM international conference on multimedia—MM'15*. ACM Press, New York, New York, USA, pp 689–692
- Viola P, Jones MJ (2004) Robust real-time face detection. *Int J Comput Vis* 57:137–154. <https://doi.org/10.1023/B:VISI.0000013087.49260.fb>
- Wedel M, Pieters R (2000) Eye fixations on advertisements and memory for brands: a model and findings. *Mark Sci* 19:297–312. <https://doi.org/10.1287/mksc.19.4.297.11794>
- Wei B, Sun X, Ren X, Xu J (2017) Minimal Effort Back Propagation for Convolutional Neural Networks. *Comput Res Repos*
- Xu X, Kakadiaris IA (2017) Joint head pose estimation and face alignment framework using global and local CNN features. In: *12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. pp 642–649
- Yan C, Zhang Y, Xu J et al (2014a) A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Process Lett* 21:573–576. <https://doi.org/10.1109/LSP.2014.2310494>
- Yan C, Zhang Y, Xu J et al (2014b) Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Trans Circuits Syst Video Technol* 24:2077–2089. <https://doi.org/10.1109/TCSVT.2014.2335852>
- Yan C, Xie H, Chen J et al (2018a) A fast Uyghur text detector for complex background images. *IEEE Trans Multimed* 20:3389–3398. <https://doi.org/10.1109/TMM.2018.2838320>
- Yan C, Xie H, Liu S et al (2018b) Effective Uyghur language text detection in complex background images for traffic prompt identification. *IEEE Trans Intell Transp Syst* 19:220–229. <https://doi.org/10.1109/TITS.2017.2749977>
- Yan C, Xie H, Yang D et al (2018c) Supervised hash coding with deep neural network for environment perception of intelligent vehicles. *IEEE Trans Intell Transp Syst* 19:284–295. <https://doi.org/10.1109/TITS.2017.2749965>
- Yan C, Li L, Zhang C et al (2019) Cross-modality bridging and knowledge transferring for image understanding. *IEEE Trans Multimed*. <https://doi.org/10.1109/tmm.2019.2903448>
- Yolcu G, Oztel I, Kazan S, et al (2017) Deep learning-based facial expression recognition for monitoring neurological disorders. In: *IEEE international conference on bioinformatics and biomedicine (BIBM)*. pp 1652–1657
- Zalewski L, Shaogang Gong (2005) 2D Statistical Models of Facial Expressions for Realistic 3D Avatar Animation. In: *IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. pp 217–222
- Zhang Y, Hua C (2015) Driver fatigue recognition based on facial expression analysis using local binary patterns. *Opt Int J Light Electron Opt* 126:4501–4505. <https://doi.org/10.1016/j.ijleo.2015.08.185>
- Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans Pattern Anal Mach Intell* 29:915–928. <https://doi.org/10.1109/TPAMI.2007.1110>
- Zhao Y, Wang X, Goubran M et al (2013) Human emotion and cognition recognition from body language of the head using soft computing techniques. *J Ambient Intell Humaniz Comput* 4:121–140. <https://doi.org/10.1007/s12652-012-0107-1>