Analysis of Eye Fixations During Emotion Recognition in Talking Faces

Houwei Cao
Department of Computer Science
New York Institute of Technology
New York, United States
hcao02@nyit.edu

Forest Elliott
Department of Computer Science
Oberlin College
Oberlin, United States
forest.elliott@oberlin.edu

***Abstract*—Research on emotion recognition from cues expressed in facial expression has a long-standing tradition. In this study, we investigate human's visual attention and fixation patterns when identifying six basic emotions on expressive talking faces. Stimuli for the current experiments consisted of 92 video clips of facial expression during talking. The whole experiments were divided into two sessions. The video stimuli in the first session were presented in random order across different face identities, while in the second session the video from the same face identity were be played sequentially. The participants' eye movements were recorded by the Tobii X3-120 screen-based eye-tracking system. We defined a set of area-of-interest (AOI) regions, including 4 AOIs of general face areas and 12 AOIs related to specific Action Units (AUs) involved in the coding of the six basic emotions. The gaze pattern analysis was done by looking at the fixation time on this predetermined set of AOIs. Based on the ANOVA analysis, we did not find significant differences in mean fixation time on any AOI for discriminating the six basic emotions, but a subset of significant AOIs was found when we sectioned the six basic emotions into positive, negative, and neutral. Next, we propose to develop a novel emotion perception classifier which can automatically classify an observer's emotion perception based on her gaze patterns and fixation sequence when identifying the basic emotions on expressive talking faces. The fixation time on the 16 predetermined AOIs were used as features to train support vector machine (SVM) models. The proposed models achieved the overall classification accuracy of 84.1% on recognizing 3-way emotions of *negative, positive and neutral*, suggesting that the proposed eye gaze patterns - the fixation time on 16 predetermined AOIs, are very promising for automatic classification of the perceived emotions. Finally, we divided the data into different gender and race groups, and discussed the diversity in gaze patterns across genders and different race groups.**

***Index Terms*—eye fixation, emotional face, gaze pattern, emotion recognition, gaze diversity**

I. Introduction

Emotions are essential to human life. They directly influence human perception and behaviors, and have big impacts on our daily tasks, such as learning, social interaction, and rational decision-making. Automatic emotion recognition has found applications in many domains, including multimedia retrieval, social media analysis, human-computer and human-robot interaction, health care, etc [1], [2].

Ekman's basic discrete emotion theory provides the perceptual basis for discriminating between distinct types of emotional expressions [3]. Based on that, facial expressions of six prototypical universal emotions (*Anger, Disgust, Fear, Happy, Surprise, and Sad*), can be produced with characteristic configurations of facial muscle movements. Action Units (AU) are defined as a contraction or relaxation of one or more muscles related to facial expression. Human coders can manually code almost any anatomically possible facial expression, and deconstruct it into the specific AUs and their temporal segments that produced the expression. For example, the facial expression of happiness consists of a flexing of mouth muscles, which include AU6 (Cheek raiser) and AU12 (Lip corner puller). The facial expression of anger consists of AU4 (Lowering of brow), AU5 (Upper lid raider), AU7 (Lid tightening) and AU23 (Lip tightening).

Research on emotion recognition from cues expressed in facial expression has a long-standing tradition. Numerous prior studies followed Ekman's basic discrete emotion theory and concentrated on emotion perception from facial cues [4], [5], [6]. They have established that prototypical basic emotions can be universally recognized by different groups of people based on the activation of specific facial expressions [7], [8], [9], [10]. Based on that, it's natural to expect that when distinguishing whether a face displays a certain emotion, some regions of the face may contain more useful information than the others. To better understand the attentional deployment in the processing of facial expression, many researchers investigated the attentional biases to emotional faces using the visual search paradigm [11], the dot-probe task [12], and by monitoring the eye-tracking face processing [13]. For example, in a study by Alshehri et al. [14] studied the attention on emotional facial expressions by examining eye movement features such as pupil size, timing of eye fixation and duration in emotional stimulation, while another study [15] analyzed EEG responses and eye gaze data in response to emotional videos.

It is well-established that when participants are presented both emotional and neutral faces, they look longer and more frequently at the emotional faces. However, we still do not have a full understanding of the impact of different emotional faces on visual attention and gaze behavior. On the other hand, many existing studies focused on analyzing the fixation and gaze pattern on still face images. Schurgin et al. [16] explored how the distribution of attention differ when examining facial

expression that correspond to emotion. The participants were asked to distinguish emotional and neutral facial expressions in the form of still images as their ocular behavior was being monitored. The results from the eye tracking analysis showed that the distribution of fixation time across regions have yielded 5 main facial regions of interest (upper nose, eyes, lower nose, nasion, and upper lip) that was accounted for 88.03% of all fixations together. However, there is little research studying the eye fixation and the movement pattern when viewing the video stimuli of facial expression during talking. It is expected that talking would introduce facial movement unrelated to emotion expression, and it may affect the attentional biases to emotional faces and disturb the recognition of facial expressions [17], [18]. Recent studies that have presented auditory and visual stimuli at the same time report that the cross-modal influences involving facial and vocal expressions are bi-directional [19], [20], [21], [22].

In this study, our goal is to investigate human's visual attention and fixation patterns when identifying six basic emotions of expressive talking faces. Stimuli for the current experiments consisted of 92 video clips of facial expression during talking. The whole experiments were divided into two sessions. The video stimuli in the first session were presented in random order across different face identities, while in the second session the video from the same face identity were played in sequential order. The participants' eye movements were recorded by the Tobii X3-120 screen-based eye-tracking system. We first defined a set of area-of-interest (AOI) regions, including 4 AOIs of general face areas and 12 AOIs related to the specific Action Units (AUs) involved in the coding of the six basic emotions. After that, the gaze pattern analysis was done by looking at their fixation time on this predetermined set of AOIs. We used the ANOVA analysis to uncover the gaze pattern, e.g., significant AOIs, on identifying positive, negative and neutral talking faces. Next, we used the fixation time on different AOIs as features to develop an advanced machine-learning based emotion classifier to identify the observer's perceived emotion. Finally, we turned to discuss the diversity in facial expressions and the gaze patterns during emotion recognition on talking faces. We investigated how facial expressions differ across different genders and races, and identified group-specific gaze patterns for different gender and race groups during emotion recognition. Furthermore, we used the gender-dependent and race-dependent gaze data to build the gender-dependent and race-dependent models, and compared their performances with the universal classifiers trained with the gaze data from all populations.

II. Eye-tracking of Emotion Recognition on Facial Expression during Talking

In this set of experiments, we recorded eye movements of participants when they were discriminating six basic emotions (*Anger, Disgust, Happy, Neutral, Sad)* on talking faces. Our goal is to investigate human's visual attention when identifying different emotions and examine the fixation patterns on emotional faces.

A. Ethics Statement

This research was reviewed and ethically approved by the Institutional Review Board at the New York Institute of Technology. Informed written consent was obtained from each participant prior to entering the study.

B. Participants

The participants of this pilot study are student and faculty volunteers from the authors' institute. There were a total of 18 participants (7 female; 11 male). Participants ranged in age from 19 to 29 years old (M = 22.33, SD = 2.60), with diverse ethnic groups in this pilot study.

C. Stimuli

Stimuli for the our experiments consisted of 92 facial expression video clips carefully selected from the *Crowd-sourced Emotional Multimodal Actors Dataset* (CREMA-D) [23]. Each clip contains a 3–5 seconds long facial expression videos of actors saying one of the two selected utterances– "Don't forget a jacket/I think I have a doctor's appointment"– in one of the six basic emotions: *Anger, Disgust, Happy, Neutral, Sad*. The actors in these clips consisted of an equal number of Caucasian, African American, Asian, and Hispanic, distributed evenly across genders (9 male, 7 female) to maximize the generalizability of our findings across these variables, which can strongly affect face recognition performance.

D. Apparatus

Stimuli were presented on a 17-inch LED monitor (resolution of 1366x768, 60Hz refresh rate) with a Dell desktop. The monitor located approximately 25-inch from the participants' eyes. We used the Tobii X3-120 screen-based eye tracking system to collect the fine-grained gaze data. It has a stable data-rate of 120Hz, that 120 gaze data points per second are collected for each eye. The drift of the system is about 0.1 degrees, spatial resolution is 0.2 degrees, and the gaze accuracy is around 0.5 degrees. The eye-tracker was mounted at the bottom of the LED monitor and connected to the Dell desktop for data collection. Responses were collected through the keyboard. The data was collected by the Tobii Pro Lab software. More specifically, we first used the Tobii Pro Lab to create a timeline of images and videos that are served as the stimulus for the experiment. After a recording has been made, we marked the areas of interest (AOI) on the timeline and recordings, and then exported the fixation time and the order of the AOIs for further analysis.

E. Procedure

The entire experiment consisted of two timelines of a total of 184 trials. We presented each stimulus twice, one in each timeline. The first timeline (random session) which presented stimulus in a completely random order of face identity and the type of expression, while the second timeline (sequential session) grouped stimulus based on their face identity and

TABLE I
ACTION UNITS AND THE CORRESPONDING FACIAL MOVEMENTS RELATED TO BASIC EMOTIONS

AU1	Inner Brow Raiser	**AU16**	Lower Lip Depressor
AU2	Outer Brow Raiser	**AU20**	Lip stretcher
AU4	Brow Lowerer	**AU23**	Lip Tightener
AU6	Cheek Raiser	**AU26**	Jaw Drop
AU7	Lid Tightener	**Left**	Left face
AU9	Nose Wrinkler	**Lower**	Lower face
AU12	Lip Corner Puller	**Right**	Right face
AU15	Lip Corner Depressor	**Upper**	Upper face
Happiness	AU6+ AU12		
Sadness	AU1+ AU4+ AU15		
Fear	AU1+ AU2+ AU4+ AU5+ AU7+ AU20+ AU26		
Ange	AU4+ AU5+ AU7+ AU23		
Disgust	AU9+ AU15+ AU16		
Neutral	AU12 + AU14		

TABLE II
PARTICIPANTS' RESPONSE ACCURACY (%) ON RANDOM AND SEQUENTIAL SESSIONS.

	Random Session	**Sequential Session**
Happiness	87.6	90.7
Anger	53.9	66.0
Neutral	74.6	77.1
Sad	49.0	54.4
Disgust	71.7	78.5
Fear	49.7	53.7
Overall	58.1	70.1

presented all stimulus from the same face identity together. Each participant could take a break (at least 30 minutes) between each session. Each session began with a 9-point eye calibration & validation procedure following an instruction screen. Each trial started with the video presentation of the facial expression along with a spoken phrase. After watching the video, participants were asked to press the number on the keyboard which corresponds to the answer they think is correct for each facial expression. The six possible answers were displayed on the screen with a corresponding number for each emotion (*0 = neutral, 1 = angry, 2 = disgust, 3 = fear, 4 = happy, 5 = sad*). Participants were asked to respond as fast and as accurately as possible. There was no time limit on each trial and the next trial would only begin after a response had been made. The entire experiment will automatically finish when all trials are completed.

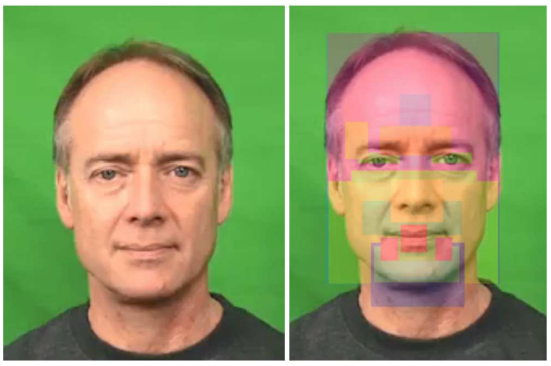

Fig. 1. Example of a stimulus before (left) and after marking the AOIs (right).

III. EYE TRACKING ANALYSIS

For each video stimulus, we defined 16 Area-of-Interest (AOI) regions as shown in Table I, and measured the percentage of the total fixation time over these ROI regions. The predetermined AOIs include 12 specific Action Units (AUs) involved in the coding of the six basic emotions, as well as the upper, lower, left and right side of the face which might show more general trends. Fig. 1 shows an example of a stimulus before and after marking the AOIs.

A. Response Accuracy

We first examined the response accuracy. An accurate response was defined as the participant selecting the correct emotional facial expression for each trial, i.e., the participant's perceived emotion matches the intended emotion from emotion expression. As shown in Table II, participants' overall accuracy is 58.1% for the random session, and 70.1% for the sequential session where the participants rated all stimulus from one actor together. The gap between the response accuracy for random versus sequential sessions is due to that the participant became more familiar with the face identity in the sequential session, thus was able to perceive the emotion more accurately. In addition, we also observed the confusion matrices to better understand which classes are most likely to be confused with each other during the emotion perception with talking face. As shown in Figure 2, *happy* is the most clear emotion, while *sad* and *fear* are the two most ambiguous emotions. Significant confusion can be observed between negative emotions and neutral, and among different negative emotion classes.

Next, we grouped all the four negative emotion classes together and examined the response accuracy in terms of detecting *Positive, Negative* and *Neutral*. As expected, the gap between the response accuracy for random versus sequential sessions reduce significantly, and the corresponding participant response accuracy was 79% for random session and 83% for sequential session.

B. Gaze Pattern Across Emotions

We used one-way ANOVA F-test to see if there are any significant differences in the mean fixation time on any subset of AOIs for facial expressions of different emotion groups. If significant difference was observed on any AOIs, we further followed-up with the post hoc test by using the Tukey's method. Based on the significant test results on this pilot sst of 18 participants, we didn't observe significant differences in the mean fixation time on any AOIs for the six basic emotions. Next, we clustered the six basic emotions into three groups: *Positive, Negative* and *Neutral*, and applied ANOVA analysis to see if there is any significant differences in the mean fixation time on any subset of AOIs for facial expressions

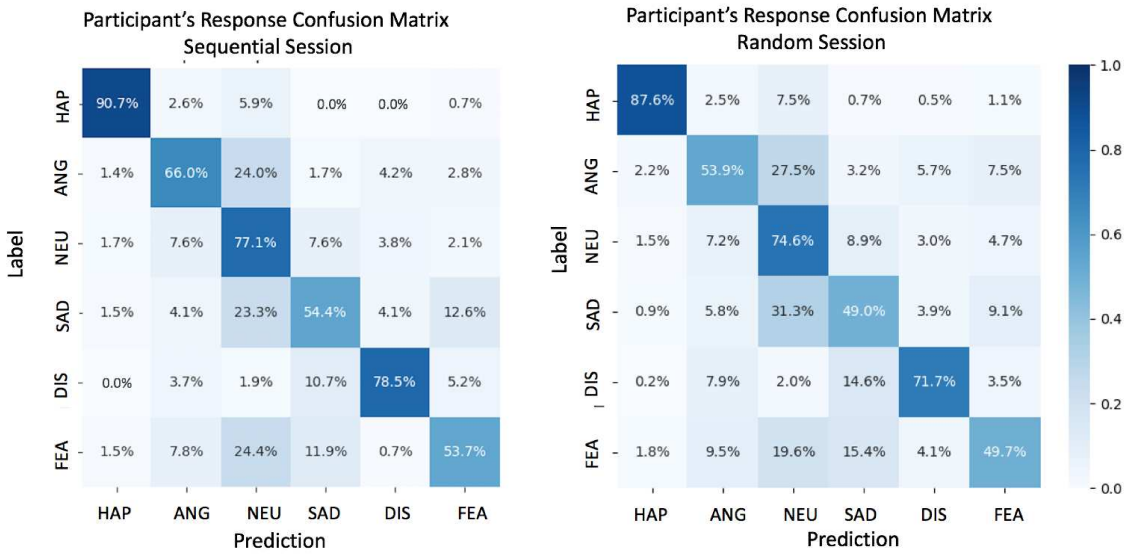

Fig. 2. Confusion matrix for participants' response accuracy (%) in random and sequential sessions.

of the three emotion groups. We showed that there indeed were significant differences in the fixation time of a subset of AOIs for the three emotion groups. For the random session, there was a significant difference of the mean fixation time for AOIs related to AU1 (Inner Brow Raiser), AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU9 (Nose Wrinkler), AU12 (Lip Corner Puller), Left face, Lower face, Right face, and Upper face with a significance level of 0.05 or less. For the sequential session, there was a significant difference of the mean fixation time for AOIs related to AU7 (Lid Tightener), AU9 (Nose Wrinkler), Left face, Right face, and Upper face with a significance level of 0.05 or less.

IV. Automatic Classification of Perceived Emotions based on Gaze Patterns

While the traditional emotion recognition task aims at analyzing the emotional status of a subject based on her facial expression, etc., it is also important to automatically classify the emotion perception of an observer. In a HCI/HRI conversation scenario, assuming the computer/robot can mimic human's emotional facial expression, it is as important for the computer/robot to study the human's perception of its mimicked emotion based on how the human's gaze pattern changes during the conversation. *For this purpose, we propose to develop a novel* **emotion perception classifier** *which can automatically classify an observer's emotion perception based on her gaze patterns and fixation sequence when identifying the basic emotions on expressive talking faces.*

Here, we focused on the 3-way classification and sectioned the six basic emotions into three groups of *Positive, Negative*, and *Neutral*. There is an imbalance of data with the *Negative* class having four times the amount of data points of the *Neutral* and *Positive*. In order to address the data imbalance problem, we down-sampled the *Negative* class data points randomly to create even sized classes.

TABLE III
Classification Accuracy (%) of Eye Fixation Models with Random Data, and Sequential Data

Data	Overall Accuracy	Positive	Negative	Neutral
Random Data	84.1	89	74	88
Sequential Data	81.7	90	69	87

We trained the Support Vector Machine (SVM) with the eye-gaze features - the fixation time on the 16 predetermined AOIs. In order to investigate the effect of the face familiarization, we divided the data into the following two main sets, and trained two different models accordingly.

1) **Random Data:** all data points from the random experiment session
2) **Sequential Data:** all data points from the sequential experiment session

We first examine the classification performance of the selected eye gaze features - the fixation time on the 16 predetermined AOIs. Table III lists the 5-fold cross-validation rates of models trained with random data and sequential data respectively. Detailed confusion matrices are provided in Figure 3 as well.

It can be seen that both models can achieve very encouraging performance. The overall classification accuracy is 84.1% with random data and 81.7% with sequential data. This suggest that the proposed eye gaze patterns - the fixation time on the 16 predetermined AOIs, are very promising features for automatic classification of the perceived emotions. It can be also observed that the proposed models obtain the highest classification accuracy on *Positive*, followed by *Neutral*. As expected, they have the lowest classification accuracy on

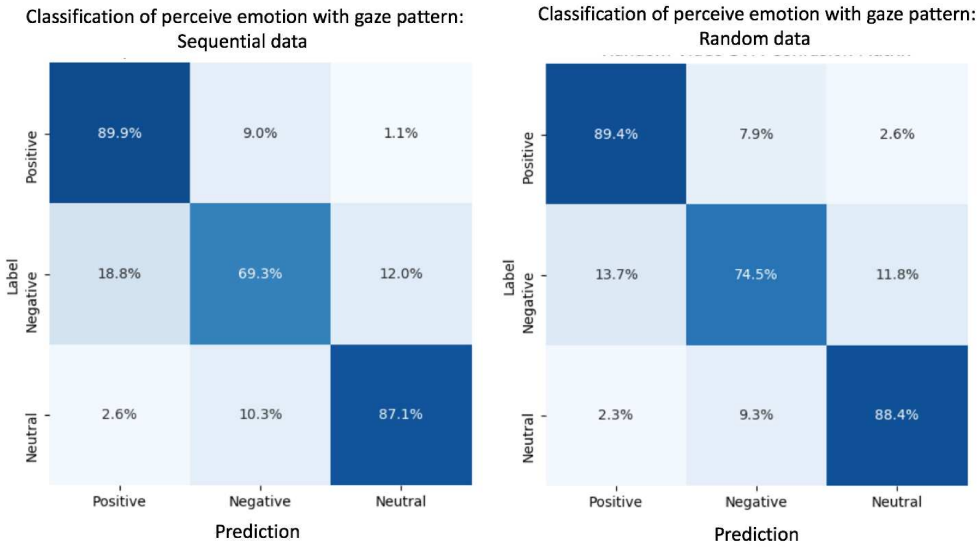

Fig. 3. Confusion matrix for eye fixation based perceived emotion classification models with random data, and sequential data respectively.

Negative, which is the group of four negative emotions of *anger, disgust, fear and sad.*

Next let's turn to discuss the effect of face familiarness. This can be done by comparing the classification rates of the random models vs the sequential models. It can be seen that the models trained with the random data slightly outperformed the model trained with the sequential data. This might because under random stimuli, participants' gaze patterns are more independent while under sequential stimuli a participant's gaze pattern for one trial is influenced by the patterns from the previous trials for the same face identity. Consequently, the random stimulus session seems to have more valuable eye fixation ROIs and the model trained from random stimuli data is slightly more accurate.

V. Diversity in Gaze Pattern During Emotion Recognition on Talking Faces

A critical aspect limiting emotion recognition performance in practice is the intrinsic personal diversity. Aspects of our heritage including race, ethnicity, culture, and our individual identity such as age, gender and visible forms of self-expression, are reflected in our emotion expression and perception. In this study, we rely on data-driven methods to create *emotion perception classifier* based on the gaze patterns. We are interesting in the following fundamental questions:

1) Are facial expressions different across different gender and race groups?
2) Can we identify group-specific gaze patterns for different genders and races?
3) Should we train different *emotion perception classifiers* for different genders and races?

A. Gaze Patterns across Gender & Race Groups

In order to answer those questions, we first investigated the facial expressions and gaze patterns during emotion perception across different gender and race groups, and tried to identify if there are any significant differences across different groups. Specifically, we performed a series of significance tests to compare the fixation time on AOIs across different Participant (P)'s gender & race and Actor (A)'s gender & race groups, on random data and sequential data respectively.

The results are summarized in Table IV. Here,"Y" means there is a significant difference of the mean fixation time for that specific AOI with a significance level of 0.05 or less , and "N" means no significant difference was detected for that AOIs. Our results showed that there are significant differences in the fixation time of a subset of AOIs for different gender and race groups. For example, we can consistently find significant difference of mean fixation time for AOIs related to AU7 (Lid Tightener), and AU12 (Lip Corner Puller), on both random and sequential data. On the other hand, no significant differences were observed on AU2 (Outer Brow Raiser), AU15 (Lip Corner Depressor), AU16 (Lower Lip Depressor), AU20 (Lip stretcher) , and AU23 (Lip Tightener), across any gender and race groups.

B. Gender-dependent Emotion Perception Classifiers

Next, in order to compare the performance of the general *emotion perception classifier* with gender- and race-dependent classifiers, we segmented the data into sections based on the gender and race of the participant, and the gender and race of the actor in the video, and further trained various gender-dependent and race-dependent *emotion perception classifiers* with each of these data segments.

We first discuss the gender-dependent classifiers. We separated the data out based on the participant and the actor gender, separately. The results in Table V show the emotion classification accuracy of eye fixation models based on participant (P) and actor (A) gender, with random and sequential

TABLE IV
SIGNIFICANT ANALYSIS TO COMPARE FIXATION TIME ON AOIS ACROSS DIFFERENT PARTICIPANT (P)'S GENDER & RACE, ACTOR (A)'S GENDER & RACE GROUPS, ON RANDOM AND SEQUENTIAL DATA. Y - THERE IS A SIGNIFICANT DIFFERENCE OF MEAN FIXATION TIME FOR THAT SPECIFIC AOI WITH A SIGNIFICANCE LEVEL OF 0.05 OR LESS; N - NO SIGNIFICANT DIFFERENCE IN THE MEAN FIXATION TIME WAS DETECTED.

	Random Sessions				Sequential Sessions			
Action Units	Gender (P)	Race (P)	Gender (A)	Race (A)	Gender (P)	Race (P)	Gender (A)	Race (A)
AU1 (Inner Brow Raiser)	N	N	N	Y	N	N	N	N
AU2 (Outer Brow Raiser)	N	N	N	N	N	N	N	N
AU4 (Brow Lowerer)	Y	N	Y	N	N	N	N	N
AU6 (Cheek Raiser)	Y	Y	Y	Y	Y	Y	N	N
AU7 (Lid Tightener)	Y	Y	Y	Y	Y	Y	Y	Y
AU9 (Nose Wrinkler)	Y	Y	Y	Y	N	N	N	Y
AU12 (Lip Corner Puller)	Y	Y	Y	Y	Y	Y	Y	Y
AU15 (Lip Corner Depressor)	N	N	N	N	N	N	N	N
AU16 (Lower Lip Depressor)	N	N	N	N	N	N	N	N
AU20 (Lip stretcher)	N	N	N	N	N	N	N	N
AU23 (Lip Tightener)	N	N	N	N	N	N	N	N
AU26 (Jaw Drop)	N	Y	N	N	N	N	N	N
Upper Face	Y	Y	Y	Y	Y	Y	N	Y
Lower Face	Y	Y	Y	Y	N	N	N	N
Left Face	Y	Y	Y	N	Y	N	Y	N
Right Face	N	Y	Y	Y	Y	Y	Y	Y

data, respectively. We also give the respective results for the general gender-independent classifier in the table, for the sake of comparison.

The results show that the performance of the emotion perception classifiers trained with male participants were higher then the classifiers trained with female participants, for both random and sequential sessions. This differs from what we observed on the actor genders, where the classification performances on female actors surpassed the male actors. Those results suggest that the female actors have more clear emotion expressions than the male actors, while male participants have more consistent gaze patterns during the emotional perception. We also notice that although the gender-dependent models with different actor (A) gender outperform the universal gender-independent models, the results from gender-dependent models with different participant (P) genders are even worse than the gender-independent models. This suggest that the gender-independent models can capture the intrinsic inter-subject variability from their gaze patterns.

C. Race-dependent Emotion Perception Classifiers

Next we turn to discuss the race-dependent classifiers. Similar as what we have done on gender-dependent analysis, we separated the data out based on the participant and actor's race. The results in Table VI show the emotion classification accuracy of eye fixation models based on participant (P) and actor (A) race, with random and sequential data, respectively. We also give the respective results for the general race-independent classifier in the table, for the sake of comparison. We notice that the classification accuracy are significantly different for different race groups, and the race-dependent models outperform the race-independent models in the most cases. This suggests that there may exist considerable differences on emotion expression and gaze patterns during emotion perception between different race groups. As this is only our initial study with a small number of participants from different race groups, we will collect more data from a larger population to confirm our finding in the future.

TABLE V
CLASSIFICATION ACCURACY OF EYE FIXATION MODELS BASED ON PARTICIPANT (P) AND ACTOR (A) GENDER, WITH RANDOM AND SEQUENTIAL DATA, RESPECTIVELY.

Random Session	Overall	Positive	Negative	Neutral
Male (P) Model	85%	89%	79%	88%
Female (P) Model	80.1%	85%	74%	81%
Male (A) Model	83.7%	85%	74%	81%
Female (A) Model	86.3%	88%	77%	86%
Gender-Independent Model	84.1%	89%	74%	88%

Sequential Session	Overall	Positive	Negative	Neutral
Male (P) Model	84.2%	88%	78%	87%
Female (P) Model	79%	85%	74%	81%
Male (A) Model	82.3%	85%	75%	86%
Female (A) Model	83.6%	88%	77%	86%
Gender-Independent Model	81.7%	90%	69%	87%

VI. CONCLUSION

In this study, we investigated human's visual attention and fixation patterns when identifying six basic emotions on expressive talking faces. We defined a set of area-of-interest (AOI) and further investigated the fixation time on this predetermined set of AOIs. Based on the ANOVA analysis, we did not find significant difference in the mean fixation time on any AOI for discriminating the six basic emotions, but a subset of significant AOIs can be found when we sectioned the six basic emotions to positive, negative, and neutral. Next, the fixation time on different AOIs were used as features to train support vector machine models to recognize 3-way emotions of negative, positive and neutral. We trained

TABLE VI
CLASSIFICATION ACCURACY OF EYE FIXATION MODELS BASED ON PARTICIPANT (P) AND ACTOR (A) RACE, WITH RANDOM AND SEQUENTIAL DATA, RESPECTIVELY.

Random Session	Overall	Positive	Negative	Neutral
Asian (A) Model	79.3%	79%	72%	88%
White (A) Model	86.8%	96%	82%	83%
Hispanic (A) Model	86.7%	88%	91%	90%
Black (A) Model	90.6%	90%	86%	96%
Asian (P) Model	82%	87%	76%	83%
White (P) Model	84.5%	88%	80%	86%
Hispanic (P) Model	85.7%	91%	81%	86%
Black (P) Model	86.2%	93%	83%	83%
Race-Independent Model	84.1%	89%	74%	88%

Sequential Session	Overall	Positive	Negative	Neutral
Asian (A) Model	84.8%	91%	79%	85%
White (A) Model	84.1%	85%	79%	89%
Hispanic (A) Model	84.5%	88%	78%	87%
Black (A) Model	79.1%	82%	72%	84%
Asian (P) Model	82.3%	89%	76%	82%
White (P) Model	82.6%	86%	77%	86%
Hispanic (P) Model	78.4%	82%	74%	80%
Black (P) Model	85.1%	89%	81%	86%
Race-Independent Model	81.7%	90%	69%	87%

different models based on sequential data and random data respectively, and the best performance of 84.1% is achieved on the random data. This suggests that humans tend to have consistent fixation pattern on the same face identity even across different expressions. Finally, we investigated the diversity in gaze pattern during emotion recognition on talking faces. We divided the data into different gender and race groups, and performed a series of significance tests to reveal the group-specific gaze patterns for different gender and race groups, and to identify the universal AOIs that can capture the intrinsic inter-subject variability. After that, we used the gender- and race-dependent gaze data to build the gender-dependent and race-dependent models, and we compare their performances with the universal perceive emotion classifiers trained with the gaze data from the whole population. We observed that female actors have more clear emotion expressions than male, while male participants (perceivers) have more consistent gaze patterns (e.g., AOIs) than female participants during the emotional perception. We also noticed that the independent models can capture the intrinsic variability on gaze patterns from different gender groups, while the race-dependent models outperform the independent models, which can better capture the group differences from different race groups. This is only our initial study and the number of participants is relatively small. In the future, we plan to recruit more participants and validate our findings on a larger population. We also plan to investigate the eye-movement trajectory to better uncover the dynamic gaze patterns during emotion recognition.

ACKNOWLEDGMENT

This work is partially supported by the US National Science Foundation (NSF) EAGER Grant IIS-2034791 and REU Grant CNS-1852316.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, pp. 32 – 80, 02 2001.

[2] C. Peter and R. Beale, Eds., *Affect and Emotion in Human-Computer Interaction, From Theory to Applications*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 4868.

[3] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of face recognition*. Springer, 2005, pp. 247–275.

[5] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[6] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 568–573.

[7] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, p. 974, 1999.

[8] J. F. Cohn and K. Schmidt, "The timing of facial motion in posed and spontaneous smiles," in *Active Media Technology*. World Scientific, 2003, pp. 57–69.

[9] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *Image and vision computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[10] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 149–149.

[11] A. Ohman, D. Lundqvist, and F. Esteves, "The face in the crowd revisited: A threat advantage with schematic stimuli," *Journal of personality and social psychology*, vol. 80, pp. 381–96, 04 2001.

[12] B. P. Bradley, K. Mogg, S. J. Falla, and L. R. Hamilton, "Attentional bias for threatening facial expressions in anxiety: Manipulation of stimulus duration," *Cognition and Emotion*, vol. 12, no. 6, pp. 737–753, 1998.

[13] M. Calvo and L. Nummenmaa, "Eye-movement assessment of the time course in facial expression recognition: Neurophysiological implications," *Cognitive, affective & behavioral neuroscience*, vol. 9, pp. 398–411, 12 2009.

[14] M. Alshehri and S. Alghowinem, "An exploratory study of detecting emotion states using eye-tracking technology," 10 2013.

[15] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, pp. 211–223, 04 2012.

[16] M. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Chiao, and S. Franconeri, "Eye movements during emotion recognition in faces," *Journal of vision*, vol. 14, 11 2014.

[17] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, and F. Lepore, "Audio-visual integration of emotion expression," *Brain research*, vol. 1242, pp. 126–35, 05 2008.

[18] S. Paulmann and M. D. Pell, "Is there an advantage for recognizing multi-modal emotional stimuli?" *Motivation and Emotion*, vol. 35, no. 2, pp. 192–201, 2011.

[19] T. Brosch, D. Grandjean, D. Sander, and K. Scherer, "Behold the voice of wrath: Cross-modal modulation of visual attention by anger prosody," *Cognition*, vol. 106, pp. 1497–503, 04 2008.

[20] G. Pourtois, B. Gelder, J. Vroomen, B. Rossion, and M. Crommelinck, "The time-course of intermodal binding between seeing and hearing affective information," *Neuroreport*, vol. 11, pp. 1329–33, 05 2000.

[21] T. Brosch, D. Grandjean, D. Sander, and K. Scherer, "Cross-modal emotional attention: Emotional voices modulate early stages of visual processing," *Journal of cognitive neuroscience*, vol. 21, pp. 1670–9, 10 2008.

[22] S. Paulmann and M. Pell, "Contextual influences of emotional speech prosody on face processing: How much is enough?" *Cognitive, affective & behavioral neuroscience*, vol. 10, pp. 230–42, 05 2010.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans Affect Comput. 2014 Oct-Dec; 5(4):377390*, 2014.