Proposing an Empirical Motion-Time Pattern of Human Gaze Behaviors in a Social Situation

Mohammad Hossein Mashaghi
Department of Mechanical Engineering
Sharif University of Technology
Tehran, Iran
mh.mashaghi@student.sharif.edu

Alireza Taheri*
Department of Mechanical Engineering
Sharif University of Technology
Tehran, Iran
***Corresponding author:**
artaheri@sharif.edu

Saeed Behzadipour
Department of Mechanical Engineering
Sharif University of Technology
Tehran, Iran
behzadipour@sharif.edu

Mehrdad Boroushaki
Department of Energy Engineering
Sharif University of Technology
Tehran, Iran
boroushaki@sharif.edu

***Abstract*— Social eye gaze is an important nonverbal behavior in human-human communications. Due to the rapid growth of the social robotics, social robots need to behave more and more humanlike. The present study strived to extract an empirical motion-time pattern of human gaze behavior while watching a certain video. After collecting the gaze data from an appropriate number of participants, we applied Gaussian mixture model (GMM) and Gaussian mixture regression (GMR) methods to elicit a pattern from their gaze behavior. A Bayesian Information Criterion (BIC) was used to determine the optimum number of components. Due to the indeterminate result of this criterion, we estimated a 60-component GMM to be suitable for this dataset. The resulting pattern from GMR method was visually acceptable. Although because of some limitations, we could not scientifically accept or reject the proposed model. Based on the survey, GMR results showed more similarity (but not significantly) to human gaze behavior rather than the mean pattern of people's actual gaze pattern.**

Keywords— Social robotics, Social eye gaze, Gaussian mixture model, Gaussian mixture regression

I. Introduction

Due to the progressive trajectory of robotic technologies, social robots will have remarkable effects on the future of the world. The main purpose of designing social robots is to have an intimate interaction with people. This interaction is applicable for educating people, certain treatments, and some fields of industry. Social robots have recently been used in the mentioned fields. Thus, these robots should be capable of performing such social behaviors in a way that their human users experience an easy and comfortable interaction [1]–[4]. Natural communication is required for this type of interaction. Although it seems that verbal communication is basic in human-human interactions, nonverbal behaviors are very important, too. Nonverbal communication includes eye gaze, gesture, etc. Using these nonverbal signals, people can express their mental mood and also improve their verbal communication. Eye gaze is even more significant than other nonverbal signals because it is proved in psychology that eyes are special cognitive stimuli with unique hardwired pathways in the brain dedicated to their interpretation [5].

In order to design a gaze control system for a social robot, it is required to find out details about the human gaze behavior at first. In this paper, we try to extract an empirical motion-time pattern from the gaze behavior of people while watching a certain video. In the first step, gaze data were collected from a number of participants, then we tried to fit a probabilistic model to the collected data. After finding the optimum model parameters, the model output was displayed on the testing video in order to visually validate the generated gaze pattern with a real human gaze pattern.

II. Background and Related Work

Studies in the field of social eye gaze can be divided into two general categories:

- Gaze pattern for diagnosis and classification
- Gaze control system for social robots

A. Gaze Pattern for Diagnosis and Classification

Studies conducted in this category tried to use the human gaze pattern to classify people or diagnose some mental illnesses. For instance, considering the three factors of human cognition, visual behavior and ongoing activity, Raptis et al. [6] were able to conclude that human cognitive characteristics are reflected in the eye movements. Using eye tracking data, they introduced two classification tests for cognitive characteristics. Hoppe et al. [7] recorded eye gaze data from 50 participants using SMI wearable glass. They firstly divided the participants into five personality groups; then used a random forest classifier to classify the gaze data into five clusters. They found that human personality features can be identified via eye gaze pattern. Rogers et al. [8] investigated the human eye gaze pattern in face-to-face conversation situation. They asked 76 participants to pair up and have a 4 minutes conversation while wearing a pair of Tobii Pro Glasses 2 to record their gaze data. Using mangold INTERACT software to analyze the gaze data, they found out that each individual spends a different amount of time looking at each part of the other person's face (e.g. eyes, mouth, etc.), but this time pattern is almost constant for each person.

On the other hand, some studies have been conducted to use eye gaze pattern to diagnose certain diseases or disorders such as Autism Spectrum Disorders (i.e. a developmental disorder with main symptoms of deficit in social interactions, communications, imaginative abilities [9], and imitation skills [10]). Liu et al. [11] asked 3 groups of 29 children to memorize 6 different human faces and then recognize them from 18 new faces. One group consisting of children with Autism Spectrum Disorders (ASD) and two other groups consisting of Typically Developing (TD) children. They recorded the children's eye gaze data using the Tobii T60 eye tracker and used the K-means and Support Vector Machine (SVM) methods to propose a quite accurate criterion for diagnosing autism disorder in the early ages. In a more extensive study, Jones et al. [12] attempted to provide a model for early detection of children with autism by taking data from 110 infants, each at 10 time points from 2 to 24 months. In these experiments, a video of a babysitter is shown to the baby and the data is collected by an ISCAN eye tracker. As a result of this study, a significant difference was observed at the age of 2 to 6 months between the patterns of gaze of healthy infants and those later diagnosed with ASD.

B. *Gaze Control System for Social Robots*

Most of the studies conducted in this category focused their effort on Human-Robot Interaction (HRI) field. For instance, Aliasghari et al. [2] used a mathematical model of human gaze behavior derived by Zaraki et al. [13], and tried to modify its coefficients by collecting gaze data from 23 participants watching a video containing 4 social cues. The data was collected by a Kinect sensor. They designed a gaze control system and implemented it on a social robot and evaluated the system's behavior by asking some other participants to interact with the robot. Lathuilière et al. [14] developed a method for controlling the robot's gaze using the reinforcement learning method. In this method, the robot's neural network learns to direct the robot's attention towards a direction that maximizes the number of people in its field of view, based on the input data consisting of the environment's audio and video. The designed network is first trained using pre-made videos and then placed in a real environment. In this study, human eye tracking data was not used and also the criteria for changing the direction of gaze are only human stimuli. Yoo et al. [15] introduced a method based on fuzzy integrals to control the direction and attention of a robotic head. In this method, 7 effective environmental factors were recorded by the robot's sensors and sent to the control unit. On the other hand, in another unit, user preferences were sent to the control unit as fuzzy criteria and the control unit selected the gaze direction. This system was tested in 6 scenarios (i.e. 5 scenarios with human presence and 1 scenario without human presence). In all 6 scenarios, the gaze control system demonstrated a natural and real-time performance.

III. Method

A. *Experiment Structure*

Due to the special conditions caused by COVID-19 pandemic, we encountered several difficulties. As a result, we were not able to conduct our own designed experiment. The dataset used in this study was provided by Djawad Mowafaghian Research Center of Intelligent Neuro-Rehabilitation Technologies (DMRCINT).

In this part, we explain the following details of the experiment conducted by DMRCINT:

- Participants
- Eye tracking setup
- Video used in the experiment
- Dataset

1) Participants

A total number of 9 adults were asked to participate in this experiment, including 2 healthy adults, 3 Parkinson's Disease (PD) patients with freezing of gait and 4 PD patients without freezing of gait.

TABLE I. Properties of the Video File

Property	Value
Resolution	1920x1080
Duration	30.560 sec
Frame rate	25 FPS

2) Eye Tracking Setup

DMRCINT is equipped with a SR-Research EyeLink 1000 Plus eye tracker. This video-based tracker is desktop-mounted, capable of sampling binocularly at up to 2000 Hz and has down to 0.15° accuracy.

During the experiment, the participant's head is fixed in its position in front of the LCD monitor (as shown in Fig. 1); he/she is asked to watch the video that is being displayed on the mentioned monitor and the eye tracker sensor captures the coordinates of the point on the screen that is being looked at by the participant (i.e. the participant's gaze data).

3) Video Used in the Experiment

The video used in this experiment is about 30 seconds and taken from inside of a subway station in Tehran, Iran. As shown in Fig. 2, the video is very crowded with people and there are lots of social stimuli in each frame.

4) Dataset

The tracking device (EyeLink 1000 plus) is capable of capturing a vast variety of parameters during the experiment including position, velocity, acceleration, pupil diameter, fixation/saccade/blink status, etc. In this study we only used the position of each participant's gaze data. The sampling rate on this experiment is 1000 Hz. The recorded data can be viewed in different diagrams and graphical plots using EyeLink Data Viewer software. It can also export the data into excel files.

B. *Pattern Extraction*

Our proposed method for extracting a pattern from the gaze data is Gaussian Mixture Model (GMM) and Gaussian Mixture Regression (GMR).

1) GMM

Gaussian Mixture Models are probabilistic models for displaying normally distributed subsets within a general set. Generally speaking, mixture models do not need to know which subset each datapoint belongs to, which allows the model to automatically learn the subsets. It is not clear in advance which subset each datapoint belongs to, so this is a kind of unsupervised learning [16]. One of the most popular

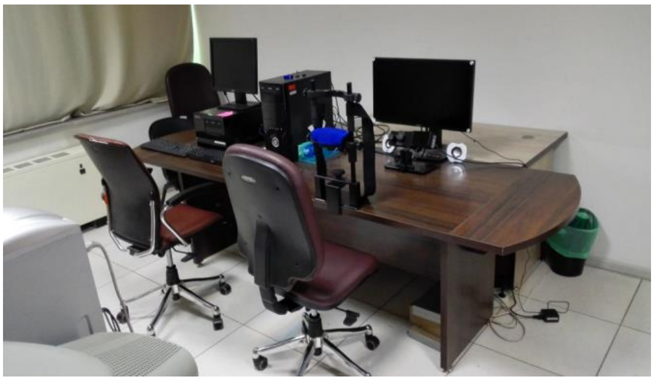

Fig. 1. Eye tracking setup at DMRCINT.

Fig. 2. one frame of the video used in the experiment

ways to approximate the density of continuous or binary data is the mixture modeling. It provides flexibility by considering an appropriate tradeoff between the complexity of the model and the variations of available training data. The definition of a mixture model consisting of K components is

$$p(\vec{x}) = \sum_{k=1}^{K} p(k)p(\vec{x}\,|k). \tag{1}$$

where $\vec{x}$ is a datapoint, $p(k)$ is the prior, and $p(\vec{x}\,|k)$ is the conditional probability density function [17]. Note that the datapoint vector in this study is $\vec{x} = \{t, x, y\}$, the time-based position of the gaze.

A Gaussian mixture model consists of two types of parameter values, the mixture component weights and the component means and covariances. For a Gaussian mixture model with K components, the k^th component has a mean of $\overrightarrow{\mu_k}$ and covariance matrix of Σ_k and mixture component weight of ϕ_k. In order to normalize the probability distribution, the following constraint rules on the mixture component weights:

$$\sum_{i=1}^{K} \phi_i = 1. \tag{2}$$

If the component weights are not learned, they can be viewed as an a-priori distribution over components so that $p(x \text{ generated by component } C_k) = \phi_k$. If they are instead learned, they are the a-posteriori estimates of the component probabilities given the data [16].

$$p(\vec{x}) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x}|\vec{\mu}_i, \Sigma_i) \tag{3}$$

$$\mathcal{N}(\vec{x}|\vec{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right) \tag{4}$$

Finally, a multi-dimensional GMM can be represented as (2), (3) and (4) together.

2) *Learning the Model*

The maximum likelihood estimation method for the mixture model parameters is performed iteratively using the standard Expectation-Maximization (EM) algorithm. EM is a simple local search technique looking for maximum likelihood. This method ensures a uniform increase in the likelihood of training set during optimization. An initial estimate is required for the EM algorithm. Firstly, a k-means clustering method is roughly applied to the data in order to provide the initial estimates and also to avoid getting caught in a weak local minimum point. Gaussian parameters are then obtained from the clusters found by k-means [17].

EM algorithm is explained in details in [18]. Here is a brief summary of the algorithm steps:

a) Step 1: Initialization

The model parameters (ϕ, μ, Σ) are initialized in this step. In order to avoid local minima, we can use the results obtained by a previous k-means run for instance, as a good starting point.

b) Step 2: Expectation

Calculate the term

$$\gamma(z_{nk}) = \frac{\phi_k \mathcal{N}(\vec{x}_n|\vec{\mu}_k, \Sigma_k)}{\sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x}_n|\vec{\mu}_i, \Sigma_i)}\,. \tag{5}$$

c) Step 3: Maximization

Calculate the term

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{6}$$

and then update the model parameters

$$\phi_k^* = \frac{N_k}{N}, \tag{7}$$

$$\vec{\mu}_k^* = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\,\vec{x}_n \tag{8}$$

and

$$\Sigma_k^* = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\,(\vec{x}_n - \vec{\mu}_k)(\vec{x}_n - \vec{\mu}_k)^T\,. \tag{9}$$

Steps 2 and 3 are iterated respectively until the model parameters are converged.

One drawback of EM is that the optimal number of *K* components in a model may not be already known. A common method for solving this problem is estimating multiple models by increasing the number of components and selecting the optimal one based on model selection criteria [17].

Therefore, there is a tradeoff between optimizing the model's likelihood (i.e. how suitable the model is for the data) and minimizing the number of parameters required to encode the data. Various criteria have been proposed: cross-validation, Akaike information criteria, Bayesian information criteria (BIC), and minimum description length are commonly found in the literature [17].

We decided to choose the BIC for this dataset. According to [17], the BIC score is calculated from

$$S_{BIC} = -\mathcal{L} + \frac{n_p}{2} log(N) \tag{10}$$

where

$$\mathcal{L} = \sum_{n=1}^{N} log\big(p(\vec{x}_n)\big) \tag{11}$$

and

$$np \ = \ (K \ - \ 1) + K\left(D \ + \left(\frac{1}{2}\right)D(D \ + \ 1)\right). \tag{12}$$

The optimum number of the components has the minimum BIC score.

3) GMR

The final step of pattern extraction is applying the Gaussian Mixture Regression on the mixture model. In fact, the GMR method reconstructs a general form for the gaze signals of all participants.

In this method, time values x_t are considered as query points (i.e. the points where you need to know the value) and the corresponding spatial values $\hat{x}_s$ are estimated by regression. For each GMM, the time and spatial values are separated, i.e., the mean and covariance matrix of the Gaussian component k are defined by

$$\vec{\mu}_k \ = \ \{\mu_{t,k}, \mu_{s,k}\}\,, \qquad \Sigma_k = \begin{pmatrix} \Sigma_{t,k} & \Sigma_{ts,k} \\ \Sigma_{st,k} & \Sigma_{s,k} \end{pmatrix} \tag{13}$$

For each Gaussian component k, the conditional expectation of $x_{s,k}$, given x_t, and the estimated conditional covariance of $x_{s,k}$, given x_t, are

$$\hat{x}_{s,k} = \mu_{s,k} + \Sigma_{st,k}\big(\Sigma_{t,k}\big)^{-1}\big(x_t - \mu_{t,k}\big) \tag{14}$$

and

$$\hat{\Sigma}_{s,k} = \Sigma_{s,k} + \Sigma_{st,k}\big(\Sigma_{t,k}\big)^{-1}\Sigma_{ts,k}\,. \tag{15}$$

$\hat{x}_{s,k}$ and $\hat{\Sigma}_{s,k}$ are mixed according to the probability that the Gaussian component k has, being responsible for x_t

$$\beta_k = \frac{p(x_t|k)}{\sum_{i=1}^{K} p(x_t|i)}. \tag{16}$$

Using (14), (15) and (16), for a mixture of K components, the condition expectation of x_s, given x_t, and the conditional covariance of x_s, given x_t, are

$$\hat{x}_s = \sum_{k=1}^{K} \beta_k \hat{x}_{s,k}\,, \qquad \hat{\Sigma}_s = \sum_{k=1}^{K} \beta_k^2 \hat{\Sigma}_{s,k}\,. \tag{17}$$

C. Model Evaluation

In order to check the validity of the output, we asked 20 people to participate in a survey. They were asked to watch (a) the original video, (b) the original video with the mean and covariance of all participants' gaze data corresponding to each frame plotted on the same frame, and (c) the original video with the GMR results of each frame plotted on the same frame, respectively. Finally, they were asked to give a Likert Scale score of 1 to 5 for each of the videos (b) and (c), based on how similar the pattern of each video was to their own gaze pattern from watching the video (a).

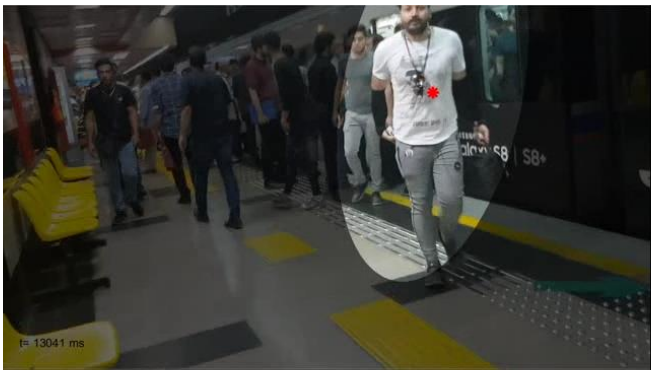

Fig. 3. A sample frame of the third video displayed at the survey.

IV. Results

We applied GMM and GMR methods to our dataset to extract an empirical motion-time pattern from human gaze behavior. At first, we calculated the BIC score for components up to 100. As shown in Fig. 4, the BIC score does not determine the optimum number of components, because the BIC score is almost uniformly decreasing by increasing K. Therefore, there is not a minimum BIC score.

Although we did not find the optimum number of components, we calculated the GMM and GMR for K=5, 22, and 60, respectively and plotted the results. As shown in Fig. 5, Fig. 6, and Fig. 7, black lines are the experimental data, red ellipses are the GMM components, thick green line is the reconstructed signal by GMR, and the transparent green tape represents the covariance at each data point.

After applying the paired t-test on the viewpoints of the participants of this study regarding the two used algorithms, although the mean value of the GMM/GMR method (with 60 components) was greater than the simple method (i.e. 3.45 (SD: 1.05) versus 3.15 (SD: 1.14)), no significant difference was observed. The calculated T-value was 1.03 and the p-value was 0.316 (>0.05). The overall viewpoint of the subjects regarding the gaze path generated via the GMM/GMR algorithm is considered to be between "3: moderate" and "4: agree" that is acceptable in this preliminary exploratory study. Including more participants in the study may also improve the results of the statistical test.

V. Limitations and Future Work

Due to the special conditions caused by COVID-19 pandemic, we were unable to conduct our desired experiment

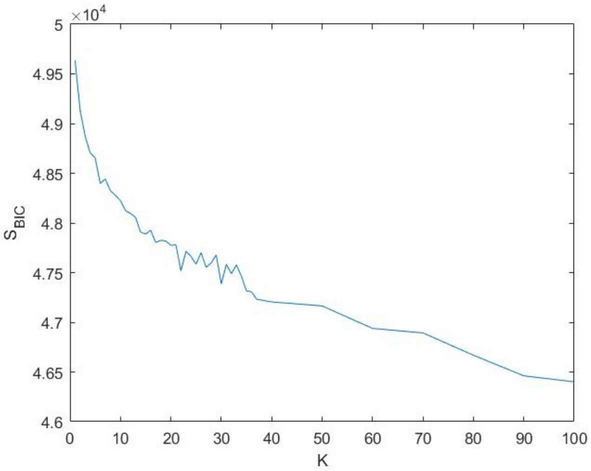

Fig. 4. BIC score of the models with *K*'s of up to 100.

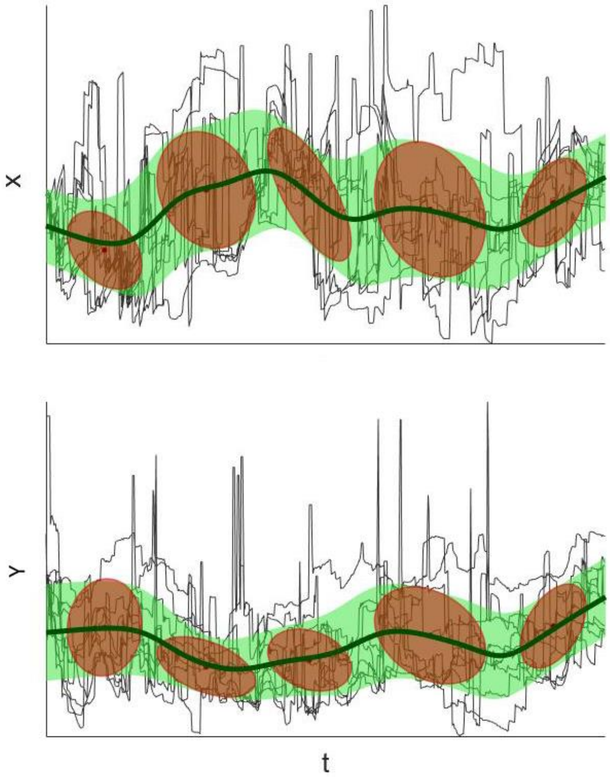

Fig. 5. GMR result for *K*=5

at DMRCINT. The video used in this study is not compatible with social cues that we would like to study systematically. The other problem is that the participants were not completely healthy and we had two groups of PD patients among the participants which makes it difficult to generalize the obtained results to the gaze patterns of the humans. The purpose of this study was not to investigate the effects of any physical or mental illness on the human gaze behavior. Unfortunately, the mentioned problems and limitations noticeably decrease the validity of the results.

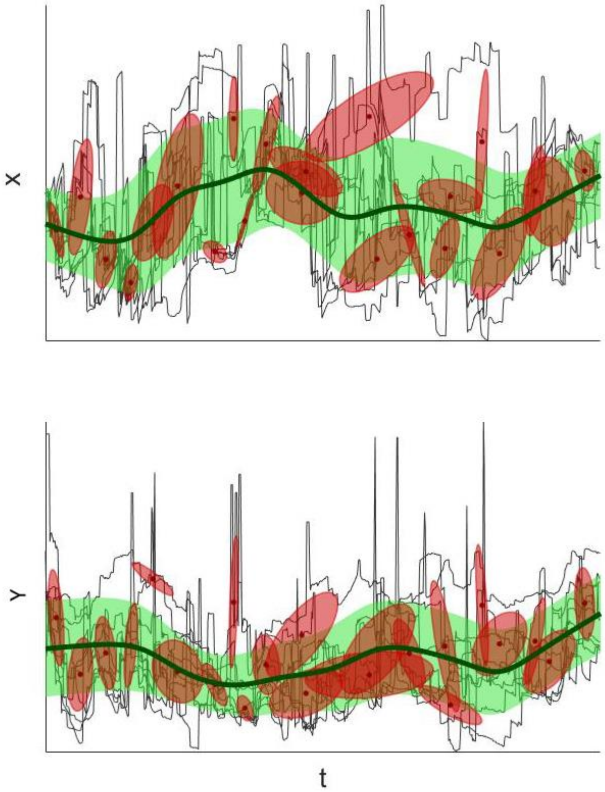

Fig. 6. GMR result for k=22

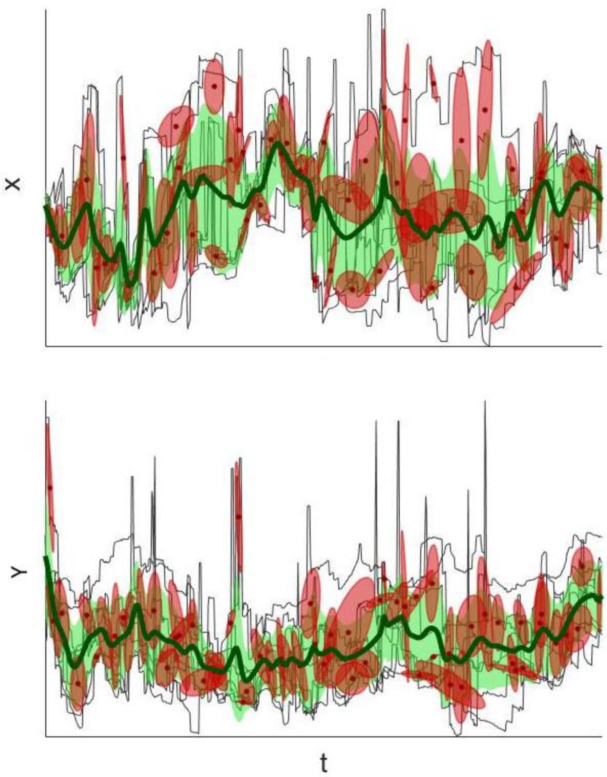

Fig. 7. GMR result for *K*=60

For the future studies, we propose to resolve all the mentioned problems and limitations as much as possible. We also suggest using more criteria to find the optimum number of the components, using other clustering methods such as Naive Bayes to compare with GMM, using Task Parameterized GMM (TP-GMM) method [19], etc. In TP-GMM method, the social cues inside the video and their position should be determined frame-by-frame. Then the trained model receives the present social cues as input and predicts the next cue that will be looked at.

VI. Conclusion

In this paper, we tried to use GMM and GMR methods to extract a motion-time pattern from human gaze behavior while watching a certain video (i.e. a 30 second video taken from inside a subway station in Tehran, Iran). Unfortunately, we couldn’t find the optimum number of Gaussian components based on the Bayesian Information Criterion. We tried 5, 22, and 60 components as instances to see the behavior of the GMR results. We plotted the GMR results on the original video and created a new video. 5 and 22 components resulted very smooth signals which are not similar to human eye’s motion at all. It also does not follow the social cues (the social cues in this video were mostly humans moving at different speeds in the station). The 60-component model has more acceptable results, because the reconstructed signal has more rapid movements (similar to human eyes) and has a better following of the social cues. Although we can visually confirm that 60 is the appropriate number for this model, it is much higher than the typical number of components used for similar studies. As seen in Fig. 7, the Gaussian components overlap at many points. This means the number of the components are very large. Except the visual observations of the results, we found no other way to accept or reject the GMM model with 60 components for this dataset.

VII. Acknowledgements

The authors would like to thank Ms. Fatemeh Sadat Daei Nejad, a PhD candidate at Sharif University of Technology, for generously providing our required dataset.

VIII. References

[1] A. Zibafar *et al.*, “State-of-the-Art Visual Merchandising Using a Fashionable Social Robot: RoMa,” *Int. J. Soc. Robot.*, 2019.

[2] P. Aliasghari, A. Taheri, A. Meghdari, and E. Maghsoodi, “Implementing a gaze control system on a social robot in multi-person interactions,” *SN Appl. Sci.*, vol. 2, no. 6, 2020.

[3] M. Alemi, A. Taheri, A. Shariati, and A. Meghdari, “Social Robotics, Education, and Religion in the Islamic World: An Iranian Perspective,” *Sci. Eng. Ethics*, vol. 26, no. 5, pp. 2709–2734, 2020.

[4] M. Tavakol Elahi *et al.*, “‘Xylotism’: A Tablet-Based Application to Teach Music to Children with Autism BT - Social Robotics,” 2017, pp. 728–738.

[5] H. Admoni and B. Scassellati, “Social Eye Gaze in Human-Robot Interaction: A Review,” *J. Human-Robot Interact.*, vol. 6, no. 1, p. 25, 2017.

[6] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris, “Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles,” pp. 164–173, 2017.

[7] S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling, “Eye Movements During Everyday Behavior Predict Personality Traits,” *Front. Hum. Neurosci.*, vol. 12, p. 105, 2018.

[8] S. L. Rogers, C. P. Speelman, O. Guidetti, and M. Longmuir, “Using dual eye tracking to uncover personal gaze patterns during social interaction,” *Sci. Rep.*, vol. 8, no. 1, Mar. 2018.

[9] M. Shahab *et al.*, “Social Virtual Reality Robot (V2R): A Novel Concept for Education and Rehabilitation of Children with Autism,” in *2017 5th RSI International Conference on Robotics and Mechatronics (ICRoM)*, 2017, pp. 82–87.

[10] A. Taheri, A. Meghdari, and M. H. Mahoor, “A Close Look at the Imitation Performance of Children with Autism and Typically Developing Children Using a Robotic System,” *Int. J. Soc. Robot.*, 2020.

[11] W. Liu, M. Li, and L. Yi, “Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework,” *Autism Res.*, vol. 9, no. 8, pp. 888–898, 2016.

[12] W. Jones and A. Klin, “Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism,” *Nature*, vol. 504, no. 7480, pp. 427–431, 2013.

[13] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, “Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot,” *IEEE Trans. Human-Machine Syst.*, vol. 44, no. 2, pp. 157–168, 2014.

[14] S. Lathuilière, B. Massé, P. Mesejo, and R. Horaud, “Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction,” *Pattern Recognit. Lett.*, vol. 118, pp. 61–71, Feb. 2019.

[15] B. Yoo and J. Kim, “Fuzzy Integral-Based Gaze Control of a Robotic Head for Human Robot Interaction,” *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1769–1783, 2015.

[16] “Gaussian Mixture Model | Brilliant Math & Science Wiki.” [Online]. Available: https://brilliant.org/wiki/gaussian-mixture-model/. [Accessed: 24-Aug-2020].

[17] S. Calinon, F. Guenter, and A. Billard, “IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, PART B 1 On Learning, Representing and Generalizing a Task in a Humanoid Robot,” vol. 37, no. 2, pp. 286–298, 2007.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[19] S. Calinon, “A Tutorial on Task-Parameterized Movement Learning and Retrieval,” *Intell. Serv. Robot.*, vol. 9, no. 1, pp. 1–29, Jan. 2016.