DEVELOPING A NEW AUTOMATED MODEL TO CLASSIFY COMBINED AND BASIC GESTURES FROM COMPLEX HEAD MOTION IN REAL TIME BY USING ALL-VS-ALL HMM

[1]Amina Atiya Dawood, [2]Dr Scott Turner, [3]Dr Prithvi Perepa
[1]PhD student,[2]Associate Professor computing and Principal Lecturer
Computing,[3]Programme leader MA SEN and Inclusion/ MA SENI (Autism)
[1]Department of Computing and Immersive Technologies,
University of Northampton,
Northampton, United Kingdom

Abstract*—Human head gestures convey a rich message, containing information deliver for peoples as a communication tool. Nodding, shacking are commonly used gestures as non-verbal signals to communicate their intent and emotions. However, the majority of head gestures classification systems focused on head nodding and shaking detection. while they ignored other head gestures which have more expressive emotional signals like rest(up and down), turn, tilt, and tilting. In this paper we developed a new model to classify all head gestures (rest, turn, tilt, node, shake, and tilting) from complex head motions. The model methodology based on distinguishing basic head movements (rest, turn, and tilt) and combined movements (nodding, shaking, and tilting). The purpose of this system is to detect and label combined and basic head movements in dynamic video. In addition, this phase of this study looking at developing an affective machine uses head movements to extract complex affective states (this work is underway). The system used 3D head rotation angles to classify relevant head gestures in-plan and out-plan of view during user interaction with computer. This system used an open source tracker to detect and track head movements. The Three angels that obtained from the tracker (pitch, yaw, and roll), wereanalyzed and packed into sequences of observation symbols or cues. Observations formed inputs to an all-vs-all discrete Hidden Markov Model (HMM) classifier. Three classifiers were used for each angle. The classifiers are trained on Boston University dataset, and tested on available mind reading data. The system evaluate on video streams in real time by webcam. The system is fully automatic without incurring any cost of technical methods and doesn't require any sensitive tools.

Keywords — Head Gestures, Head Pose, Nod, Shake, Tilt, Rest, Turn,All-vs-All HMM, HMM, Real Time.

I. INTRODUCTION

Human head gestures usually attempt to convey a message that carries out different meanings. The can be considered as a communication tool, particularly in conversations. For instance, head nodding gesture during a conversation is a sign of agreement or acceptance, that is to say, a pointer to listener attention, and understanding. Also, head shaking gesture refers to conceptual disagreement [1][2]. The ability to detect head motion by people is an effortless task [2]. However, it deemed difficult and challenging task in both computer vision context and human-computer interaction. Computational head poses estimation requires extensive processing to infer head rotation and translation. These transformations can be either taken with respect to a camera or world coordinates [16].

The existing literature on head gestures as a tool is extensive and focuses particularly on detecting head nod and shake gestures. Therefore, the majority of works investigated in this paper focused on the aforementioned points. While other head gestures get less attention, although it can observe specific emotions, for example head up movement related with pride, boredom and contempt, head down linked with greeting, guilt, embarrassment, and shame [3]. The overall aim of this paper is to develop a solid and robust system based on 3DOF, by using All-vs-All HMM classification to categorize combined head gestures (i.e., nod, shake, and tilting) and basic gestures (rest, turn, and tilt) in real time. This approach uses an open source head tracker and a webcam. The head tracking model is based on 3D head model aligned with 68 detected face feature points to detect head orientation (pitch, yaw, and roll). Other researchers used methods based on the coordinates of face feature points like (eye corner or nose tip) and use geometric calculations to estimate head orientation. The limitations of using features points are in incidents of occlusion due to changes in illumination. This system performs in an uncontrolled environment with variable illumination. Also, it does not assume that the head should start with neutral pose against other approaches. The rest paper structure as follow; in section I we introduce a briefreview of related works and important methods that used in head tracking and detection. Section III describes the methodology used for head tracking and extracting head actions/cues, which then used as observations in HMM. Section IV presents general review of HMM model and its elements with its problems. Model training methods can be found in section V and Classificationmethods in section VI. Results and system evaluation can be found in section VII and section VIII is for conclusions.

II. RELATED WORKS

Face Detection

Face detection is one of the first steps in any automated facial expressions recognition systems or head gesture detection. It is an automated process to extract and get the face region from a background in a still images or videos. This process is called face detection in process of finding a face from an image or a frame, otherwise, it called face tracking when it is tried to estimate the new position of detected face in a sequence of images or video. However, face detection is a difficult task due to the dynamics of face as a non-rigid object, and variations in location (in-plane), pose (out-of-plane), illumination conditions, and occlusions [17]. The techniques of face detection are based on the type of face models such as 3D or 2D. A number of methods have been developed to detect a face, for example in [18] these methods were classified into:

1. Knowledge-based methods, which use predefined knowledge rule to find face based on previous knowledge of human.
2. Feature invariant methods, these methods are more robust due to their aims to find the structure and shape of face features.
3. Template matching methods that try to recognize the face in the image based on pre-defined face model.

Appearance-based methods to describe the differences in appearance of facial features in images like (skin colour, face shape, and eye colour).

Generally, different methods have been developed in the area of face detection. One direction focused on In-plane of face detection or near frontal views of the face [42][19][31]. Furthermore, a method to detect a frontal face in a still grey image by using a component-based, trainable system [42]. Robust real-time face detector also developed by [20]. Basically, they used a cascade of binary linear classifiers. Beside to adequate face, detectors have been developed to detect faces from different poses. For example, Eigen-space methods to detect the face from different poses [34]. Deformable models attracted more attention in face detection and tracking methods [15]. For more information and details, a comprehensive survey about face detection and tracking can be found in [21][17].

Head Pose Estimation methods

Different techniques considered human head as a rigid object. Head pose estimation methods can track whole head region(2D or 3D) [28][53][30][29][45]. Examples of these methods include, 3D cylindrical head model to recover 6DOF (3-orientations and 3-translations) of head motion [38][22], and morphable model for synthesis faces of 3D dataset [4]. In [40] used 2D data to recognize head pose in real time based on random forests. Also, depth information of 2D face was used to estimate head pose in interaction between a robot and human by using stereo vision camera [29]. While 3D head model is more robust and accurate, but it consumes more time and manual working [13].

Other methods based on geometrical analysis of shape and location for a cloud of facial feature points to detect head orientation [44][32][36][23][46]. An exemplar of this method in [39], they used the location of both eyes corners (inner and outer) and nose tip to estimate head pose. In [24] another method used 5 facial landmarks (tip of the nose, left and right corner of mouth, and the out corner of eyes). Also, the same previous features were used to detect head in images [6]. Robust estimation of head can be obtained by geometric methods with limited features, but, the difficulty resides with miss-detected and out of context features, also when the detection requires high accuracy and precision [2]. Hybrid methods consist from more than one method to overcome their limitations used in [41].

Head gestures classifications methods

In this section, I will provide a brief review of the most used techniques to classify head gestures. These methods differ on the total of gestures classified, input symbols, and the training methods required by the model.

For example,HMM model used in [7] to detect head nod and shaking. With accuracy 81.08% of nod recognition, and 75.0% of shake. They used infrared camera equipped with infrared LEDs, to detect head direction from positions of eyes pupil. Then head direction was used as input observations to recognize head node, and shake. Two HMMs were used, each model has three states and five symbols. Similarly, discrete HMM used in [8] to detect head's nod and shake from video stream in real time. This system was based on extracted eye location that used as a pointer to head direction. The nodding accuracy is 82% and shaking accuracy is 89%.

In [37] HMM was used to detect four head gestures (yes, no, maybe, and hello). The model based on seven observation symbols (up, down, left, right, in, out, and rest). Each HMM trained against one of these gestures. Another method to classify nod and shake developed in [25]. This method was based on detecting eyes locations in videos, these locations then used to classify head node and shake by HMM. In [9] alternative method proposed by adopting a finite state machine to detect user's acknowledgements from head gestures captured by IBM Pupil Cam. Using aspects in [9], a dialog box tool was developed in [54]to recognize head nod and shake relative to yes and no respectively. Head nod machine used in this system had four states (start, up, down, and finish). The process starts with either nod up or down. In [52] automated model was developed to detect nod and shake gestures in American Sign Language video sequences. They used shape-based approach with 6DOF (orientation and translation), this work close to our work. However, the study in this paper based on different approaches like OpenFace tracker, 3DOF, and multiclass-HMM to detect all head gestures not only nod and shake.

In contrast to all works that mentioned above, we introduce a new system to detect all basic and combined head gestures from complex head movements in video and real time. Based on [10] they stated head convey information of emotions are not restrictedon nod and shake but include others additional gestures like (rest, turn, tilt, and tilting). All head gestures have the same important in emotion recognition.

The HMM classifier trained on BU dataset and tested on complex movements of available mind reading data. In previous works, the x and y coordinates are more control as pointers to head direction, these coordinates extracted either based on eyes coordinates or based on special features like nose tip. In this study the model training based on observation of 3D coordinates x, y, and z with 3 symbols. The HMM classifier trained on BU dataset, each video in BU dataset have one movement, for example basic movement or combined like nod, or shake, or tilting. Testing was done on complex movement of available mind reading dataset.

III. SYSTEM PERFORMANCE

Head Pose Estimation

Head plays an important role in human communication, it reveals emotions through its movements such as nod, tilt, and shake [26]. However, limited works focused on head role to extract human emotion [13][14].

In computer vision and HCI environments, head pose estimation can be defined as a set of operations to infer and interpret human head direction, specifically pitch, yaw, and roll (figure.1) from still or dynamic images in 2D or 3D spaces [27][41]. Head pose representation requires sequence of operations to transform 2-D face feature points obtained from face tracker to a 6-DoF (6 degree of freedom) vector. Six Degree of Freedom(6-DoF) vector composed of three rotation angles(Yaw, Pitch, Roll) and 3 translations (Tx, Ty, Tz) along spatial axis of x, y, and z as in figure.1. The overall architecture of this section can be shown in figure.2.

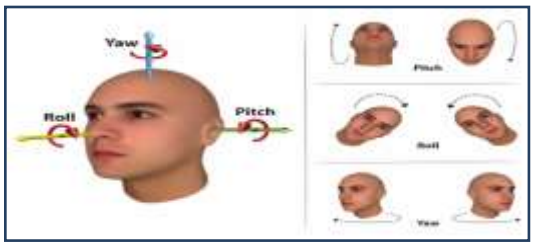

Figure 1: Head Pose Angles

Head detection and tracking should be applied at first, to detect the head in a given video stream. Head tracking is the process of monitoring the continuous changes in tempo-spatial domains. I used library functions in OpenCV and publicly available OpenFace tracker. I applied these functions in C++ to detect and track head pose. OpenFace is a state-of-art tool can perform major tracking tasks in real time. It has ability to track head pose, land mark detection, and eye gaze tracking. In spite of the significant progress in facial behaviors analysis, we can rarely find publically available state-of-art open source tools that perform all above tracking tasks, especially in real time. Additionally, even there are freely available source tools, they require more efforts to re-initialization and optimization [14]. OpenFace tracker was used to track the motion of the points over a live or recorded video stream, the tracker uses Conditional Local Neural Fields (CLNF) for detecting and tracking. CLNF is an example of a Constrained Local Model (CLM). CLM algorithm is easy touse for tracking face and head pose in 3D model with variable appearance and shape. Also, it has ability to work in an uncontrolled environment such as low illumination, head movement in different direction, and occlusion[15].

The dataset used is Boston University (BU) a publicly available dataset as a 3D head training data [22]. This dataset consists of 45 videos recorded with uniform and varying lighting. Each video consist of 200 frames and freely head motion under in-plane and out-of-plane rotations.

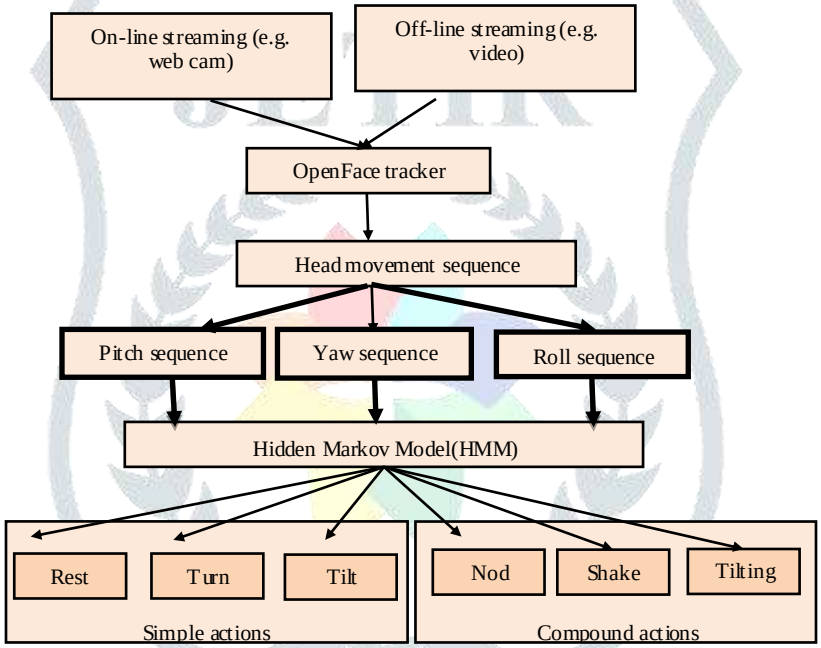

Figure 2 Project's Schematic View

Head Action and Cues

This research uses the term head actions and head pose actions interchangeably to indicate a continuous movement in one direction in a fixed slice of time. On the other hand, Head cues are a prefixed sequence of homogeneous and heterogeneous head actions used by the classifier to extract the head's movement class.

In order to indicate cues time, we estimate that recognizable cue will last for 1 second. BU dataset videos were digitized at 30 frames per second [22], in this case one cue will consume 30 frames. Cues can be classified by the set of actions they hold, so cues must be further divided into a set of fixed time slices, each slice is an action. The number of these time-slices are 6; i.e., each action consume 5 frames and last for approximately 0.166ms.

Classification can take place after reading a whole cue. That means the first cue class, covering frames 0, to 30, will be obtained after 1 second from the beginning of the video, due to first cue reading overhead. The subsequent cues are formed from moving a sliding window of length 30 frames with offset of 1 frame from the start of the previous queue. This approach helps with generating classes equals to the number of the video cues.

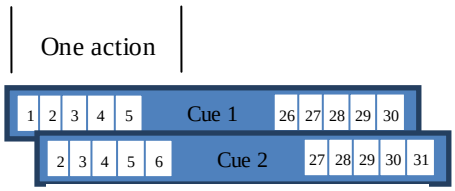

Figure 3: Video actions and cues

The main features to estimate head pose is the pose angles through time. These angles are passed by the head tracker while reading training videos. Head tracker emits two vectors regarding head pose; i.e., translation vector and rotation vector. In this study our concern focused on the rotation vector, which consist of 3 degrees of freedom donating the head rotation about the X-, Y-, and Z-axis. During the rest of this research these angles will be called pitch, yaw, and roll angles presenting rotations on X, Y, and Z axis respectively. These angles act as the primitive components for basic and combined actions. Head tracker's emitted angles are signaled for each frame, which eventually conform a time series of three independent random variables$T(Rx, Ry, Rz)_t$. This series breakdown into three sub series, i.e., TRx(t), TRy(t), and TRz(t). Sub series are used to build the head pose action features vector.

After acquiring an action, the next step is to extract set of symbols to be fed into HMM classifier.To label an action we first categorized the action into two categories, Table 1 describes these actions.-

1- basic head actions (rest, turn, tilt),

2- Compound head actions (nod, shake, tilting).

Table 1: Simple and Combined Actions

Action	**Action direction**	**Repetition**
Basic action		
Rest	Up or Down	Single
Turn	Left or Right	Single
Tilt	Left or Right	Single
Combined action		
Nod	Up and Down	Cyclic
Shake	Left and Right	Cyclic
Tilting	Left and Right	Cyclic

To extract head actions; first the related angle is extracted from the 3DOF rotation vector obtained from the tracker. These angles are labelled per action. The labelling process produce 6 symbols per cue which are used as input to HMM.

IV. HIDDEN MARKOV MODELS (HMM)

HMM is a probabilistic model with random processing that generate outputs of random sequence of observations. HMM is an extension of Markov chain model which has visible states in a direct way for observer. But in HMM these states invisible (i.e., hidden). More explanation about HMM and its application can be found in in Rabiner [11]. HMM is well-known as a probabilistic mathematical model and used for modelling observations sequence in time series. HMMs haven a significant role in applications of speech recognition. Recently, more applications have been adapted HMM like face, head gesture, and hand writing recognition [48][51][37].

There are two main types of HMM; Ergodic, and left-right models or called Bakis model [12]. These two types of are the essential and standard types of HMM, despite the variations in implementing these models. For example, two left-right models are connected together to produce a parallel model. Basically, there is no theoretical evidence about the optimal choice of the type of HMM (ergodic, left-right, or another type), or the number of states (model size), or select of observation symbols (discrete, continuous, single, or multi-mixture). The choices are based on the modeling action [11].

Elements of HMM

The basic elements of HMM are described below:

HMM can be described by $\lambda = (A, B, \pi)$

N: Number of states, $S = \{S_1, \ldots, S_N\}$. Where the state at time t can be denoted as qt

M: Number of different observation symbols, can be denoted as $V = \{v_1, v_2, \ldots v_M\}$

A: state transition matrix (N x N).

$A = \{ a_{ij} \}$

$aij = P\{qt + 1 = Sj | qt = Si\}$…… $1 \leq i, j \leq N$

B: Observation Probability Matrix (N x M)

$B = \{bj(k)\}$

$bj(k) = P[U_k \text{ at } t \mid qt = Si]$

$1 \leq k \leq M$ and $1 \leq j \leq N$

π: initial state probability

$\pi i = p[qi = si]$…… $1 \leq i \leq N$

HMM can be used as a generator to generate sequence of observation O = O1, O2, …,Ot (where each Oi represent one of the symbols in V), this process depends on a convenience value of N, M, A, B, π. For best and useful model that apply in real time, there are three problems that must be solved.

HMM problems

1- The evaluation problem: Given observation O = {O1, O2, O3, …, Ot}, and $\lambda = (A, B, \pi)$. How to compute the probability $P(o \mid \lambda)$ that gave observation O?

2- The Decoding Problem: Given observation O = {O1, O2, O3, …, Ot}, and $\lambda = (A, B, \pi)$. Compute the optimal state sequences?

3- The training problem: Given observation O = {O1, O2, O3, …, Ot}, and $\lambda = (A, B, \pi)$. Adjust $\lambda = (A, B, \pi)$ to maximise $P(o \mid \lambda)$.

Our goal is training a model to solve problem 3, to recognize head gestures.

V. MODEL TRAINING

In this section the training scheme of HMM is based on the Baum-Welch algorithm to ensure that the convergence of the log-likelihood function to a local maximum is always achieved.

Training process starts with collecting tracking data from the videos dataset. The tracking data is separated into three head pose action features vectors (HPAFVpitch, HPAFVyaw, HPAFVroll). The HMM classification module contains three multiclass classifiers, one for each angle to predict simple and combined cues. The classification process is to estimate the class with the maximum likelihood among N classes with known probability pi(X).

$$f(X) = \arg \max_{i \in 1,\dots,N} p_i(X)$$

This function describes the general classification method, it performs with a one crucial condition, the classes' probabilities densities must be known beforehand.

Finding class probabilities is the main purpose for using HMM. Basically Hidden Markov models covers two principles; a model topology, and the statistical parameters matrices. The model topology is designed by the modular to depict and serve the purpose of the underlying scheme. In section(IV) we described the main types of HMM, the model type used in this research is ergodic model due to the little number of states that we need and little number of symbols of observations.

The method deployed in this research is to build one classifier for each angle. Each classifier contains two models (classes), which means a sum of six models all share the same full connected topology with; number of states = 2, and number of symbols = 3.

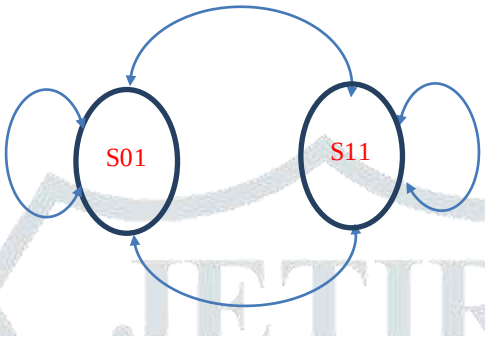

Figure 4: The suggested ergodic model.

The input for the classifiers are sequences of observations (cues) containing six HPAFVs

Opitch = [HPAFVpitch1, HPAFVpitch2, HPAFVpitch3, HPAFVpitch4, HPAFVpitch5, HPAFVpitch6]

Pitch cue symbols alphabet are:

$$\sum \boldsymbol{O}_{pitch} = \{head - up, nill, head - down\}$$

Oyaw = [HPAFVyaw1, HPAFVyaw2, HPAFVyaw3, HPAFVyaw4, HPAFVyaw5, HPAFVyaw6]

Yaw cue symbols alphabet are:

$$\sum \boldsymbol{O}_{yaw} = \{turn - left, nill, turn - right\}$$

Oroll = [HPAFVroll1, HPAFVroll2, HPAFVroll3, HPAFVroll4, HPAFVroll5, HPAFVroll6]

Roll cue symbols alphabet are:

$$\sum \boldsymbol{O}_{roll} = \{tilt - left, nill, tilt - right\}$$

After defining the HMM topology, the model must be learned to estimate the statistical parameters will be used in prediction process.

HMM training data consist of 35 video which 23 of them are in uniform lighting and 12 in varying lighting (table-2).

Preparing the training videos produces 1149 unique cues (457 pitch cues, 335 yaw cues, and 357 roll cues (table-3).

Table 2: Training videos properties

Angle	**#videos**	**#videos with Uniform-light**	**#videos with Varying-light**
Pitch	11	7	4
Yaw	14	9	5
Roll	10	7	3
Total	35	23	12

Table 3: Number of training cues

Angle	**#Unique sequences**
Pitch	457
Yaw	335
Roll	357
Total	1149

The learning process is estimating the statistical parameters matrices. The learning process implement Baum-Welch algorithm with forward-backward passes.HMM learning start with an estimated uniform distributed model parameters, for my model:

$$\lambda = (\mathcal{A}, \mathcal{B}, \pi)$$

Given: $S = \{S0, S1\}$

$$\mathcal{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

$$\mathcal{B} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\pi = |1/2 \quad 1/2|$$

Where λ is the hidden Markov model, S is the model states, $\mathcal{A}$ is the states' transition matrix, $\mathcal{B}$ is the emission matrix, and π is the model's initial probabilities.

The algorithm first perform a forward pass to update the transitions probabilities matrix. Given a sub-observation o1,…,ot that output state i at time t, as defined by $a_i(t) = P(O_1 = o_1, \dots, O_t = o_t, Q_t = i|\lambda)$.Calculating $a_i(t)$ is a recursive process in a sense that calculating a state frequency at current time depends on the frequency at previous time. Recursive processes always start with an initialisation step, then updating the obtained value depending on backward iteration method. A new iteration will now started with the updated model. The model continues to be updated until convergence. That means there is no significant difference between the previous and current iteration. After the all parameters for each model are estimated, then can be used them in the classifier module.

VI. GESTURES CLASSIFICATION

There are two main categories to classify a set of inter-related classes; One-vs-All classification, and All-vs-All classification. One-vs-All classification suits binary classifier with classes hold only the positive training points for this class. The method needs classifiers equal to the number of model classes and classify using

$$f(X) = \arg\max_i f_i(X)$$

All-vs-All method uses N(N-1) classifiers, each classifier dedicated to differentiate pairs of classes i and j. the training points are separated by placing positive points into class i, and the negative points into class j. The classification process is

$$f(X) = \arg\max_i \sum_i f_{i,j}(X)$$

The classifiers work after getting the tracking information for each angle and getting cues symbols. Each angle symbols cue enters their designated classifier to get the class with max log-likelihood for the cue Fig.5.

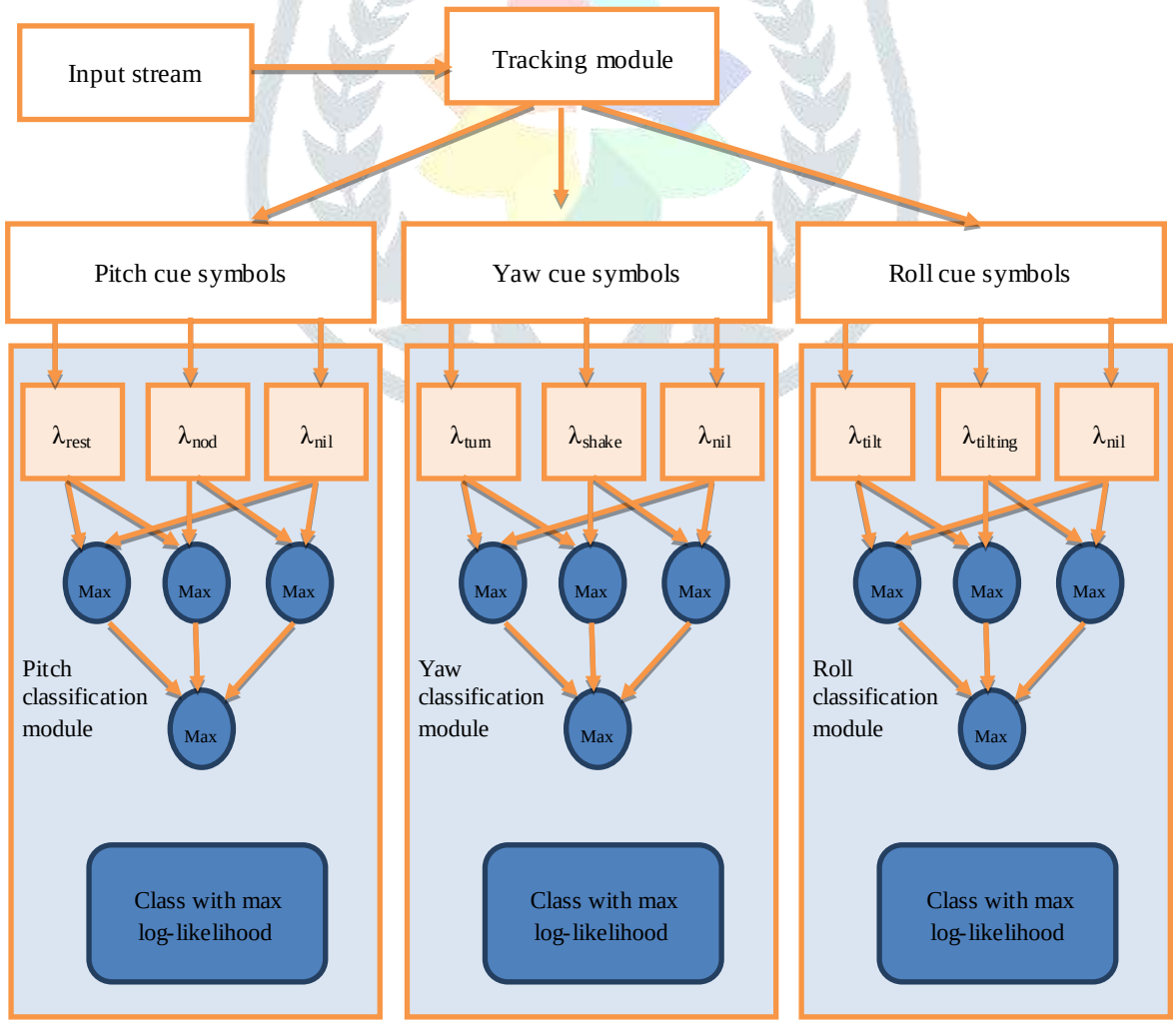

Figure 5: Classification workflow.

VII. RESULTS AND EVALUATION

I used free available videos from "Mind Reading" dataset to carry out the tests on HMMs classifiers (node, shake, tilting, rest, turn, and tilt). These videos contain complex movement of head; the challenge was to classify each gesture with respect to its pattern. Example of the results is shown in Fig. 6 of an available free video from mind reading DVD on https://www.autismresearchcentre.com/arc_tests.The classifier detects head basic movement rests, turns, and tilt. Also, classification of combined movements to identify nod, shake, and tilting. Classification results are based on maximum log-likelihood of the HMM. The video shows different head movements happened during variety of durations. The video presents reassured emotion, the classifier detected nodding, shaking, rest, tilt, and tilting from this video. The classification results approve the importance of head features to support emotion recognition in this video or any other dynamic video. From the figure, each combined cue consists of a cyclic period of consecutive actions at least two actions. Each action has onset, apex, and offset periods. Onset refers to the stronger starting of action. Apex is the action at peak and there are no more changes. Offset refers to the slow-down of action where there is no sign of action. Basic cues represented by a single action.

To evaluate the system efficiency, I tested the system online, the test carried out on 11 live streams sessions. The streams contain a combination of simple and combined cues throughout Fig.7. The figure below depicts the results for one video stream in real time, the video sampled at 30fps.

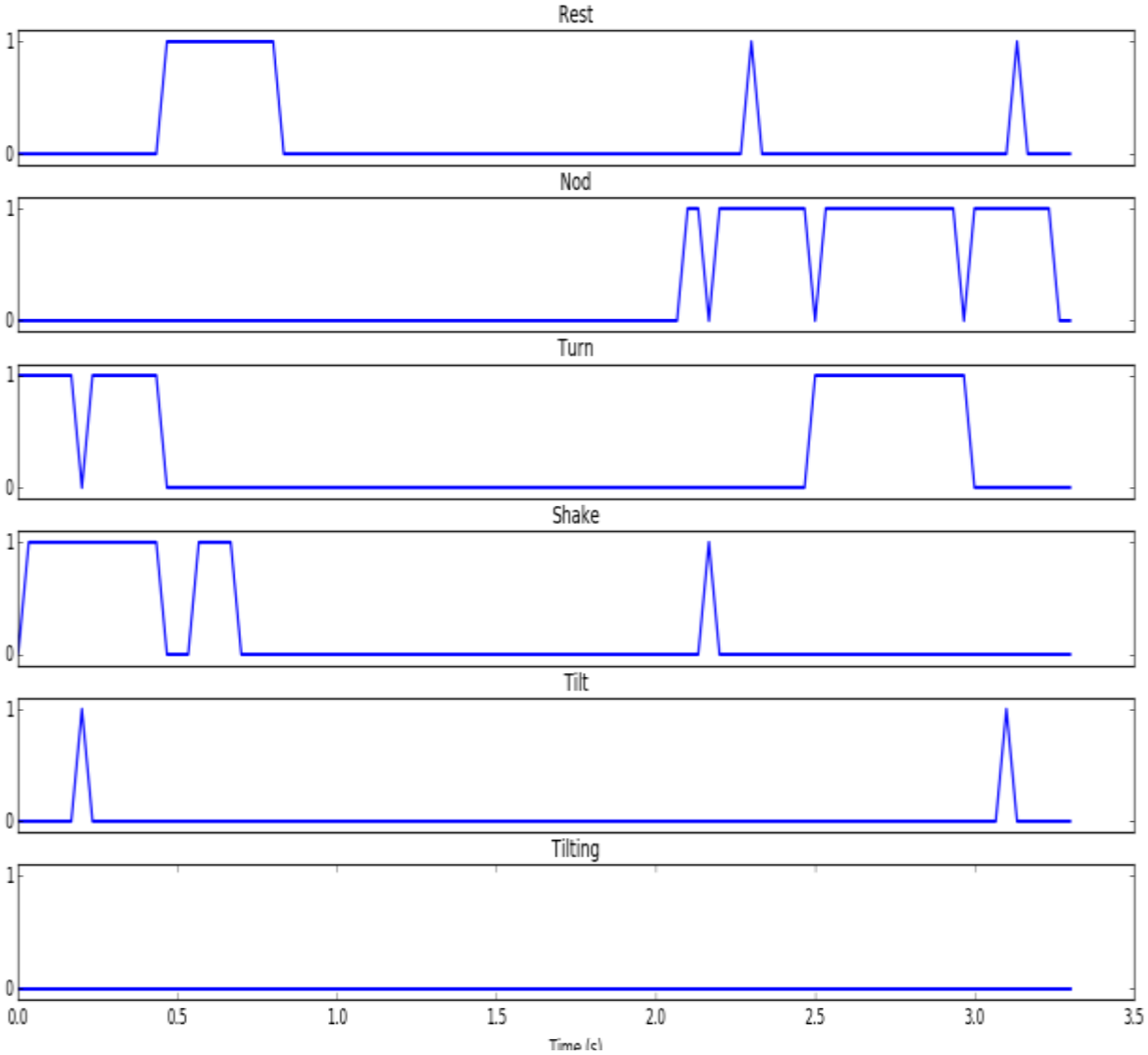

Figure 6: Classification results from a video of mind reading DVD. the video represents reasured emotion. Each cue represented by 0 and 1, number 1 refer to the action acquire and 0 refer there is no Off-line streaming (e.g. video)

The lost classification happened when the head movement of pitch, yaw, and roll exceeds the determined system's threshold. The classifier accuracy can show in table-4. Column 1 presents the true positive (TP), that depicts classification rate for each class from combined and basic gestures. 3For example, TP for nod represents number of nod correctly classified as node. Second column states the false positive (FP), is the number of gestures that incorrectly classified as nod.

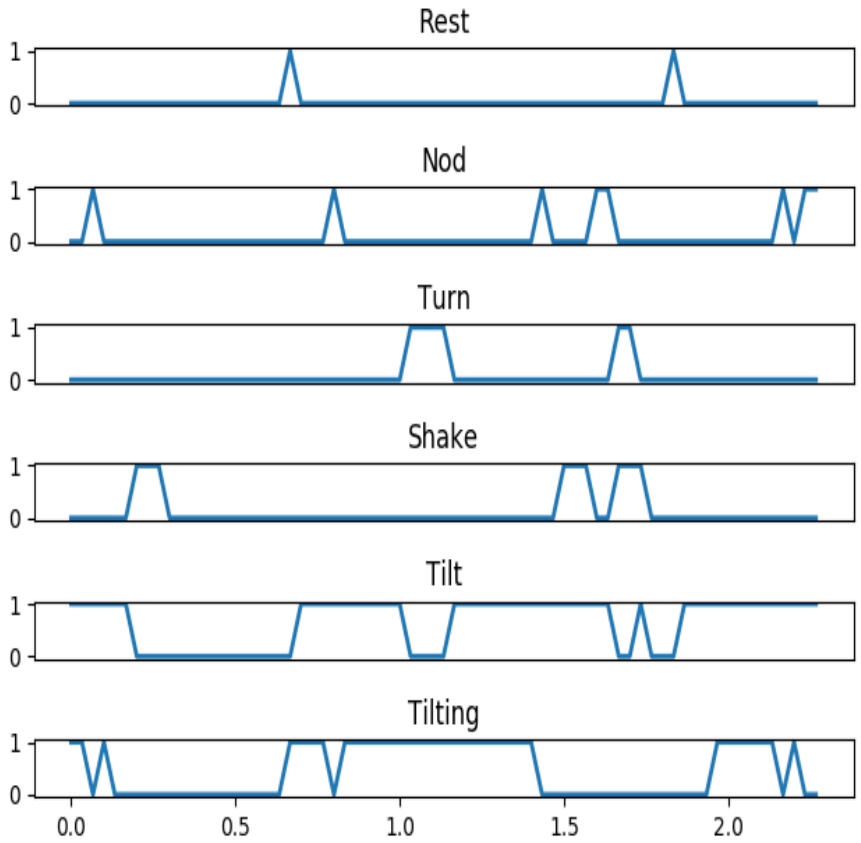

Figure 7: Classification results from online video stream.

These results show a good performance achieved using only head pose (pitch, yaw, and roll) and a webcam. The difficulty faced this work is scarce datasets for head combined and basic movements. The work based on training on the simple movement of head and testing on complex head movements from mind reading data. Other challenge was little sources of data of head tilting, as most works focused on head nodding and shaking. The model efficiency depend on quantity of trained data, where high efficiency required more data.

Table 4: Accuracy results: TP is rate of correctly classified. FP is rate of incorrectly classified.

Gestures	**TP**	**FP**
Nod	0.99	0.007
Shake	0.87	0.12
Tilting	0.96	0.039
Rest	0.94	0.05
Turn	0.83	0.16
Tilt	0.92	0.07

VIII. CONCLUSIONS

We have developed a new model that have ability to classify basic and combined movements from complex head motion using 3DoF(pitch, yaw, and roll). These angles represents head direction in consecutive frames to form a set of observations to input the all-vs-all HMMs classifier. The system used OpenFace tracker and trained and tested on freely available BU dataset and mind reading videos respectively. For system evaluating, we tested the classifier on online video stream in real time by using laptop webcam under varying illumination and uncontrolled environment. The system developed and evaluated by using free and simple techniques without incurring any cost. So, the system accuracy can be best achievement comparing with accuracy for other systems, which extracted head nod and shake based on eye location and specific feature points by using sensitive instruments.

For the future work, we need to increase the classifier accuracy by getting more complex data for training and testing. Also, this system can be used as a seed to detect complex affective state in dynamic video and video stream.

REFERENCES

[1] F. Althoff, R. Lindl, L. Walchshausl and S. Hoch, "Robust multimodal hand-and head gesture recognition for controlling automotive infotainment systems," VDI BERICHTE, vol. 1919, p. 187, 2005.

[2] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 4, pp. 607--626, 2009.

[3] A. Mignault and A. Chaudhuri, "The many faces of a neutral face: Head tilt and perception of dominance and emotion," Journal of nonverbal behavior, vol. 27, no. 2, pp. 111-132, 2003.

[4] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187--194, 1999.

[5] R. A. El Kaliouby, Mind-reading machines: automated inference of complex mental states, Cambridge: PhD, thesis, Citeseer, 2005.

[6] J.-G. Wang and E. Sung, "EM enhancement of 3D head pose estimated by point at infinity," Image and Vision Computing, vol. 25, no. 12, pp. 1864--1874, 2007.

[7] A. Kapoor and R. W. Picard, "A real-time head nod and shake detector," Proceedings of the 2001 workshop on Perceptive user interfaces, pp. 1-5, 2001.

[8] W. Tan and G. Rong, "A real-time head nod and shake detector using HMMs," Expert Systems with Applications, vol. 25, no. 3, pp. 461--466, 2003.

[9] J. W. Davis and S. Vaks, "A perceptual user interface for recognizing head gesture acknowledgements," Proceedings of the 2001 workshop on Perceptive user interfaces, pp. 1--7, 2001.

[10] A. Adams, M. Mahmoud, T. Baltrusaitis and P. Robinson, "Decoupling facial expressions and head motions in complex emotions," International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 274--280, 2015.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257--286, 1989.

[12] R. Bakis, "Continuous speech recognition via centisecond acoustic states," The Journal of the Acoustical Society of America, vol. 59, no. S1, pp. S97--S97, 1976.

[13] R. A. El Kaliouby, "Mind-Reading Machines: automated inference of complex mental states," University of Cambridge, Cambridge, United Kingdom, p. 185, 2005.

[14] T. Baltru, P. Robinson, L.-P. Morency and others, "OpenFace: an open source facial behavior analysis toolkit," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1--10, 2016.

[15] T. Baltrusaitis and C. T. Baltrusaitis, "Automatic facial expression analysis," University of Cambridge, Computer Laboratory, Technical Report, no. UCAM-CL-TR-861, 2014.

[16] R. P. Gaur and K. N. Jariwala, "A survey on methods and models of eye tracking, head pose and gaze estimation," India, 2014.

[17] C. a. Z. Z. Zhang, A survey of recent advances in face detection, Technical report, Microsoft Research, 2010.

[18] M.-H. Yang, D. J. Kriegman and N. Ahuja, "Detecting faces in images: A survey," IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 1, pp. 34--58, 2002.

[19] S. Li, X. Zou, Y. Hu and Z. Zhang, "Real-time multi-view face detection, tracking, pose estimation, alignment, and recognition," IEEE International Conference on Computer Vision and Pattern Recognition, 2001.

[20] M. J. Jones and P. Viola, "Robust real-time object detection," Workshop on statistical and computational theories of vision, p. 56, 2001.

[21] R. Chellappa, C. L. Wilson and S. Sirohey, "Human and machine recognition of faces: A survey," Proceedings of the IEEE, vol. 83, no. 5, pp. 705--741, 1995.

[22] M. La Cascia, S. Sclaroff and V. Athitsos, "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 4, pp. 322--336, 2000.

[23] J. Xiao, S. Baker, I. Matthews and T. Kanade, "Real-time combined 2D+ 3D active appearance models," CVPR (2), pp. 535--542, 2004.

[24] A. Gee and R. Cipolla, "Determining the gaze of faces in images," Image and Vision Computing, vol. 12, no. 10, pp. 639--647, 1994.

[25] Y. G. Kang, H. J. Joo and P. K. Rhee, "Real time head nod and shake detection using HMMs," International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 707-714, 2006.

[26] C. Chris, "Our head movements convey emotions," 2015. [Online]. Available: https://www.mcgill.ca/newsroom/channels/news/our-head-movements-convey-emotions-256366. [Accessed 17 8 2016].

[27] J. Foytik and V. K. Asari, "A two-layer framework for piecewise linear manifold-based head pose estimation," International journal of computer vision, vol. 101, no. 2, pp. 270--287, 2013.

[28] S. Srinivasan and K. L. Boyer, "Head pose estimation using view based eigenspaces," 16th International Conference on Pattern Recognition, pp. 302--305, 2002.

[29] E. Seemann, K. Nickel and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings, pp. 626-631, May 2004.

[30] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," Fifth IEEE International Conference on Automatic Face and Gesture Recognition. Proceedings. , pp. 255--260, 2002.

[31] H. a. K. T. Schneiderman, "A statistical method for 3D object detection applied to faces and cars," IEEE Conference on Computer Vision and Pattern Recognition., pp. 746--751, 2000.

[32] J. M. Rehg, M. Loughlin and K. Waters, "Vision for a smart kiosk," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 690--696, 1997.

[33] D. Paul, "A speaker-stress resistant HMM isolated word recognizer," IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'87, pp. 713--716, 1987.

[34] A. Pentland, B. Moghaddam and T. Starner, "View-based and modular eigenspaces for face recognition," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'94, pp. 84--91, 1994.

[35] K. Nickel, E. Scemann and R. Stiefelhagen, "3D-tracking of head and hands for pointing gesture recognition in a human-robot

interaction scenario," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 565--570, 2004.

[36] F. Moreno, A. Tarrida, J. Andrade-Cetto and A. Sanfeliu, "3D real-time head tracking fusing color histograms and stereovision," International Conference on Pattern Recognition, pp. 368--371, 2002.

[37] C. Morimoto, Y. Yacoob and L. Davis, "Recognition of head gestures using hidden Markov models," International Conference on Pattern Recognition, pp. 461--465, 1996.

[38] O. Kwon, J. Chun and P. Park, "Cylindrical model-based head tracking and 3D pose recovery from sequential face images," International Conference on Hybrid Information Technology, ICHIT'06., 2006.

[39] T. Horprasert, Y. Yacoob and L. S. Davis, "Computing 3-d head orientation from a monocular image sequence," International Conference on Automatic Face and Gesture Recognition, pp. 242--247, 1996.

[40] C. Huang, X. Ding and C. Fang, "Head pose estimation based on random forests for multiclass classification," International Conference on Pattern Recognition (ICPR), pp. 934--937, 2010.

[41] G. Guo, Y. Fu, C. R. Dyer and T. S. Huang, "Head pose estimation: Classification or regression?," 19th International Conference on Pattern Recognition, pp. 1--4, 2008.

[42] B. Heisele, P. Ho and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," IEEE International Conference on Computer Vision, pp. 688--694, 2001.

[43] B. Heiselet, T. Serre, M. Pontil and T. Poggio, "Component-based face detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I--657, 2001.

[44] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 232--237, 1998.

[45] L. M. Brown, "3D head tracking using motion adaptive texture-mapping," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I--998, 2001.

[46] G. J. Edwards, C. J. Taylor and T. F. Cootes, "Interpreting face images using active appearance models," IEEE International Conference on Automatic Face and Gesture Recognition, pp. 300--305, 1998.

[47] R. El Kaliouby and P. Robinson, "Real-Time Inference of Complex Mantal States from Facial Expressions and Head Gestures," Computer Vision and Pattern Recognition workshop, pp. 154 - 154, 27 July 2004.

[48] N. Oliver, A. P. Pentland and F. Berard, "Lafter: Lips and face real time tracker," Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pp. 123--129, 1997.

[49] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," Automatic Face \& Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 30--35, 2011.

[50] G. Littlewort, M. S. Bartlett and I. Fasel, "Dynamics of facial expression extracted automatically from video," Image and Vision Computing, vol. 24, no. 6, pp. 615--625, 2006.

[51] J. J. Lien, T. Kanade, J. F. Cohn and C.-C. Li, "Automated facial expression recognition based on FACS action units," Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pp. 390--395, 1998.

[52] U. M. Erdem and S. Sclaroff, "Automatic detection of relevant head gestures in American Sign Language communication," International Conference on Pattern Recognition, pp. 460--463, 2002.

[53] L. P. Morency, P. Sundberg and T. Darrell, "Pose estimation using 3D view-based eigenspaces," IEEE International Workshop on Analysis and Modeling of Faces and Gestures, pp. 45--52, 2003.

[54] R. El Kaliouby and P. Robinson, "Real Time Head Gesture Recognition in Affective Interfaces," Interact, 2003.