Machine Learning for Single and Complex 3D Head Gestures: Classification in Human-Computer Interaction

Dr. Amina Atiya Dawood
Computer Center, University of Babylon, Iraq.
Email: amina@uobabylon.edu.iq

Balasem Alawi Hussain
Computer Center, University of Babylon, Iraq, Email: balasem@uobabylon.edu.iq

Received August 27, 2021; Accepted December 05, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19095

Abstract

This paper presents a new Hidden Markov Model based approach for fast and automatic detection and classification of head movements in real time dynamic videos. The model has been developed to utilize human-computer interaction applications by using only the laptop webcam. The proposed model has the ability to predict single head and combined simultaneously in fast responses. Other models paid more attention to classify head nod and shake only, but our model contribute the role of other head movements. The model proposed here doesn’t need any user intervention or previous knowledge of its environment. In addition, there is no limitation on illumination changes and occlusions, as well as no restrictions on head movements ranges. The model achieved significant results and efficient performances when tested on unseen data. As the model accuracies were 94%, 99%, 83%, 87%, 93%, 96% for all head gestures (rest, nod, turn, shake, tilt and tilting) respectively. On the other hand, the model accuracy was 99% and 88% for combined and single cues respectively. The aim of this model is to provide a fast application to infer and predict human emotions and affective states in real time through head gestures.

Keywords

Machine Learning, Head Movements, Head Pose Estimation, Classification, Head Gestures, HCI.

Introduction

Recently, computers had a great role in all our daily life, and people tend to spend more time with computers. Computing has motivated scientists and researchers to develop more interactive and responsive computers by calibrating and facilitating machine hardware and
applications (Baltrusaitis & Baltrusaitis, 2014) (Busso, Mariooryad, Metallinou, & Narayanan, 2013). Research attentions were toward making computer interface more natural and responsive. Therefore, the focus was to make computers that can detect and infer emotions and behaviors of their users. People use different communication channels to communicate with their peers or with computers. These communication channels include non-verbal cues, such as facial expressions, head movement, and eye gaze, in addition to posture and body movements which are used to interact with others (Baltrusaitis & Baltrusaitis, 2014) (Tian, Kanade, & Cohn, 2005). In recent years, researchers and psychologists developed different state-of-the-art HCI techniques. Almost all of these techniques focused on using faces as a communication channel tool, due to the fact that human faces convey more effective expressive information (Zeng, Pantic, Roisman, & Huang, 2009). Great efforts have been devoted to analysis human facial expressions, posture, and gestures, which interested in analyzing facial expressions by using specific facial feature points such us eyes, eyebrows, and mouth corners. In spite of the important role of faces in human-machine interaction, head can be considered as a valuable communication tool. Head movements involve different meanings of information, especially during interacting with others or computers. For example, studies have indicated that head nodding during interacting or speaking means agree while head shaking means disagreement (Althoff, Lindl, Walchshausl, & Hoch, 2005). As well as the head provides valuable information in understanding the emotions, affective state, and people behaviors. This information considers assistive means in different applications such as HCI, affective computing, computer vision, and virtual reality. Prediction and classification of human head movements are difficult and need more efforts (Murphy-Chutorian & Trivedi, 2009). However, inferring head rotation and direction through head pose needs a lot of computational process for head estimation (Gaur & Jariwala, 2014). Therefore, advances of different techniques were deployed to predict and classify head features (Althoff, Lindl, Walchshausl, & Hoch, 2005) (Anitta & others, 2021) (Chris, 2015) (Huang, Ding, & Fang, 2010) (Tan & Rong, 2003). However, most of these studies have focused on extracting specific head movements such as head nodding and shaking, and don’t take into account the other head movements which have the same role and importance, like (tilting, tilt, rest, up, dawn). For example, head dawn movement is related with shame, anxiety, and disengagement (Mignault & Chaudhuri, 2003). Therefore, this paper proposes a new model has the ability to detect and predict all head gestures (combined and single cues). The proposed model based on Hidden Markov Model (HMM) to classify all head movements from sequences of videos in real time. Tracking the head in real time depended on three degrees of freedom (3DoF) with three head axes (X, Y, Z). The goal of building such a model that has an ability to predict head movements is to serve as an assistance tool that
recognize and classify an emotional and affective state in real world applications that interact with computers. It estimates the relative user situations through the head movements in consecutive video frames.

The remainder of this paper is organized as follow; section II described the most relative and important previous works that focused on head movements detection and prediction, the methods which were used in this experiment and model building and designing was analyzed and discussed in section III. Study results are presented in IV section; and section V concluded this work.

Previous Works

To detect and classify head gestures, first, head should be tracked in space. A lot of works used different methods to track head pose. These methods considered the head as a rigid object has two or three dimension (rotation and translation) or by using the facial features points for face to track and estimate head pose in plane-in and plane-out. For instance the methods which considered head as a rigid object, stated that we should track the whole head in two region (two dimensional or three dimensional) to detect head orientation and translation (Srinivasan & Boyer, 2002) (Raja, Vali, Palipana, Michelson, & Sigg, 2020). One of the first examples of these methods can be seen in the works of (Kwon, Chun, & Park, 2006) (La Cascia, Sclaroff, & Athitsos, 2000), the researchers used cylindrical model for 3D head rotations to extract three head orientations and translations. Another approach depended on head as a rigid object clearly appeared in a work developed by Blanz and Vetter (Blanz & Vetter, 1999). They used morphable model (3D model) for synthesis faces from a 3D dataset. While the researchers in (Huang, Ding, & Fang, 2010) suggested a random forest technique with linear discriminate analysis to track head pose from 2D head dataset.

In (Luo, Zhang, Yu, & Wang, 2019) they estimated the direction and translation of peoples’ head in 3D. The researchers used depth image technique to obtain 3D head pose and face model.

3D head tracking was used to monitor the driver situation, the researcher joined features reduction and time frequency analysis from millimeter-wave Doppler radar (Raja, Vali, Palipana, Michelson, & Sigg, 2020). Another study used stereo vision camera to capture head pose between people and a robot, this method depended on depth image of 2D head (Seemann, Nickel, & Stiefelhagen, 2004). According to these works, it was observed that
the head pose estimation is more robust and accurate in 3D than 2D. However, it requires more efforts to label head manually and also time consuming (El Kaliouby R.A., 2005).

Generally, it can be shown that many studies relied on the methods that depended on facial features points, where the researchers in these methods adapted geometrical analysis of face shape to find head pose (orientation and translation). They extracted facial features from head tracking data such as eyes corners (inner and outer), tip of nose, and mouth corner (Moreno, Tarrida, Andrade-Cetto, & Sanfeliu, 2002) (Xiao, Baker, Matthews, & Kanade, 2004). These methods like the before mentioned methods have a good accuracy to find head rotation, but they faced different challenges like illumination changes, and head motion out of plane which lead to features misclassification (Wang & Sung, 2007) (Guo, Fu, Dyer, & Huang, 2008) (Murphy-Chutorian & Trivedi, 2009).

To overcome the limitations of the previous methods, hybrid methods were used by combining head rigid in 3D or 2D with facial features points like in (Anitta & others, 2021) (Giannakakis, Manousos, Chaniotakis, & Tsiknakis, 2018) (Guo, Fu, Dyer, & Huang, 2008) (Madrigal & Lerasle, 2020).

To classify head gestures like pitch, yaw, and roll and their combinations in hybrid methods, we should use different technical tools to achieve this classification. These tools varied based on number of gestures; types of inputs; and the methods which will be used for data training. The tracker was used by this study depended on hybrid methods in head tracking and detecting.

After introducing the tracking and detecting of head pose estimation. Now, we will present the previous works that were developed for head gestures classification. For instance, nodding and shaking were detected from video streams in (Tan & Rong, 2003), Tan and Rong model based on eye features like the location of inner and outer eyes corners to detect head orientation. The model designed using HMM with accuracy 89% and 82% for shaking and nodding respectively. In similar way, HMM was used in (Kapoor & Picard, 2001) to classify head nod and shake. The researchers used infrared camera to determine eyes position which were adapted to detect head pose, then classify node and shake movements.

Facial features points were extracted from video in (Turabzadeh, Meng, Swash, Pleva, & Juhar, 2018) to classify single movement of head, the researchers used local binary point technique to capture features points. These features then moved as input for K-nearest algorithm to classify head motion. As well as, another work used HMM for each one of these observations of head (left, right, up, down, in, out, and rest) to classify head gestures.
which recognized by the words (yes, no, hello, and maybe) (Morimoto, Yacoob, & Davis, 1996). Other study used finite state machine to extract acknowledgements of users through their head gestures. The gestures were captured using IBM pupil camera (Davis & Vaks, 2001). (El Kaliouby & Robinson, 2003), they used the same idea of Davis & Vaks, to design a model which can detect nod and shake according to yes or no movements respectively. The researchers in this study used four actions like begin, finish, up, and down to determine head gestures. The movements of head in this model should start with up or down movement.

Head gestures can be used as an important factor to recognize the stress affect, like in study developed by (Giannakakis, Manousos, Chaniotakis, & Tsiknakis, 2018). The researchers used active appearance model (AAM) to detect head movement which are relative to stress situation. The model used the facial features points and 3D head pose for classification.

However, to the authors best knowledge most of previous studies have classified head nod and shake only, they have also some restrictions on head movements like head should be neutral or in plane-in or the head motion should start from rest, in addition, they assumed there is no changes in illumination, and occlusion.

Researchers like (Adams, Mahmoud, Baltrusaitis, & Robinson, 2015) stated that head gestures are not only nodding and shaking but extend beyond these gestures like tilting, rest, tilt and turn. In addition, all gestures have the same important role to express and communicate. Therefore, this work was coined to develop HMM model to extract all head gestures like single (rest, tilt, up, dawn, left, right, and turn) and combined gestures (nod, shake, tilting). The developing model based on natural-spontaneous affective dataset which have no limitation on head pose (direction and rotation) or face movement, and illumination changes. The model has the ability to predict and classify all head movements in real time by using laptop webcam and without any extra cost and intervention.

**System Performance**

**Head Pose Estimation**

Firstly, we worked to track a head in a video stream, as the tracking process helps us to captures the spatial-tempo features of continuous changes. OpenFace tracker was used in this work, it is an available state-of-art head tracker tool in real time. The tracker based on facial features of face combined with 3D head pose - more information about this tracker are in (Baltru, Robinson, Morency, & others, 2016). We didn’t use all the toolkit for this tracker, only its (C++) functions were used for tracking and detecting head posing. There
is a significant development in the field of facial expressions and emotions recognition, but it is difficult to find an open-source tracker available that serve all tracker functions in real time. There are many head tracker tools available, Nevertheless, there is no open-source code doesn’t need re-implement and optimization to work in real world (Baltru, Robinson, Morency, & others, 2016).

The proposed model was trained on the publicly available Natural Affective-Cognitive Dataset (NACD). The dataset comprises natural spontaneous data for students with Asperger Syndrome (AS) and Typical Development (TD), interacting with a computer (Dawood, Turner, & Perepa, Natural-Spontaneous Affective-Cognitive Dataset for Adult Students With and Without Asperger Syndrome, 2019). The data was collected and annotated for affective-cognitive states prediction in real time (Dawood, Turner, & Perepa, Affective computational model to extract natural affective states of students with Asperger syndrome (AS) in computer-based learning environment, 2018). This dataset was recorded in a computer-based learning environment without using any wearable sensors. It characterized by an uncontrolled environment, without any limitation on head pose changes, talking, hair and facial styles, occlusions, glasses, and background and illumination changes - more information about this dataset can be found in (Dawood, Turner, & Perepa, Natural-Spontaneous Affective-Cognitive Dataset for Adult Students With and Without Asperger Syndrome, 2019).

**Head Action and Cues**

Head pose and head action terms are used interchangeably as an indicator for the continuous of head cues in a fixed time interval with one direction. Head cues consist of symmetric and asymmetric group of actions, which can be used in head movements classification. Estimation time of one cue to occur and classify by a classifier requires one second. The videos in NACD dataset contained 30 frames per second, in our model that means each cue required 30 frames to happen. Her one cue can be divided into a group of fixed time parts, and one part represents one action. Each action in this study consist of 5 frames and one cue requires 6 actions and continue for approximately 0.166ms.

The classification will start after reading one entire cue, which is covering frames 0 to 30. The detection of the first class of the cue will happen after 1 second from the start of first cue action. To analyze continuous head changes, a window of size 30 frames/second was selected, the choosing of this window was based on different trials to predict the best class from gestures. The cues reading occurred by sliding the window sequentially frame by frame from the beginning of the previous queue. Starting from the previous frame a window
of 30 frames was slides with offset of 1 frame. This way, number of classes yielded equal to the video cues.

Varies in head pose angles, generally, help to extract temporal features. When the head tracker reads the training video, it transmits the captured pose angles conveyed in two vectors, which are, specifically, translation and rotation vectors. The transformation vector concerns our study was the rotations vectors. The rotations considered were around the X, Y, and Z axes. Pose angles referred to as pitch, yaw, and roll to correspond to X, Y, and Z rotations, respectively, and this distinction will be followed throughout this paper. Complex and elementary actions are composed from the before mentioned angles. For the purpose of pose computations, the captured pose is transit to help in forming a three independent multivariable time series, T(RX, RY, RZ)(t). The three components are also time series with a value for each time instance, it implies that the pos at time t is the feature point [RX(t), RY(t), RZ(t)].

The head pose classifier used in this paper is the HMM classifier. The classifier needs well defined labeled states. Therefore, the next step was to define the states labels from the dynamics of head movements as a 3DOF system. Actions were grouped into two categories, i.e., basic head actions, and compound head actions.

The classifier used in this paper is a probabilistic model used discrete-states and discrete-time modeling of a Markovian chain, called Hidden Markov Model (HMM) but here there are states which are hidden or unobserved Rabiner (Rabiner, 1989). HMM model is best suited for timeseries forecasting, and proven its feasibility in fields like speech recognition, hand gesture, and handwriting recognition (Oliver, Pentland, & Berard, 1997) (Lien, Kanade, Cohn, & Li, 1998) (Morimoto, Yacoob, & Davis, 1996).

Hidden Markov model can be represented by the parametric model λ= (A, B, π); where, A is the state transition matrix, B is the observation transition probability matrix of states, and π is the probability of the initial state. The ability of generating sequences from the defined HMM as it is also a generative model, the output here is the observations sequence.

HMM has several well-known restrictions, the first one is computing the observations probabilities matrix B from the set of given observation, O. Computing the optimum sequence of states from the given observations O and model λ, is the next problem. Finally, maximizing the probability of predicting the input observations from the model parameters, which is an objective to solve, in order to predict head gestures.
Model Training

Extracting complex features to classify head gestures require solid and robust classifier. Therefore, Baum-Welch algorithm was used in this study due to its solidity in real-time sequences, and to estimate the maximum likelihood of the features in HMM. The classification process modeling starts with a training phase using different samples of features -which are extracted from head tracker. These features were divided into three vectors for angles: pitch, yaw, and roll. For best prediction, we assigned three multiclass HMMs for each vector. The goal of the classification algorithm is to obtain the best prediction of the class with highest probability \( P(x_i) \) of \( N \) classes. Classification algorithms like Baum-Welch algorithm ensures that a good log-likelihood convergence to a local-maxima is always obtained. The initiation of the training process begins after gathering tracking features in the dataset videos. The training data is to be feed to the three HMM, one for each head angle vector. The goal of the classifier is to reach the maximum log-likelihood for the true class.

\[
f(X) = \arg \max_{i \in 1, \ldots, N} p_i(X|\lambda)
\]  

(1)

The classification function (Eq. 1) returns the class with highest likelihood, \( p_i(X|\lambda) \), where the likelihoods are the probabilities that the model predicted. Primarily, the aim of implementing the HMM is to find the predictions of the classes with highest probability to discriminate the true class.

HMM models have several types, in the subsequent section a description of the main types of HMM is provided and an indication of the model used in this paper.

The concept followed in this research assigns one ensemble model for each head pose angle. Furthermore, each ensemble model comprises of three HMM models, i.e., one for each head pose angle, with three states and three symbols. The ensemble model takes its input in the form of sequences of timed head pose cues -six slices of time.

The construction of the observation cue \( O_{\theta}(t) \) indicate the observation cue at time \( t \) for head pose angle \( \theta \), where \( \theta \) can take the values; pitch, yaw, roll. Consequently, the obtained cues take the following shapes.

\[
O_{\text{pitch}}(t) = [pitch(t), pitch(t+1), pitch(t+2), pitch(t+3), pitch(t+4), pitch(t+5)]
\]

(2)

\[
O_{\text{yaw}}(t) = [yaw(t), yaw(t+1), yaw(t+2), yaw(t+3), yaw(t+4), yaw(t+5)]
\]

(3)

\[
O_{\text{roll}}(t) = [roll(t), roll(t+1), roll(t+2), roll(t+3), roll(t+4), roll(t+5)]
\]

(4)
The corresponding cue symbols alphabet, \( \Sigma \), for each angle cue is illustrated below

\[
\Sigma O_{\text{pitch}} = \{ \text{head – up, nill, head – down} \} \tag{5}
\]
\[
\Sigma O_{\text{yaw}} = \{ \text{turn – left, nill, turn – right} \} \tag{6}
\]
\[
\Sigma O_{\text{roll}} = \{ \text{tilt – left, nill, tilt – right} \} \tag{7}
\]

Note that until this stage, the model cannot do any useful work. The designed model should be learned to be able to do useful classification tasks. Generally, the learning task is to fit the model parameters around the input observation cues data. The implemented learning algorithm, i.e., Baum-Welch algorithm, has forward and backward passes. First the model is provided with initial values for the parameters -assuming an equiprobability is reasonable because the head can start from any state with equal probability. The model assumptions and specifications are as follows. For the model \( \lambda(T, E, \pi) \), given:

\[
S = \{ S0, S1, S2 \} \tag{8}
\]
\[
T = \begin{bmatrix}
1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3
\end{bmatrix} \tag{9}
\]
\[
E = \begin{bmatrix}
1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3 \\
1/3 & 1/3 & 1/3
\end{bmatrix} \tag{10}
\]
\[
\pi = [1/3 \ 1/3 \ 1/3] \tag{11}
\]

The hidden Markov model \( \lambda \) have the following entities; \( S \) is the model states, \( T \) is the state transitions matrix, \( E \) is the emission matrix, and \( \pi \) is the model’s initial probabilities.

The initial transition probabilities are updated in the first forward pass. Furthermore, calculating an output state in time \( t \) depends only on the previous state (Markov property for Markovian chains) considered a recursive process -the feedback pass.

**Gestures Classification**

Multiclass environments with inter related classes can be classified using one-vs-all or all-vs-all classifications. Generally, binary classification tasks use one-vs-all classification because it follows the logic of the problem specifically, we need one class to be nominated among many other classes. The classification model in Eq. 12, the required class is the one with the highest probability, i.e., having maximum-likelihood.

\[
f(X) = \arg \max_i f_i(X) \tag{12}
\]
On the other hand, all-vs-all method uses $N(N-1)/2$ classifiers, using class dedicated classifiers to construct 2-tuple classifiers. The target classification will be the arg-max of their result. Finally, the resulting classification will be Eq 13.

\[
f(X) = \arg\max_i \sum_j f_{i,j}(X) \quad (13)
\]

The illustration in Fig. 1 is the HMM classification model used in this paper.

![HMM classification model used in this paper](image)

**Results and Evaluation**

The number of videos used in this study was a dataset of 569 videos. For training and testing the proposed model, we first set aside 5% of the dataset for generality and test then the remaining dataset was divided into two parts (80% and 20%) for training and validation purposes. The purpose of the model was to classify head movements based on combined movement variations (nod, shake, and tilting) and single movements (turn, tilt, left, right, up, dawn and rest). Furthermore, the function of the model was to capture intellectual features from the input video-clip and produce the likelihood probability for each class in parallel to the time-series of the video-clip frames. The class with maximum probability considered as the current gestures of the video-clip under consideration.

The 5% ratio that was excluded from the training dataset before the model starts its training phase and considered as the unseen data. Principally, the purpose of the unseen data was to challenge the model performance to predict the desired classes from the data, providing that this data was not used in both training and validation phases. Divers head actions occurred with different duration appeared in each video. Moreover, classification of head gestures has a significant role in supporting affective state and emotions prediction in real time. Classifying the combined features consists of two or more episodically repeated actions,
and they labeled as onset, offset, and apex periods. The onset marks the beginning of head movement and the offset means there is no head action. While, apex means the peak of the head action and there are no more changes.

The proposed model achieved efficient performance and best results on online and offline data tested in real time. The prediction accuracy for this model can be seen in the Table 1 below. Table 1 depicts the true positive rate (TP%) for the correctly classified action categorized by action label. While the false positive rate (FP%) is for the false classification (or misclassified) actions. Misclassifications take place when exceeding the threshold of the systems (i.e., detected the falsely classified classes). The results show that the highest classification rate was (99%) for nod motion, while the lowest classification rate was (83%) for turn motion. These results refer to the solid performance of the model and high generality of the model towards unseen data. Another method to represents the model performance and its accuracy can also be shown by using precision-recall curve (PRC) as in figures 2 and 3. Figure 2 was to gather the predictions of each class and calculate the area under a Precision-Recall Curve. While the average performance in figure 3 was calculated using the micro-average metric to ensure that the weighted contributions of each class performance is included in the calculations rather than considering all the classes contributions in equal shares (as in macro-average metric).

Table 1 Accuracy results for the model, TP is correctly classified, and FP is incorrectly classified

<table>
<thead>
<tr>
<th>Action</th>
<th>TP%</th>
<th>FP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>94</td>
<td>0.05</td>
</tr>
<tr>
<td>Nod</td>
<td>99</td>
<td>0.007</td>
</tr>
<tr>
<td>Turn</td>
<td>83</td>
<td>0.17</td>
</tr>
<tr>
<td>Shake</td>
<td>87</td>
<td>0.13</td>
</tr>
<tr>
<td>Tilt</td>
<td>93</td>
<td>0.07</td>
</tr>
<tr>
<td>Tilting</td>
<td>96</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Figure 2 Precision-Recall Curves for each class
In conclusion, these results show a significant performance was achieved by HMM in real time through using 3-head rotation angels (yaw, pitch, and roll) with only a webcam. The limitation was the small size of the obtained dataset, considering the model efficiency depends heavily on the quantity of the trained data, hence higher efficiencies require more data.

**Figure 3** Precision-Recall Curves for combined and basic models (ensembled)

**Conclusions**

This study presented the role of head in human-computer interaction and its applications. A new model based on HMM was developed to extract and classify head gestures. The head movements, in this paper, divided into two groups of head movements. First group include single or basic head movements (up, dawn, turn, left, right, tilt, and rest). While, the second group comprise the combined head movements such as (nod, shake, tilting). Collectively, the two of these groups have the same important role in expressing specific meaning during human interact with computer. The proposed system used the OpenFace tracker to obtain head rotation and translation. This tracker is an open source and freely available for scientific research. The results of the tracker are 3DoF and three head rotation axes (X, Y, and Z), which represented head features in sequences of frames to assemble a set of cues to serve as input for the HMM classifier. The best class prediction of gestures was done by choosing the maximum likelihood probability produced by the classifier. The model was trained and tested on the NACD dataset which is an available and public dataset. It categorized in two folders of videos, one for students with Asperger Syndrome and the other folder contains videos for typical students. The data which used in this paper focused on videos of students with typical development. The 5% of data set aside as unseen data, to test the model efficiency and generality toward unseen data. The remaining data sliced into 80% for training purposes and 20% for model validation. The model generality tests
happened after the model training and validation process ending, the generality process carried out without need to repeat the training process. The model achieved a significant result comparing with the results of other models. The model accuracy reaches 99% and 88% for combined and single cues respectively. As well as, the model accuracy was (94%, 99%, 83%, 87%, 93%, 96%) for all captured head gestures (rest, nod, turn, shake, tilt and tilting) respectively. The model can be used in online interactions and offline through dynamic videos. The model doesn’t require expensive tools except a webcam, and it works under varying illumination, without any limitations and occlusion, and in an uncontrolled environment. In addition, the head pose can be in plane-in and plane-out. Some limitation faced this work represented in the small size of the data. Though, it did not influence the model performance, since the model manage to produce sufficient number of cues to successfully derive the proposed model.

For the future work, we need to increase the size of the sample data. The goal of this model will help to use the output of this model as a seed to predict emotions in dynamic video and online video streams.

References


