Multimodal Technologies and Interaction

MDPI

Article

Three-Dimensional, Kinematic, Human Behavioral Pattern-Based Features for Multimodal Emotion Recognition

Amol Patwardhan

AssetMark Inc., Concord, CA 94520, USA; amolpatty@gmail.com

Received: 17 August 2017; Accepted: 8 September 2017; Published: 11 September 2017

Abstract: This paper presents a multimodal emotion recognition method that uses a feature-level combination of three-dimensional (3D) geometric features (coordinates, distance and angle of joints), kinematic features such as velocity and displacement of joints, and features extracted from daily behavioral patterns such as frequency of head nod, hand wave, and body gestures that represent specific emotions. Head, face, hand, body, and speech data were captured from 15 participants using an infrared sensor (Microsoft Kinect). The 3D geometric and kinematic features were developed using raw feature data from the visual channel. Human emotional behavior-based features were developed using inter-annotator agreement and commonly observed expressions, movements and postures associated to specific emotions. The features from each modality and the behavioral pattern-based features (head shake, arm retraction, body forward movement depicting anger) were combined to train the multimodal classifier for the emotion recognition system. The classifier was trained using 10-fold cross validation and support vector machine (SVM) to predict six basic emotions. The results showed improvement in emotion recognition accuracy (The precision increased by 3.28% and the recall rate by 3.17%) when the 3D geometric, kinematic, and human behavioral pattern-based features were combined for multimodal emotion recognition using supervised classification.

Keywords: multimodal emotion recognition; depth sensing; infrared sensor; affective computing; three-dimensional features; geometric features; kinematic features; feature-level fusion

1. Introduction

Emotion responsiveness in automated systems, computers, and assistive robotics greatly improves the quality of interaction with humans [1]. For these interactions to be successful it is important that highly accurate emotion recognition systems exist. Emotions can be detected using the data from audiovisual channels also known as modalities. In the past decade, research interest has shifted towards multimodal emotion recognition [2–4] as opposed to unimodal or bimodal emotion recognition. Researchers [5–10] have shown that the integration of various modalities for affect recognition provides better accuracy over individual modalities. In addition to detecting emotions from the facial features, studies [4,11] have identified that the features from hand modality and the body modality, also contribute significantly during the emotion recognition process. Hence, this research adopts the multimodal emotion recognition approach and uses three-dimensional data from the visual channel combined with the audio features. Specifically, the implementation in this research uses features extracted from the head, face, hand, body, and speech input channels for the multimodal emotion recognition system.

Firstly, this paper proposes the development of novel features deduced from human behavioral patterns to recognize emotions. The motivation to use behavioral features for emotion estimation was drawn from research done by Wallbott [12], which found that body form, movement, gestures and postures are indicative of specific emotions. A comprehensive survey [13] on emotional bodily expression discussed several studies that used behavioral patterns representing emotions. For example, forward whole body movement shows hot anger and symmetric up and down hand movement shows joy. Thus, daily gestures such as a head nod, head shake, shoulder shrug, raising eyebrows, and rolling one's eyes can be attributed to specific emotions. Very few studies [14–21] have attempted to translate daily behavioral patterns into behavioral rule-based features. For instance, some of the questions that can be asked to estimate anger are: (1) Did the person raise both arms above the shoulder? (2) Did the person frown and clench his teeth? These behavioral patterns can be evaluated to obtain (yes = 1, no = 0) answers to the rules above and can be converted into a series of binary values. Thus, a binary feature vector for each modality (head, face, hand, and body) can be constructed and concatenated to geometric and kinematic features. Therefore, this paper uses the movement of tracked points on face, head, hand, and body to extract behavioral pattern-based features from behavioral rules to represent specific emotions. There have been other studies that have examined rule-based feature dynamics, behavior, and gestures for unimodal emotion recognition [18] and adaptive rule-based facial expression recognition [19]. The studies have demonstrated unimodal affect recognition by using a limited set of gesture-based rules and low-level descriptors extracted from various facial expression profiles. Coulson [20] used computer-generated mannequins from shoulder, hand, and head descriptors and showed that each posture and movement can be attributed to one of the six basic emotions [21]. The study demonstrated that knowledge-based rules for emotion recognition can be developed using annotator agreement. Bone et al. [14] have shown the successful use of rule-based framework for arousal rating. The study focused only on vocal features and proposed using rule-based features as an alternative to supervised learning. On the contrary, instead of using the behavioral pattern-based features as an alternative, this research proposes combining the novel features with the geometric and kinematic features to create a joint feature vector. Furthermore, this research examines whether the multimodal emotion recognition accuracy improved by using the feature-level fusion of the behavioral features with the geometric and kinematic features. Finally, this research verifies whether the proposed feature combination improved the emotion recognition accuracy, on spontaneous emotional displays instead of posed emotions. This study also evaluates the emotion recognition accuracy on external datasets to measure the generalizability of the multimodal emotion recognition system. 3D data sets such as Microsoft Research Cambridge 12 (MSRC-12) [22], UCFKinect [23], and MSR Action 3D [24] dataset contain annotated human activity data. In this research, the annotated actions were mapped to one of the six basic emotions using inter-annotator agreement. The multimodal emotion recognition system implementation was evaluated on the mapped dataset.

This research hypothesizes that the concatenation of three-dimensional geometric, temporal, and kinematic features with behavioral pattern-based features would improve the precision and recall rates for multimodal affect recognition. The behavioral features, geometric features, and kinematic 3D data were captured from head movement, facial expressions, hand gestures, and body posture and evaluated on spontaneous emotional display. Thus, the contributions of this research are:

(1) Development of human behavioral pattern-based features from face, head, hand, and body for multimodal emotion recognition.

(2) Using the concatenation of behavioral features and 3D geometric and kinematic features to augment the multimodal emotion recognition system accuracy using supervised learning technique.

2. Related Work

Researchers have used single modalities such as audio signal [25], facial expressions [26], and body modality [27] for the recognition of emotions. Subsequent emotion recognition studies examined bimodal systems that combined face and body modality [28] or face and audio channels [5,6]. In the last decade emotion recognition research, has evolved from unimodal (face modality only, audio modality only) or bimodal systems (face and body modality, face and voice modality) into multimodal systems (three or more combinations of face, head, shoulder, hand, body, voice modality) [2,9,29–33]. The emotion recognition studies have mostly focused on feature extraction from facial geometry or movement of tracked features on the body. For instance, [6,10,11,17] have used coordinates of tracked facial and shoulder points, distance between the facial pair of points to create a feature vector based on facial geometry. Similarly, [4,28] used the movement of tracked joints on the body to create temporal feature vectors. Even though the movement-based temporal features were used, the above studies only tracked features across few consecutive frames and short bursts (under 1 s). On the other hand, this paper calculates the kinematic features using multiple frames (13 data frames) for the entire duration of the emotion specific gestures, actions, and facial expressions (average: 3 s).

Kleinsmith and Bianchi-Berthouze [13] indicated that several behavioral science and psychology studies have shown that specific emotions can be associated with commonly known human behavior, actions, static poses, body language, body form, hand gestures, and facial expressions. However, very few automatic emotion recognition studies have attempted to translate these daily behavioral patterns into behavioral rules and extracted features from these rules. For instance, some of the questions that can be asked to describe and recognize anger are: (1) Did the person throw a punch? (2) Did the person shake their head? These behavioral patterns can be collectively treated as behavioral rules representing a specific emotion [12]. The rules are nothing but the measurements of the coordinates, angles, speed and displacement of tracked points and can be evaluated to obtain (yes = 1, no = 0) answers in the form of a series of binary values. Thus, a binary feature vector for each modality (head, face, hand, and body) can be constructed and combined with geometric and kinematic features. Therefore, this paper uses the movement of tracked points on the face, head, hand, and body into behavioral pattern-based features or behavioral rules to represent specific emotions. The motivation to use features from behavioral patterns for emotion estimation was drawn from behavioral science research on emotional gesture recognition [18,20,21,34,35] and adaptive rule-based facial expression recognitions [19]. The studies have demonstrated unimodal affect recognition by using limited set of gesture-based rules and rules extracted from various facial expression profiles. Coulson [20] used 176 computer generated mannequins from shoulder, hand, and head descriptors and showed that each posture and movement can be attributed to one of the six basic emotions. Dael et al. [35] developed the Body Action and Posture Coding System (BAP) as a tool for expressing emotions.

A survey by Kleinsmith and Bianchi-Berthouze [13] contains a list of gestures from various behavioral studies that have been attributed to specific emotions. The survey identified the potential to map these emotional behavioral patterns into features and found that very few automatic emotion recognition studies have examined such features. Hence, this paper translates these behavioral patterns and cues into features by modality (face, head, hand, and body), for automatic affect recognition. The affect recognition studies [2,3] have found that many studies have employed supervised learning techniques for emotion recognition. Researchers have successfully used classification methods such as Bayesian classifiers [4], Hidden Markov Models (HMM) [15], and SVM [16,17] for affect recognition. Results in [6,10,11] and the classification techniques discussed in multimodal research surveys [32] showed that the performance improved in affect recognition by using SVM classifiers. Hence this paper uses SVM classification for multimodal emotion recognition. Feature-level fusion (also known as early binding [32]) consists of the extraction of features from various modalities and combining the features into a joint feature vector.

Feature-level fusion produces good results when the emotional display is tightly coupled and the sensory information is highly synchronized. Researchers [11,15,36] have used facial and audio data fusion at the feature level for affect recognition. Hence, this paper uses concatenation of features from face, hand, head, body, audio, and the behavioral rule-based features to form a joint feature vector for multimodal emotion recognition system implementation. In recent years, researchers [37–39] have increasingly used Kinect for emotion recognition studies. Konstantinidis et al. [37] used the sensor for emotions among elders in assisted living; however, the study evaluated the emotions in controlled lab conditions. Researchers [38] have proposed deep learning techniques for emotion recognition; however, the study did not address the absence of features from certain modalities at an instance of time. Zhang et al. [39] used Kinect 3D data for emotion recognition; however, only facial features were used for classifier training. In contrast, this paper makes major advancements from a prior study [40] that only focused on affect intensity estimation. Furthermore, this work made significant improvements to the global behavioral rule-based features, calculations of low-level local thresholds used for evaluation of rules from the raw data and list of actions included in the prior work [29,31,41] by using spontaneous emotional responses instead of enacted recordings.

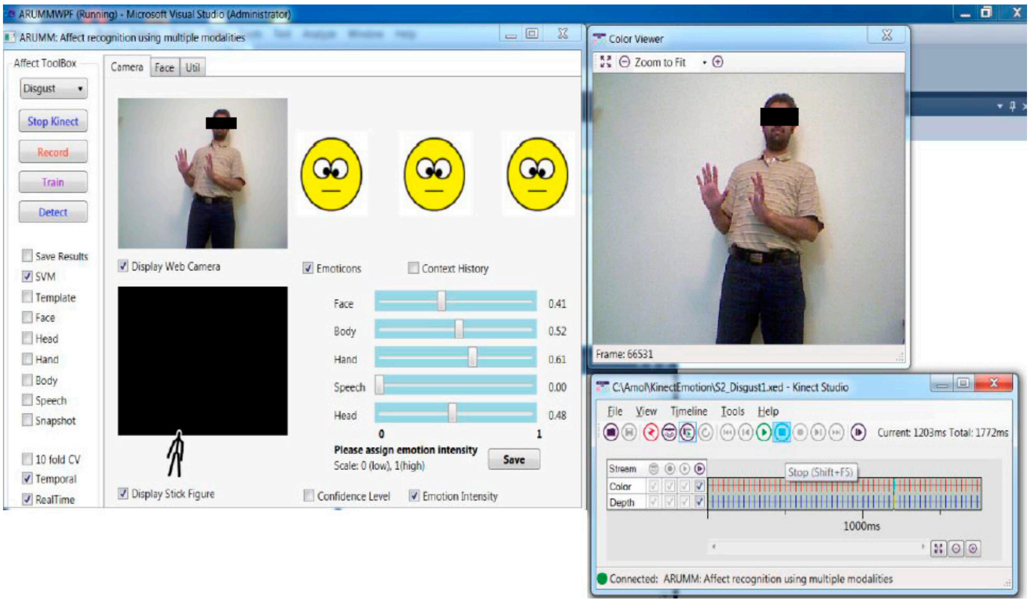
Recently, studies [42,43] have examined rule-based emotion recognition from audiovisual data in the context of fusing outcomes from individual modalities (rule-based decision-level emotion recognition or late binding). However, the rules were not used to extract features in these studies. In contrast, this paper examines extracting behavioral features using rules for low-level raw features. Researchers [44] used a combination of recurrent and convolution neural networks and multiple kernel learning for audio visual sentiment analysis. However, the study extracted only facial features from every 10th frame of the visual channel and used a dataset with conversations done while sitting. In contrast, this paper used actions containing higher degree of movement while standing. A fuzzy rule-based method [45] was developed to detection emotion polarity. However, the study only examined text and movie review data. Researchers [46,47] used three-dimensional convolution network cascaded deep belief networks for emotion recognition from audiovisual streams. However, the study only extracted facial features and did not include features from hand or body movement. Although recent research on multimodal emotion recognition is moving towards deep learning, most of the studies still rely only on facial features from the visual channel. Thus, the overall combination of features from head, face, hand, and body movement in emotional display is largely unexplored. The studies from psychology and human behavioral analysis discussed above have identified the potential emotional cues contained by human behavioral pattern. This study proposed extracting features based in these emotional behavior patterns using three-dimensional, kinematic, rule-based features extracted from the tracked points across face, head, hands, body, and audio channels.

3. System Overview

In this research, a multimodal emotion recognition system was developed (Figure 1) using an infrared sensor from Microsoft called Kinect [37–41]. The 3D data from face, head, hand gestures, and body movement were used for the visual channel. This research used the openEar toolkit by Eyben et al. [25] for capturing audio data. The data from the various modalities was combined using feature-level fusion. The classifiers for the multiple modalities were trained using SVM supervised learning technique. This research used a data mining tool called Weka [48] for training and evaluating the multimodal classification process.

The system was implemented using C# programming language (v4.0) [49], Microsoft. Net framework (v4.0) [50], Kinect software development kit (SDK) (v1.8) [51] and Windows Presentation Foundation (WPF v4.0) [52] for the user interface. The system contains a module for capturing continuous 3D data from various audiovisual channels. An important module of the system was a component called behavioral pattern-based rule evaluation engine (see Section 5.2 for the feature extraction process). This component measured the 3D coordinates, distance between tracked features, speed and direction from the continuous stream of facial expressions, hand, head, and body data. These measurements were then evaluated using behavioral pattern-based rules (also called low-level descriptors, see Section 5.2).

The development of the behavioral patterns used as rules and the mapping between each rule and the specific emotion represented by the rule was done based on inter-annotator agreement and existing studies from behavioral science. Several rules were formulated to check a series of conditions (comparison with threshold) related to location, movement, speed, and frequency of various features. Each of these rules represented a commonly known facial expression or gesture such as rolling the eyes, shoulder shrug, head nod, and raising the arms above shoulder. For instance, the system would evaluate whether the shoulder moved by distance y in the vertical direction and record the outcome of the evaluation as a yes (value = 1) or a no (value = 0) to form a feature vector consisting of binary feature points. This behavioral rule-based feature was then concatenated with the features extracted from the audio–video channels to form a joint feature vector. After this step, the system used the feature for a supervised learning-based classification process implemented using an open-source machine learning library called EmguCV. The system user interface contains a dropdown field for annotating the facial expressions, gestures, and postures to a specific emotion class. The system also provided replay capability for the annotators so that they could review a portion of the video segment. The annotated data were stored in a format (arff file extension) that was compatible to the WEKA data mining tool [48].


ARUMMWPF (Running) - Microsoft Visual Studio (Administrator)

ARUMM: Affect recognition using multiple modalities

Affect ToolBox: Disgust

[redacted]

□ Save Results
☑ SVM
□ Template
□ Face
□ Head
□ Hand
□ Body
□ Speech
□ Snapshot

□ 10 fold CV
☑ Temporal
☑ RealTime

☑ Display Web Camera
☑ Emoticons
□ Context History

Face	0.41
Body	0.52
Hand	0.61
Speech	0.00
Head	0.48

Please assign emotion intensity
Scale: 0 (low), 1(high)

☑ Display Stick Figure
□ Confidence Level
☑ Emotion Intensity

Color Viewer

[redacted]

Frame: 66531

C:\Amol\KinectEmotion\S2_Disgust1.xed - Kinect Studio

Current: 1203ms Total: 1772ms

Stream				
Color	✓	✓	✓	☑
Depth	✓	✓	✓	☑

Stop (Shift+F5)

Connected: ARUMM: Affect recognition using multiple modalities

Figure 1. Implementation of multimodal affect recognition system.

4. Experimental Setup

4.1. Participant Details

The six basic emotions: anger, surprise, disgust, sadness, happiness, and fear were used as the candidate emotions for the recognition process. Each of the six emotions was portrayed by 15 different individuals. The age of the participants was between 25 and 45 years. 5 participants were female and 10 participants were male. Five participants were American and 10 were Asian. All the experiments were conducted in controlled lighting conditions and fully frontal position. The distance between the sensor and the participant was 1.5 to 4 m.

4.2. Data Collection

Fifteen volunteers were used for spontaneous emotional display in the study. These participants displayed naturalistic emotions while standing in front of the Kinect sensor [37–41] at 1.5 to 4 m (Figure 2). The audiovisual color and depth data were used to capture spontaneously expressed emotions from each of the six basic emotion categories. To aid in the emotional display, the subjects were given a list of actions and dialogs as a cue. This list of actions was prepared based on ideas from existing surveys [2,3,13] and inter-annotator agreement. The principal researcher discussed the emotion evoking topic with the participants for 5 min. The facial expressions, voice, body language, hand gestures of the participants, and how they reacted during the dialogue was captured using the sensor. The participants were given freedom to spontaneously change topic, or share stories and experiences that evoked various basic emotions during the dialogs. Thus, the data captured spontaneous and natural emotional display even though they were guided by the scripts, primarily because the participants had the freedom to improvise. The open-ended discussions on topics of the participant's interest allowed the individuals to express their emotions more naturally, without getting self-conscious about the sensor and the recording process. The participants were shown a video related to the following topics and asked about their opinion to evoke emotions corresponding to the expected emotions shown in parenthesis.

- Discussion of the presidential election (Anger, Disgust, Sadness, Happiness)
- Discussion of the NBA finals (Anger, Disgust, Sadness, Happiness)
- Discussion of the NFL Super Bowl (Anger, Disgust, Sadness, Happiness)
- Discussion of *Star Wars* movie (Anger, Disgust, Sadness, Happiness)
- Reaction to viral cat videos (Happiness)
- Reaction to funny viral videos (Happiness)
- Reaction to disgusting viral videos (Disgust)
- Reaction to viral music videos (Disgust, Happiness, Anger)
- Reaction to videos on violence (Disgust, sad, Anger)
- Reaction to sad news (Sadness, Anger)
- Reaction to shocking news (Sadness, Surprise, Anger)
- Reaction to seeing an insect (Fear, Surprise)

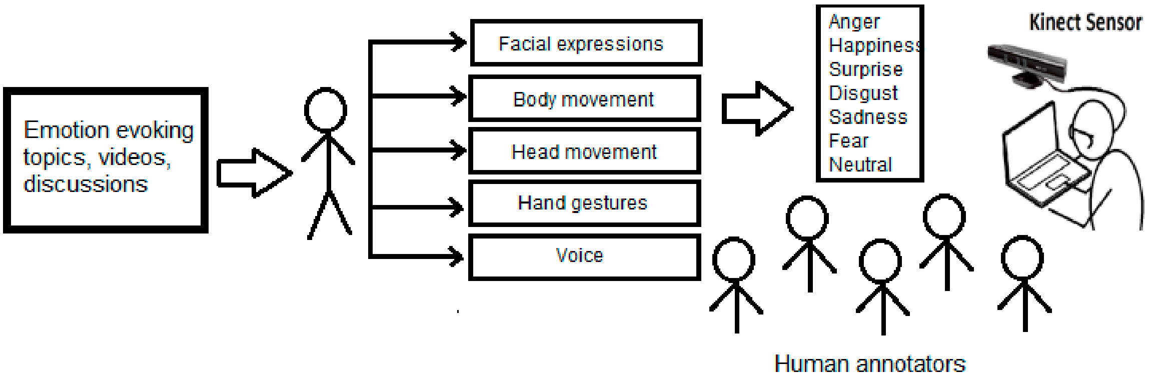


Figure 2. Spontaneous emotion data collection and annotation process using Kinect sensor.

The participant's behavior, facial expressions, body movement, hand gestures, and voice were recorded during the discussion.

4.3. Annotation Process

Five expert annotators from the behavioral science field were asked to label each emotional action, expressed by the participant, with one of the six basic emotions. The panel of five annotators was shown the recording and asked to annotate the emotion for each video segment. The annotators only labeled the high-level emotional actions, speech, and gestures in the video segment. The underlying measurements of micro-expressions (low-level descriptors), the 3D and kinematic feature extraction, and the behavioral rules-based feature calculations were the responsibility of the automated multimodal system's data capturing module. The emotion recognition system calculated the frequency of hand movement, facial expressions, body movement, direction of head gesture, and direction of hand movement at the rate of 13 calculated data frames per second. The sensor frame rate was 30 frames per second. The average action, expression or gestures lasted for 3 s (based on readings from the training set). The resulting 90 frames were used to calculate 13 data frames that contained kinematic feature information such as speed, displacement, direction, and frequency of actions, expressions, and gestures. Thus, the annotators only had to annotate the video segment with an emotion label, while the feature extraction and association with the emotion label was handled by the emotion recognition system's data gathering module. The quality of annotation process was measured using Fleiss kappa value [53] for inter-annotator agreement. To test the effectiveness of the approach on inter-corpus test data evaluation was done using MSRC-12 dataset, UCFKinect dataset, and MSR Action 3D dataset. These datasets are not directly annotated with one of the six basic emotions. The format of the features is also different from the feature definitions of this study. To overcome this limitation before testing the proposed method against the external datasets, each activity in the dataset was mapped to an emotion class as a preprocessing step. Five experts from the field of behavioral science acted as annotators to label each activity with an emotion class. The mapping of emotions and various facial expressions, gestures, postures, and behavioral patterns is described in detail in Section 5.

5. Methods

5.1. Three-Dimensional and Kinematic Feature Extraction

Facial features were extracted by the face recognition application programming interface (API) available in the Kinect SDK [51]. Sixty non-rigid features out of the 121 features were used (Figure 3). The intuition behind the initial selection of features was that only the features from the expressive part of the face were considered.

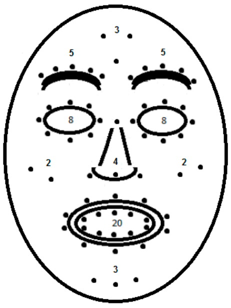


Figure 3. Tracked features using face tracking API. The 60 features were chosen from the expressive part of the face.

The 60 non-rigid features included x, y, and z coordinates of the eyes, nose, lips, eyelids, chin, cheeks, and forehead. In addition to the coordinates, the distance between each pair of features and the angle made by each pair with the horizontal axis was calculated. The movement of each of these features was captured for a window of 3 s, which resulted in 39 calculated data frames. The velocity and displacement of each feature were calculated and used as kinematic features. The features were stored in a format called arff used by the Weka data mining tool [48].

For tracking the head position and movement, 12 features along the border of the skull, out of the 121 extracted features, were chosen (Figure 4). The reason for choosing these features was that they would define the shape of the head as well as capture movements such as a pitch, yaw, roll, nod, shake, lateral, backward, or forward motion of the head. Additionally, the distance between each pair of features, the angle with the horizontal, and the movement of features across 90 consecutive frames were calculated.

In the case of hand gestures, palms, wrist, elbow and shoulder joints of both hands were tracked resulting in eight features. These features were selected because they capture the vigorous movement of the arms along all three axes. The distance between each pair of joints, the angle with the horizontal, and the velocity and displacement of each joint across 90 frames were calculated to create a feature vector.

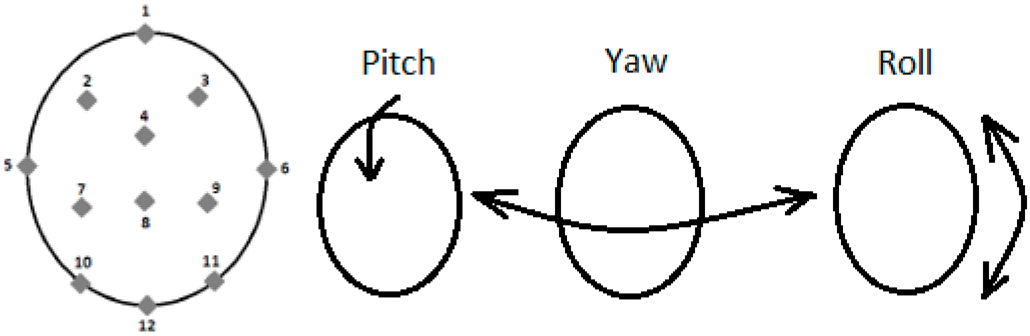


Figure 4. Tracked features for head position and movement.

For the body posture, the center of the spine, hip, and left and right hip joints were tracked in addition to the joints of the hand (Figure 5). The feature vector was constructed using the distance between pairs of joints, the angle with the horizontal, and the velocity and displacement of each joint across 90 frames. For the audio modality, the openEar toolkit [25] was used to extract the features and the pre-built SVM-based classifiers were not used for emotion recognition.

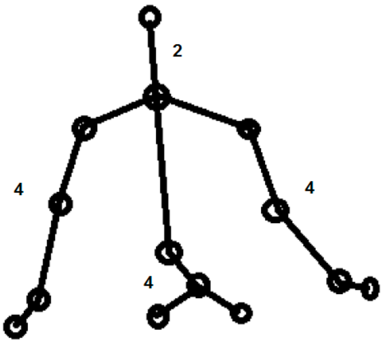


Figure 5. Tracked features for hand and body position and movement.

5.2. Human Behavioral Features

This research developed low-level descriptors to capture daily behavioral patterns such as a head nod, head shake, shoulder shrug, or raising the arms in the air. The research developed emotion recognition behavioral pattern-based features from raw 3D geometric and kinematic data. To develop these behavioral features, five experts were asked to annotate static poses (Figure 6) and video segments containing gestures, actions, and expressions to a specific emotion class. For each emotion to action mapping and development of behavioral pattern-based rule, the Fleiss Kappa value [53] was calculated as a measurement of inter-annotator agreement. The quality of agreement was measured using the interpretation scale provided by Landis & Koch [54].

Both the motionless body form as well as movement-based spontaneous emotional displays were used for creating behavioral pattern-based features. The annotators used the video segments containing emotional actions to create behavioral features by evaluating facial expressions and head, hand, and body movements. Each movement and pose was mapped to a set of behavioral features defined using the coordinates, angle, and distance of the features from various modalities. Similarly, each action, facial expression, and gesture was mapped to a set of behavioral features by calculating the movement along a certain axis, the frequency of movement, the velocity, and the displacement of the features discussed earlier. The multimodal emotion recognition system enabled the annotators to pause, replay, rewind, fast-forward, and play the video segments for detailed evaluation and labeling. Hence, the annotators could pause the video and evaluate the pose or could replay the emotional actions for a few seconds before associating the behavioral pattern with a specific emotion.

The experts could accurately annotate and examine the static poses using the pause and replay features. This process simplified the annotation process because the annotators did not have to analyze each individual micro-expression (low-level descriptor) and calculation of each feature associated with the emotion.

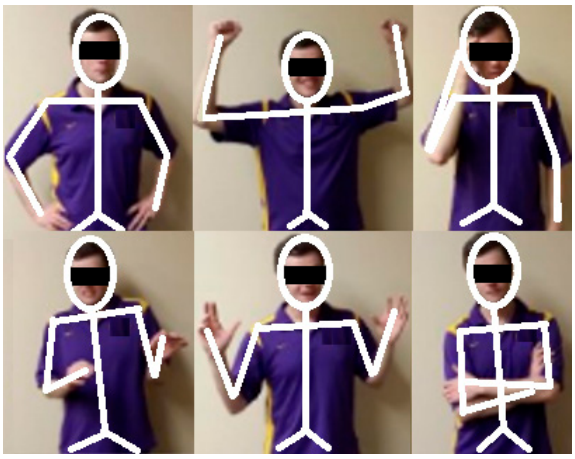


Figure 6. Spontaneous emotions and corresponding body form.

The feature extraction, measurements (coordinates, angles, distance, and speed), rule evaluation, and threshold calculations were calculated as described in Section 5.1 and Table 1 during the video segment playback and the annotators only had to apply the emotion label to the action that was being examined. Once the annotation was complete, the threshold values for each descriptor were calculated using the average of each measured angle, distance, velocity, and frequency. The threshold distances and T angles were calculated for each of the six basic emotions. Measurements were normalized using the max width and height distances of the Kinect sensor's field of vision. Different humans have different heights, sizes, and shapes. Taking the average to calculate thresholds and using normalized measurements (using screen size) accounted for the variations in shape and size. The feature reduction was performed using Information Gain and Ranker search method in the Weka tool. Table 1 shows a list of low-level measurements for computing behavioral pattern-based features. Table 2 contains the high-level expressions, poses, gestures, and actions associated with specific emotions by the panel of experts. Please refer to Appendix A for a complete list of low-level descriptors, with the corresponding threshold values per each modality and emotion combination.

For each action, facial expression, and gesture, the rule evaluation component used the 69 measurements shown in Table 1 and compared these values with the thresholds from Appendix A. This comparison was performed for each emotional behavioral pattern expressed by the participants. If the measurement was above a threshold or satisfied a condition, the result of the evaluation was a yes (value = 1) or a no (value = 0). The outcome of the series of such comparisons was a binary feature vector of size 69. The computed behavioral feature vector was then fed into the multimodal system for feature-level fusion with the geometric, 3D, and kinematic features from facial expression, head, hand, body, and audio channel. Thus, each action in the video segment resulted in a joint feature vector that was the concatenation of 3D coordinates, distance, velocity, and binary feature points. The list of high-level actions, facial expressions, gestures [12,13,20,34,55], and the corresponding emotion based on manual inter-annotator agreement is as follows:

Table 1. Low-level descriptors for behavioral pattern-based feature extraction.

Angle-, Frequency-, and Distance-Based List of Features	
Angle of left elbow = T	Shaking head sideways (yaw) frequency
Angle of right elbow = T	Head bob (roll) frequency
Angle (left shoulder and arm) = T	Body forward movement frequency
Angle (right shoulder and arm) = T	Body backward movement frequency
Angle of spine = T (x, y, z axis)	Sideways movement frequency
Angle of head = T (x, y, z axis)	Wrist movement (x, y, z axis) frequency
Y (wrist)–Y (elbow)	Elbow movement (x, y, z axis) frequency
Y (elbow)–Y (shoulder)	Hip movement (x, y, z axis) frequency
X (wrist)–X (elbow)	Forehead movement (x, y, z axis) frequency
X (elbow)–X (shoulder)	Spine tilting (x, y, z axis) frequency
Z (wrist)—Z (elbow)	X (shoulder) movement frequency
Z coordinate (elbow)–Z (shoulder)	Waving hand frequency
X coordinate (wrist)–X (shoulder)	Head nod (pitch) frequency
X (elbow)–X (spine)	Eyebrow movement (x, y, z axis) frequency
Y (elbow)–Y (spine)	Upper lip movement (x, y, z axis) frequency
Distance between right cheek and lip corner	Lower lip movement (x, y, z axis) frequency
Distance between nose tip and upper lip	Cheek movement (x, y, z axis) frequency
Distance between corners of lip	Lip corner movement (x, y, z axis) frequency
Distance between upper and lower eyelid	Distance between eyebrow and eyes
Distance between left cheek and lip corner	Distance between upper and lower lip
Distance between nose tip and forehead	Distance between left wrist and head top
Distance between upper lip and forehead	Distance between right wrist and head top
Distance between left and right eyebrow	

Table 2. High-level poses and actions associated with spontaneous emotions.

Emotion	Action	Fleiss Kappa	Agreement Level
Anger	Throwing an object.	1	Perfect
	Punching.	1	Perfect
	Holding head in frustration.	0.8	Substantial
	Folding hands.	0.6	Moderate
	Holding hands on waist.	0.6	Moderate
	Moving forward threateningly.	0.8	Substantial
	Throwing a fit.	0.8	Substantial
	Raising arms in rage.	0.6	Moderate
	Moving around aggressively.	0.6	Moderate
	Shouting in rage.	1	Perfect
	Pointing a finger at someone.	0.6	Moderate
	Threatening someone.	0.8	Substantial
	Scowling.	1	Perfect
Happiness	Jumping for joy.	1	Perfect
	Fist pumping in joy.	1	Perfect
	Raising arms in air in happiness.	1	Perfect
	Laughing.	1	Perfect
	Smiling.	1	Perfect
Disgust	Holding nose.	1	Perfect
	Looking down, expressing disgust.	1	Perfect
	Moving back in disgust.	0.6	Moderate
Sadness	Looking down, leaning against wall.	0.8	Substantial
	Looking down with hands on waist.	0.6	Moderate
	Looking down with hands folded.	0.8	Substantial
	Crying.	1	Perfect
	Toothache.	0.8	Substantial
	Head hurting.	0.8	Substantial

Table 2. *Cont.*

Emotion	Action	Fleiss Kappa	Agreement Level
Surprise	Raising arms in surprise.	0.8	Substantial
	Moving back in surprise.	0.8	Substantial
	Walking forward and getting startled.	1	Perfect
	Holding arms near chest in surprise.	0.8	Substantial
	Covering mouth with hands.	0.8	Substantial
Fear	Moving backwards, trying to evade.	1	Perfect
	Moving sideways.	0.8	Substantial
	Looking up and running away.	0.8	Substantial
	Getting rid of an insect on shirt.	0.8	Substantial
Neutral	Standing straight with no facial expression.	1	Perfect

For testing the generalizability of the proposed method, the human activity datasets were mapped to emotion labels using inter-annotator agreement between five experts. The mapping between each activity from the external datasets and an emotion class along with the inter-annotator ratings are shown in Table 3 (MSRC-12), Table 4 (UCFKinect), and Table 5 (MSR Action) below:

Table 3. Mapping for MSRC-12.

Action	Emotion	Fleiss Kappa	Agreement Level
Crouch or hide	Fear	1	Perfect
Shoot with a pistol	Angry	0.8	Substantial
Throw an object	Angry	0.8	Substantial
Change weapon	Angry	0.2	Slight
Kick to attack an enemy	Angry	1	Perfect
Put on night vision goggles	Neutral	0.4	Fair
Had enough gesture	Angry	0.6	Moderate
Music-based gestures	Neutral	0.4	Fair

Table 4. Mapping for UCFKinect.

Action	Emotion	Fleiss Kappa	Agreement Level
Balance, climb ladder, climb up	Neutral	1	Perfect
Duck	Fear	0.8	Substantial
Hop	Surprise	0.6	Moderate
Kick	Anger	0.8	Substantial
Leap	Surprise	0.6	Moderate
Punch	Anger	1	Perfect
Run	Fear	0.8	Substantial
Step back	Fear	0.8	Substantial
Step back	Disgust	0.6	Moderate
Step front	Anger	0.6	Moderate
Step left	Disgust	0.8	Substantial
Step right	Disgust	0.8	Substantial
Turn left, Turn right, Vault	Neutral	1	Perfect

Table 5. Mapping for MSRAction.

Action	Emotion	Fleiss Kappa	Agreement Level
Sit down, stand up	Neutral	1	Perfect
Hand clapping	Happiness	1	Perfect
Hand waving	Happiness	0.6	Moderate
Fist pump gesture	Happiness	1	Perfect
Boxing	Anger	1	Perfect
Toss a paper in frustration	Anger	0.8	Substantial

The MSRC-12, UCFKinect, and MSRAction 3D datasets are intended for human activities and not specifically for emotion representation. However, this limitation was overcome by mapping the actions to specific emotions using inter-annotator agreement between five experts. Based on the results of the mapping, a version of each dataset was created so that it could be tested with the emotion recognition implementation in this paper. If the dataset did not contain an action representing a specific emotion, then the actions from the general list of emotional behavioral patterns were included to complete the dataset for all the six basic emotions.

5.3. Classification Process

Researchers [12,13,20,34,55] have shown that static head, hand, and body positions can be associated with specific emotions. Hence, this research associates daily human behavioral patterns with six basic emotions using human annotator agreement. Next, the high-level behavioral patterns were evaluated as micro-expressions that involved comparison between tracked point coordinates, distance, and velocity measurements and the threshold values. The behavioral pattern-based features were developed based on the position of the hands and body posture as well as kinematic movement-based features such as speed, displacement, and frequency of the tracked points from the head, facial expressions, hand, and body. This resulted in a behavioral pattern-based binary feature vector (69 data points). This behavioral pattern-based feature vector computation was then implemented in the multimodal emotion recognition system and was concatenated with geometric, 3D, kinematic features using feature-level fusion. The resulting joint feature vector was used to train the multimodal emotion recognition classifier. The classification process was performed using SVM, which is a popular supervised learning technique. The data were split into 80% training and 20% test data. The multimodal classifier was trained using the data obtained from spontaneous emotions and evaluated using 10-fold-cross validation and the radial basis function as the non-linear kernel function. The optimal parameters were computed using the GridSearch method and the values (Table 6) for each modality; the specific classifiers are as follows:

Table 6. SVM parameters for baseline.

Modality	C	Gamma
Head	1	1/12
Face	1	1/60
Hand	1	1/8
Body	1	1/14
Multimodal	1	0.09
Audio	1	0.009

6. Results and Discussion

Fifteen participants displayed spontaneous emotions while discussing topics of their choice. The extracted feature vectors were annotated with the emotion class using the action–emotion mapping based on daily behavioral patterns by a panel of five experts. First, the baseline multimodal emotion recognition results were obtained using the affect recognition technique described in [17]. The study used 2D geometric features (coordinates and distance) and kinematic features (speed and displacement). Next, the multimodal emotion recognition results using the proposed method were measured. These included features extracted from face, head, hand, body, and audio channels. The feature vector was the concatenation of 3D geometric features (position, distance, and angle across three dimensions) and kinematic features (speed, displacement across 13 consecutive data frames) and the behavioral rule-based features (daily human behavioral patterns). The classification was done using SVM classification. The precision results of the SVM-based multimodal emotion recognition using the state-of-the-art 2D geometric and kinematic features, the proposed method, and an evaluation of external datasets are shown in Table 7.

Table 7. Precision comparison for multimodal emotion recognition.

2D Geometric Features	Proposed Method	MSRC-12	UCF Kinect	MSR Action	Label	Emotion
80.5	86.2	81.7	77.4	84.6	0	Anger
82.8	87.2	77.3	79.8	84.6	1	Happiness
79.5	82	78.9	73.9	75.1	2	Surprise
81.1	88.3	83.1	73.5	80.9	3	Disgust
83.2	85.2	75.3	72.2	76.7	4	Fear
85	84.1	78.4	75.3	80.4	5	Sadness
94.7	96.8	94.6	88.6	88.9	6	Neutral

Table 7 shows that the emotion recognition precision rate increased by an average of 3.28% when using the proposed method for all emotions as compared to the precision rate obtained using the 2D geometric and kinematic features. The proposed method was evaluated on three different datasets (MSRC-12, UCF-Kinect, and MSR Action). The precision rate decreased by an average of 5.78% for the MSRC-12 dataset, 9.78% for UCF-Kinect, and 5.51% for the MSR-Action dataset, respectively.

Table 8 shows that the emotion recognition recall rate increased by an average of 3.17% when using the proposed method for all emotions as compared to the recall rate obtained using the 2D geometric and kinematic features. The proposed method was evaluated on three different datasets (MSRC-12, UCF-Kinect, and MSR Action). The recall rate decreased by an average of 6% for the MSRC-12 dataset, 10% for UCF-Kinect, and 5.71% for the MSR Action dataset, respectively.

Table 8. Recall comparison for multimodal emotion recognition.

2D Geometric Features	Proposed Method	MSRC-12	UCF Kinect	MSR Action	Label	Emotion
80.1	86	78.6	78.4	83.4	0	Anger
80.4	86.3	80.6	72.5	79.8	1	Happiness
80.5	85.2	81.2	79.7	83.4	2	Surprise
79.7	82.4	79.7	79.5	84.8	3	Disgust
84.8	87	80.1	78.3	80.1	4	Fear
88.5	90.9	83.8	75.5	79.9	5	Sadness
92.9	91.3	83.1	73.9	77.7	6	Neutral

Table 9 shows that the emotion recognition F-score increased by an average of 3.18% using the proposed method for all emotions as compared to the F-score obtained using the 2D geometric and kinematic features. The proposed method was evaluated on three different datasets (MSRC-12, UCF-Kinect, and MSR Action). The F-score decreased by an average 5.94% for the MSRC-12 dataset, 10.14% for UCF-Kinect, and 5.68% for the MSR Action dataset, respectively.

Table 9. F-Score comparison for multimodal emotion recognition.

2D Geometric Features	Proposed Method	MSRC-12	UCF Kinect	MSR Action	Label	Emotion
80.3	86.1	80.1	77.9	84	0	Anger
81.6	86.7	78.9	76	82.1	1	Happiness
80	83.6	80	76.7	79	2	Surprise
80.4	85.2	81.4	76.4	82.8	3	Disgust
84	86.1	77.6	75.1	78.4	4	Fear
86.7	87.4	81	75.4	80.1	5	Sadness
93.8	94	88.5	80.6	82.9	6	Neutral

Figure 7 shows a consistent F-score value greater than 75% for all the emotions (including neutral) when the proposed method was used on external datasets. The results also indicated that the concatenation of 3D, kinematic, and human behavioral pattern-based features improved the multimodal emotion recognition accuracy compared to techniques that use 2D geometric and kinematic features alone. The F-score was lowest for the UCF-Kinect dataset, which mostly contained complex facial expressions and fewer emotional displays using overall body movement. Even though the accuracy dropped when the proposed method was evaluated on external datasets, the consistent F-score value (>75%) indicates that the proposed method was robust and generalizable to inter-corpus test data.

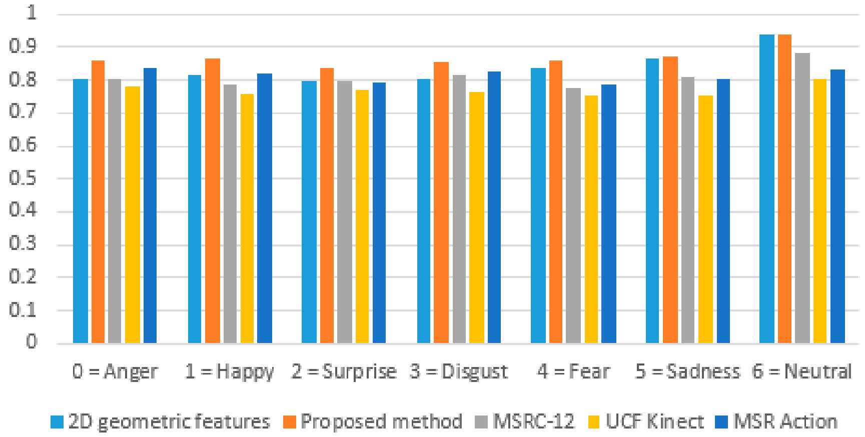


Figure 7. Comparison of F-score for multimodal emotion recognition.

The values in Table 10 along the diagonal represent the recall rate for each emotion. The overall recall rate was 83.84 (average of recall rate across all emotions) for multimodal emotion recognition using 2D geometric and kinematic features. Anger (80.1%) was most often misclassified as surprise (9.6%). Happiness (80.4%) was most often misclassified as anger (9.7%). Surprise (80.5%) was most often misclassified as happiness (10.8%). Disgust (79.7%) was misclassified as fear (8.9%). Fear (84.8%) was sometimes misclassified as disgust (6.2%). Sadness (88.5%) was most often misclassified as neutral (2.2%).

Table 10. Confusion matrix using geometric and kinematic features.

0	1	2	3	4	5	6	Label	Emotion
80.1	4.9	9.6	4.8	0.3	0.2	0	0	Anger
9.7	**80.4**	6	3.6	0.1	0.1	0.1	1	Happiness
7.4	10.8	**80.5**	0.4	0.2	0.5	0.1	2	Surprise
1	0.6	0.6	**79.7**	8.9	8.4	0.9	3	Disgust
0.8	0.3	3.1	6.2	**84.8**	3.3	1.5	4	Fear
0.5	0.5	0.1	3.2	5	**88.5**	2.2	5	Sadness
0.4	0.1	0.8	0.8	2.6	2.4	**92.9**	6	Neutral

The values in Table 11 along the diagonal represent the recall rate for each emotion. The overall recall rate was 87% (average of recall rate across all emotions) for multimodal emotion recognition using the proposed method. Anger (86%) was most often misclassified as surprise (8.9%). Happiness (86.3%) was most often misclassified as anger (6.8%). Surprise (85.2%) was most often misclassified as happiness (9.3%). Disgust (82.4%) was most often misclassified as sadness (8.4%). Fear (87%) was sometimes misclassified as sadness (3.2%). Sadness (90.9%) was most often misclassified as disgust (4.5%). Thus, the recall rates for each emotion improved when the joint feature vector of 3D geometric, kinematic, and behavioral pattern-based features was used. Misclassification was observed among anger, surprise, and happiness. This can be explained by the fact that the actions and gestures for anger, surprise, and happiness involve a higher degree of movement, jerks, and intensity. On the other hand, the misclassifications among disgust, fear, sadness, and neutral were observed because these emotions are displayed with low-intensity actions, subtle gestures and facial expressions, and less movement of the body.

Table 11. Confusion matrix using the proposed method.

0	1	2	3	4	5	6	Label	Emotion
86	3.1	8.9	1.9	0.1	0	0	0	Anger
6.8	**86.3**	6.2	0.6	0	0	0.1	1	Happiness
4.6	9.3	**85.2**	0.4	0.1	0.4	0.1	2	Surprise
0.9	0.1	0.6	**82.4**	7.2	8.4	0.4	3	Disgust
0.6	0.3	2.7	5	**87**	3.2	1.3	4	Fear
0.4	0.1	0	2.8	4.5	**90.9**	1.3	5	Sadness
0.2	0	0.4	0.6	3.5	4	**91.3**	6	Neutral

Novel behavioral pattern-based features were extracted from daily facial expressions and gesture patterns representing emotions. The annotators labeled high-level, commonly known emotional behaviors with emotion classes that were converted to corresponding low-level descriptor-based feature calculations. The behavioral pattern-based features were concatenated with 3D geometric and kinematic features extracted from the face, hands, head, body, and voice. The proposed method improved the multimodal emotion recognition results when compared to 2D geometric and kinematic features used in the state-of-the-art emotion recognition studies. To test the generalizability, the multimodal emotion recognition system was evaluated on spontaneous emotion recognition data and external datasets such as MSRC-12, MSR Action, and UCF Kinect.

The results showed that the accuracy was not significantly reduced and thus indicated that the proposed behavioral pattern-based features were robust.

7. Threats to Validity

Even though the number of participants was only 15, the research ensured that the participants belonged to a diverse demographic in terms of gender, age group, and culture. Many existing studies, according to the surveys [2,3,9,32], are limited to posed and enacted emotions, whereas this research uses spontaneous emotional display. The participants chose topics they were passionate about and engaged in a natural discussion, expressing their opinions and emotions through a range of actions, expressions, gestures, and dialogue. The gestures, actions, and expressions were mapped to various emotions for the development of behavioral pattern-based features. This mapping was done using an expert panel of five human annotators and by using a list of actions from existing behavioral science research [12,20,34,55]. The diverse group of participants enabled us to evaluate the emotional displays and behavioral pattern-based features irrespective of the culture, gender, and age group. Finally, this study mostly focuses on extracting features from behavioral patterns using visual channel data. In the future, the audio-based emotional patterns should be further explored to see whether the vocal features and rule-based method proposed in [14] can be combined during the feature-level fusion with the features in this paper.

8. Conclusions

This research developed behavioral pattern-based features using (1) Static poses for commonly identified emotional body forms; (2) kinematic data extracted from the facial expressions, hand gestures, and body movement; and (3) 3D data from infrared sensor depth and skeleton frames. The research used daily actions, facial expressions, and gestures from existing behavioral science studies. Existing studies have shown that the common behavioral patterns can be associated with specific emotions. The study used these lists of actions from the field of behavioral science and psychology and translated the patterns into a binary feature vector. The spontaneous emotional display from 15 participants was annotated by a panel of five experts. Each video segment contained natural actions, facial expressions, and gestures with specific behavioral patterns that could be attributed to one of the six basic emotions. The study encoded the participant responses into behavioral features using conditional rules that evaluated measurements such as angle, frequency, speed, displacement, direction, and coordinates.

The concatenation of behavioral pattern-based features, 3D geometric, and kinematic features was used to recognize emotions using SVM. A multimodal system implementation that combined these tightly coupled features was employed to detect emotions using the supervised classification technique. The precision increased by 3.28% and the recall rate by 3.17% when the 3D geometric, kinematic, and human behavioral pattern-based features were concatenated at the feature level. The results indicate that the behavioral pattern-based features can be used in multimodal emotion recognition systems and the joint feature vector improved the accuracy and generalizability across external datasets developed from spontaneous and naturalistic emotions. While this study used novel features and spontaneous emotions, future studies could examine the influence of context on the behavioral pattern-based features. As a future direction, automatic multimodal emotion recognition using 3D geometric, kinematic, and behavioral pattern-based features from multi-view data needs further exploration. The effectiveness of the proposed behavioral pattern-based features in multimodal emotion recognition using deep learning and convolution neural networks will be examined as a subsequent phase of the study.

Conflicts of Interest: The author declare no conflict of interest.

Appendix A

Rule Descriptor	0	1	2	3	4	5	6
Units: Angle measurements are in degrees. Distances are in normalized pixels. Frequency of moving head, joints, facial parts is the number of times the tracked joint or feature point moves along reference *x*, *y*, *z* axis or crosses the axis per video segment.							
Angle of left elbow = T	43	172	31	16	32	124	187
Angle of right elbow = T	45	178	33	11	36	121	185
Angle between left shoulder and arm = T	37	93	122	6	212	254	276
Angle between right shoulder and arm = T	39	87	126	7	214	257	271
Angle of spine = T (*x*, *y*, *z* axis)	92, 2, 89	113, 2, 94	101,1,96	105, 2, 93	108, 2, 97	85, 3, 96	92, 1, 91
Angle of head = T (*x*, *y*, *z* axis)	90, 2, 85	92, 3, 114	91, 2, 106	89, 4, 107	91, 2, 102	93, 1, 86	89, 1, 90
Y (wrist)–Y (elbow)	0.12	0.13	0.09	0.12	0.15	0.13	0.16
Y (elbow)–Y(shoulder)	0.07	0.09	0.08	0.10	0.06	0.08	0.07
X (wrist)–X (elbow)	0.6	0.7	0.6	0.5	0.3	0.2	0.2
X (elbow)–X (shoulder)	0.12	0.14	0.16	0.6	0.8	0.6	0.4
Z (wrist)–Z (elbow)	0.15	0.16	0.11	0.7	0. 5	0.7	0. 5
Z co-ordinate (elbow)–Z (shoulder)	0.14	0.17	0.13	0.6	0.6	0.8	0. 7
X co-ordinate (wrist)–X (shoulder)	0.19	0.18	0.16	0.8	0.9	0.6	0.6
Frequency of waving hand	5	3	2	2	2	2	0
Frequency of head nod (pitch)	6	4	2	1	2	1	0
Frequency of shaking head sideways (yaw)	4	3	2	1	2	1	0
Frequency of head bob (roll)	3	3	2	2	2	1	0
Frequency of body forward movement	4	3	2	1	3	1	0
Frequency of body backward movement	4	3	2	1	2	1	0
Frequency of sideways movement	3	3	2	1	2	2	1
Frequency of wrist movement (*x*, *y*, *z* axis)	6,3,3	4,2,3	3,3,2	1,1,1	2,1,1	1,1,2	0,0,1
Frequency of elbow movement (*x*, *y*, *z* axis)	5,2,2	3,3,2	2,3,1	2,1,2	2,1,1	1,1,1	0,0,1
Frequency of hip movement (*x*, *y*, *z* axis)	3,2,1	3,1,1	1,2,1	1,1,2	2,1,2	1,0,1	0,1,1
Frequency of forehead movement (*x*, *y*, *z* axis)	3,2,3	3,3,3	2,3,2	1,1,2	2,2,1	1,1,1	0,0,0
Frequency of spine tilting (*x*, *y*, *z* axis)	3,1,1	3,1,1	2,2,1	1,2,1	3,1,1	1,1,1	0,0,1
Frequency of X (shoulder) movement	3	3	2	1	2	1	0
Distance between left wrist and head top	0.13	0.15	0.08	0.23	0.08	0.36	0.42
Distance between right wrist and head top	0.14	0.18	0.09	0.24	0.07	0.38	0.44
X (elbow)–X (spine)	0.25	0.12	0.14	0.18	0.15	0.07	0.51
Y (elbow)–Y (spine)	0.18	0.19	0.22	0.24	0.21	0.23	0.22
Distance between eyebrow and eyes	0.003	0.002	0.004	0.001	0.005	0.001	0.004
Distance between upper and lower lip	0.005	0.006	0.005	0.003	0.005	0.002	0.002
Distance between nose tip and upper lip	0.001	0.003	0.003	0.002	0.002	0.003	0.001
Distance between corners of lip	0.02	0.04	0.02	0.02	0.03	0.03	0.03
Distance between upper and lower eyelid	0.012	0.015	0.14	0.14	0.13	0.12	0.15
Distance between right cheek and lip corner	0.07	0.08	0.04	0.05	0.04	0.04	0.04
Distance between left cheek and lip corner	0.04	0.06	0.04	0.04	0.04	0.04	0.04
Distance between upper lip and forehead	0.03	0.06	0.05	0.05	0.03	0.05	0.15
Distance between left and right eyebrow	0.02	0.02	0.02	0.03	0.03	0.03	0.03
Distance between nose tip and forehead	0.03	0.05	0.05	0.05	0.03	0.05	0.15
Frequency of movement of lip corners (*x*, *y*, *z* axis)	3, 3, 2	2, 2, 2	2, 3, 1	1, 2, 1	2, 2, 1	1, 2, 1	0, 0, 0
Frequency of cheek movement (*x*, *y*, *z* axis)	3, 2, 2	2, 3, 1	2, 2, 1	2, 2, 1	2, 1, 1	2, 2, 1	0, 0, 0
Frequency of eyebrow movement (*x*, *y*, *z* axis)	2, 3, 2	2, 3, 3	2, 3, 2	1, 2, 2	3, 2, 1	1, 2, 1	0, 0, 0
Frequency of upper lip movement (*x*, *y*, *z* axis)	2, 3, 2	3, 3, 2	2, 2, 2	1, 3, 1	2, 3, 1	1, 2, 1	0, 0, 0
Frequency of lower lip movement (*x*, *y*, *z* axis)	2, 3, 2	2, 3, 2	2, 3, 3	1, 3, 1	2, 2, 1	1, 2, 1	0, 0, 0

References

1. Picard, R.W.; Picard, R. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
2. Pantic, M.; Rothkrantz, L.J. Toward an Affect-Sensitive Multimodal Human–Computer Interaction. *IEEE Proc.* **2003**, *91*, 1370–1390. [CrossRef]
3. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 39–58. [CrossRef] [PubMed]
4. Castellano, G.; Kessous, L.; Caridakis, G. Multimodal emotion recognition from expressive faces, body gestures and speech. In Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 12–14 September 2007; Volume 247, pp. 375–388.
5. Emerich, S.; Lupu, E.; Apatean, A. Bimodal approach in emotion recognition using speech and facial expressions. In Proceedings of the International Symposium on Signals, Circuits and Systems, Iasi, Romania, 9–10 July 2009; pp. 1–4.
6. Chen, L.; Huang, T.; Miyasato, T.; Nakatsu, R. Multimodal human emotion/expression recognition. In Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 366–371.
7. Scherer, K.; Ellgring, H. Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion* **2007**, *7*, 158–171. [CrossRef] [PubMed]
8. Kapoor, A.; Picard, R.W. Multimodal affect recognition in learning environments. In Proceedings of the 13th Annual ACM International Conference on Multimedia, Singapore, 6–12 November 2005; pp. 677–682.
9. D'Mello, S.; Graesser, A. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Model. User Adapt. Interact.* **2010**, *10*, 147–187. [CrossRef]
10. Baenziger, T.; Grandjean, D.; Scherer, K.R. Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion* **2009**, *9*, 691–704. [CrossRef] [PubMed]
11. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–15 October 2004; pp. 205–211.
12. Wallbott, H.G. Bodily expression of emotion. *Eur. J. Soc. Psychol.* **1998**, *28*, 879–896. [CrossRef]
13. Kleinsmith, A.; Bianchi-Berthouze, N. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* **2013**, *4*, 15–33. [CrossRef]
14. Bone, D.; Lee, C.; Narayan, S. Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. *IEEE Trans. Affect. Comput.* **2014**, *5*, 201–213. [CrossRef] [PubMed]
15. Zeng, Z.; Jilin, T.; Pianfetti, B.M.; Huang, T.S. Audio-visual affective expression recognition through multistream fused HMM. *IEEE Trans. Multimedia* **2008**, *10*, 570–577. [CrossRef]
16. Takahashi, K. Remarks on SVM-based emotion recognition from multi-modal bio-potential signals. In Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication, Kurashiki, Japan, 20–22 September 2004; pp. 95–100.
17. Valstar, M.F.; Gunes, H.; Pantic, M. How to distinguish posed from spontaneous smiles using geometric features. In Proceedings of the ACM International Conference on Multimodal Interfaces, Nagoya, Japan, 12–15 November 2007; pp. 38–45.
18. Zhang, L.; Yap, B. Affect Detection from text-based virtual improvisation and emotional gesture recognition. *Adv. Hum. Comput. Interact.* **2012**, *2012*, 461247. [CrossRef]
19. Ioannou, S.; Raouzaiou, A.; Karpouzis, K.; Pertselakis, M.; Tsapatsoulis, N.; Kollias, S. Adaptive rule-based facial expression recognition. *Lect. Notes Artif. Intell.* **2004**, *3025*, 466–475.
20. Coulson, M. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *J. Nonverbal Behav.* **2010**, *28*, 117–139. [CrossRef]
21. Ekman, P. An argument for basic emotions. *Cognit. Emot.* **1992**, *6*, 169–200. [CrossRef]
22. Fothergill, S.; Mentis, H.M.; Kohli, P.; Nowozin, S. Instructing people for training gestural interactive systems CHI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012; pp. 1737–1746.
23. Masood, S.Z.; Ellis, C.; Tappen, M.F.; LaViola, J.J., Jr.; Sukthankar, R. Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vis.* **2013**, *101*, 420–436.

24. Yuang, J.; Liu, Z.; Wu, Y. Discriminative subvolume search for efficient action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 22–24.
25. Eyben, F.; Wöllmer, M.; Schuller, B. OpenEAR—Introducing the munich open-source emotion and affect recognition toolkit. In Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–6.
26. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275. [CrossRef]
27. Wang, W.; Enescu, V.; Sahli, H. Towards Real-Time Continuous Emotion Recognition from Body Movements, Human Behavior Understanding. *Lect. Notes Comput. Sci.* **2013**, *8212*, 235–245. [CrossRef]
28. Gunes, H.; Piccardi, M. Affect recognition from face and body: Early fusion versus late fusion. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC '05), Banff, AB, Canada, 5–8 October 2017; pp. 3437–3443.
29. Patwardhan, A.S.; Knapp, G.M. Aggressive action and anger detection from multiple modalities using Kinect. *arXiv* **2016**, arXiv:1607.01076.
30. Patwardhan, A.S.; Knapp, G.M. EmoFit: Affect Monitoring System for Sedentary Jobs. *arXiv* **2016**, arXiv:1607.01077.
31. Patwardhan, A.S. Multimodal affect recognition using Kinect. *arXiv* **2016**, arXiv:1607.02652.
32. Sebe, N.; Cohen, I.; Huang, T.S. Multimodal emotion recognition. In *Handbook of Pattern Recognition and Computer Vision*; World Scientific Publishing Co.: Boston MA, USA, 2005; Volume 4, pp. 387–410.
33. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
34. De Meijer, M. The contribution of general features of body movement to the attribution of emotions. *J. Nonverbal Behav.* **1989**, *13*, 247–268. [CrossRef]
35. Dael, N.; Mortillaro, M.; Scherer, K.R. The Body Action and Posture Coding System (BAP): Development and Reliability. *J. Nonverbal Behav.* **2012**, *36*, 97–121. [CrossRef]
36. Schuller, B.; Muller, R.; Hornler, B.; Hothker, A.; Konosu, H.; Rigoll, G. Audiovisual recognition of spontaneous interest within conversations. In Proceedings of the 9th ACM International Conference on Multimodal Interfaces (ICMI'07), Nagoya, Japan, 12–15 November 2007; pp. 30–37.
37. Konstantinidis, E.I.; Billis, A.; Savvidis, T.; Xefteris, S.; Bamidis, P.D. *Emotion Recognition in the Wild: Results and Limitations from Active and Healthy Ageing Cases in a Living Lab, eHealth 360°*; Springer: Budapest, Hungary, 2017.
38. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Multimodal emotion recognition using deep learning architectures. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–9 March 2016; pp. 1–9. [CrossRef]
39. Zhang, Z.; Cui, L.; Liu, X.; Zhu, T. Emotion Detection Using Kinect 3D Facial Points. In Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, USA, 13–16 October 2016; pp. 407–410. [CrossRef]
40. Patwardhan, A.S.; Knapp, G.M. Affect Intensity Estimation Using Multiple Modalities. In Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference, Pensacola Beach, FL, USA, 21–23 May 2014.
41. Patwardhan, A.S.; Knapp, G.M. Multimodal Affect Analysis for Product Feedback Assessment. In Proceedings of the 2013 IIE Annual Conference, San Juan, Puerto Rico, 18–22 May 2013.
42. Sahoo, S.; Routray, A. Emotion recognition from audio-visual data using rule based decision level fusion. In Proceedings of the 2016 IEEE Students' Technology Symposium (TechSym), Kharagpur, India, 30 September–2 October 2016; pp. 7–12. [CrossRef]
43. Seng, K.; Ang, L.M.; Ooi, C. A Combined Rule-Based and Machine Learning Audio-Visual Emotion Recognition Approach. *IEEE Trans. Affect. Comput.* **2016**, *PP*, 1–11. [CrossRef]
44. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In Proceedings of the IEEE 16th International Conference on Data Mining, Barcelona, Spain, 12–15 December 2016.

45. Liu, H.; Cocea, M. Fuzzy rule based systems for interpretable sentiment analysis. In Proceedings of the 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, Qatar, 4–6 February 2017; pp. 129–136. [CrossRef]
46. Nguyen, D.; Nguyen, K.; Sridharan, S.; Ghasemi, A.; Dean, D.; Fookes, C. Deep Spatio-Temporal Features for Multimodal Emotion Recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 1215–1223. [CrossRef]
47. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Audio-Visual Emotion Recognition in Video Clips. *IEEE Trans. Affect. Comput.* **2017**, *PP*, 1. [CrossRef]
48. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–18. [CrossRef]
49. C# Reference. Available online: https://docs.microsoft.com/en-us/dotnet/csharp/language-reference/index (accessed on 10 August 2017).
50. .NET Framework 4. Available online: https://msdn.microsoft.com/en-us/library/w0x726c2(v=vs.100).aspx (accessed on 10 August 2017).
51. Kinect SDK 1.8. Available online: https://msdn.microsoft.com/en-us/library/hh855347.aspx (accessed on 10 August 2017).
52. Windows Presentation Foundation 4. Available online: https://msdn.microsoft.com/en-us/library/ms754130(v=vs.100).aspx (accessed on 10 August 2017).
53. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378–382. [CrossRef]
54. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1997**, *33*, 159–174. [CrossRef]
55. Ekman, P.; Friesen, W. Head and Body Cues in the Judgment of Emotion: A Reformulation. *Percept. Motor Skills* **1967**, *24*, 711–724. [CrossRef] [PubMed]