Multimodal Analysis of the JESTKOD database for Affective Dyadic Interactions

by

Sinan Keçeci

A Thesis Submitted to the
Graduate School Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Electrical and Electronics Engineering

Koç University

September 2016

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Sinan Keçeci

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Engin Erzin (Advisor)

_____

Assoc. Prof. Yücel Yemez

_____

Prof. Çiğdem Eroğlu Erdem

Date: _____

ABSTRACT

Gesticulation, together with the speech, is an important part of natural and affective human-human interaction. In human-computer interaction systems, natural, affective and believable use of gestures would be a valuable key component in adopting and emphasizing human-centered aspects. However, natural and affective multimodal data, for studying computational models of gesture and speech, is limited. In this study, the JESTKOD database is introduced, which consists of speech and full-body motion capture data recordings in dyadic interaction setting under agreement and disagreement scenarios. Each participant of the recordings are rated in dimensional affect space. This study includes the evaluations that are performed on the annotations of the database. These evaluations suggest important findings of usability of the JESTKOD database for investigating gesture and speech. Accordingly it is a valuable asset for designing more natural and affective human-computer interaction systems. Second part of this study also investigates mapping of human body motion observation across different sensor technologies to adapt the JESTKOD database knowledge to other sensors, such as the Kinect.

ÖZET

Jestler, konuşma ile beraber insan-insan arası iletişimin doğallığı ve duygusal algısı açısından önemli bir unsurdur. Vücut jestleri incelemesi adına yapılan bir çalışmanın insan-makine iletişiminin doğallığının arttırılmasına katkıda bulunacağını düşünüyoruz. Bu çalışmada jest, konuşma ve bunlarla beraber duygu değişimlerini incelediği JESTKOD veri tabanı için yapılmış olan duygu etiketleme çalışmaları ve bu çalışmalar sonucunda elde ettiğimiz verilerin değerlendirilme aşamalarını anlatılmaktadır. Bu çalışmada sunduğumuz değerlendirme sonuçları JESTKOD veri tabanının kullanılabileceği analiz ve modelleme çalışmaları açısından önem taşımaktadır. Aynı zamanda bu çalışmada JESTKOD veri tabanının Kinect gibi diğer sensör teknolojilerine de uyum sağlaması için yapılmış olan eşleştirme çalışmalarına yer verilmiştir.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and profound respect to Assoc. Prof. Engin Erzin for his support, guidance, patience and valuable teachings throughout my master program. I would also like to thank Assoc. Prof. Yücel Yemez for his important guidance and support. I feel myself privileged and lucky to be able to work with them.

I would also want to thank my colleagues in MVGL laboratory. I would like to thank especially Elif Bozkurt, Hossein Khaki, Bekir Berker Türker and Mehmet Ali Tuğtekin Turan for their valuable support and guidance. I deeply appreciate your help.

I would also want to thank my family for providing me the most positive family and home environment possible. Thank you for being the most supportive people in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Gesticulation is an essential part of face-to-face communication and it adds naturalness to the perception of human-to-human communication. Gestures which include planned hand, arm and head body gestures, co-exist with speech in time with a tight synchrony and they are shaped by a given emotional state. We tent to use gestures to complement or to emphasize speech. Although gesticulation is an important part of human-human communication, it is often a missing component of human-computer interaction (HCI) for virtual character animation and for multimodal tracking of an emotional state. Technological advancements and applications have created a demand for research on multimodal analysis and modeling of gesticulation and speech for HCI to attain interactions as natural and believable as human-human interactions. In the scope of the project "JESTKOD: Vücut Jestleri ile Konuşmanın Duygu ve Etkileşim Senaroylarında Çok Kipli Analizi ve Modellenmesi" we aim to develop models of affective and interactive human-human communication within a multimodal framework of gesticulation and speech. These models could be used in developing novel applications on emotional state tracking and affective virtual character animation.

Kendon [1] suggested a very commonly accepted gesture model which consists of multiple states. He defined gesture as a pulse which has a main part, a phase and a dynamic peak. This pulse might have a preparation and a recovery stage as well. In the higher steps of the hierarchy these pulses come together to build units that might refer to certain sentences of gestures. There are also certain studies that focus on the connection between the sound structure of the dialogues and the gestures. Kendon [1] states that there is a synchronization between gesture pulses and accented parts of the dialogues. Later this synchronism was defined by McNeill [1]: The peak of the gesture phase exists either before or together with the stressed part of the dialogue that it belongs to. According to McNeill [2], gestures can be classified under four classes: Iconic, metaphorical, indicative and cyclic.

Iconic gestures are performed to reference certain events or objects. Metaphorical gestures define the abstract events and indicative gestures are used to physically show certain objects. The cyclic gestures include the body movements which happen together with the progress of dialogue and emotion rhythm. The cyclic gestures have a big role in revealing the state of emotions in human nature. We started our project by examining these cyclic gestures to create a multi-modal understanding of the relation between speech, emotion and gesture.

In the literature, the modelling of the human communication channels using multimodal databases has created many research areas in speech processing, computer vision and machine learning. The multimodal analysis of affective dyadic interactions could create more natural human to computer interactions. Some of the important studies in this area include Casell's [3] Embodied Conversational Agent (ECA) which is one of the leaders in rule-based gesture synthesis. The approach to the gesture synthesis can be separated to two different categories: rule-oriented and data-oriented approaches. A rule-based system makes decisions according to a previously defined state diagram of gestures. In the data-based system, the decisions are made according to the statistical data that is obtained from 3D motion capture systems. Annotations of these motion-capture data are performed to define phases, groups and types of gestures. Statistical results are acquired from these annotations. The VirtualHuman project [4] and Neff's [5] work are examples of the animations that were performed according to statistical data obtained from 3D multimedia recordings. The aim of the VirtualHuman project [4] was to create a virtual character that had a personality profile. Neff [5] created character animations according to the statistical data acquired from annotations of multimedia recordings. The ECA [3] and the VirtualHuman project [4] map iconic and indicative gestures with the context of the dialogue together with predefined rules. On the other hand, Neff [5] related metaphorical gestures and dialogues using the statistical data. Recently Levine [6] and his friends created a new approach that uses gesture controllers. These controllers assign gesture phases to dialogue sentences by using references from a gesture library.

In our research we wanted a multimodal database that includes natural and affective monologues of a single subject as well as affective and interactive dialogues of two participating subjects. In building the multimodal database, we started our work by examining the existing resources which are the gesture modelling studies of USC SAIL group. These studies includes "IEMOCAP: Interactive emotional dyadic motion capture" [7] and CreativeIT [8] databases.

When we examined these databases we saw that the IEMOCAP [7] database is not sufficient for us to investigate body gestures. On the other hand the CreativeIT [8] database needed a further development in order to make the recorded gestures more natural. After these examinations we decided to create our own database for our needs.

In this study JESTKOD database which includes multimodal affective recordings of spontaneous dyadic interactions under agreement and disagreement scenarios is introduced. The JESTKOD database contains high-quality audio, video and motion-capture recordings of dyadic interactions and provides a valuable asset to investigate gesture and speech signals in natural and affective interaction settings. It has potential to create useful models for affective human-computer interaction systems.

Dyadic interaction requires social interactions, such as coordination and calibration. Bavelas [9] point out that person who has the speaking turn in a dialog constantly includes the addressee, and hand gestures can help the interlocutors coordinate their dialog and serve the special conversational demands of talking in dialog rather than in monologue. Supporting this point of view, Yang [10] show that individual's attitude as well as the interaction type can be predicted as friendly or unfriendly using only the dynamics of the hand gesture patterns over an interaction. Moreover, humans are able to distinguish statements of agreement from disagreement and neutral utterances on the basis of low-level nonverbal cues alone [11].

JESTKOD is a multimodal speech and body motion database, which is collected in natural dyadic interaction settings. Dyadic interactions are set with agreement and disagreement tasks to create affective variability. Recordings of each participant in the JESTKOD are annotated using the dimensional attributes of activation, valence and dominance (AVD). The main contribution of the JESTKOD is to present speech and motion data in natural dyadic interactions. Furthermore, use of the agreement/disagreement task to create the targeted variability in AVD attribute space is a secondary contribution.

In this sense, the JESTKOD database provides a valuable asset for investigating gesture and speech signals in affective interaction scenarios. As well it constructs a rich body motion repertoire in natural dyadic interactions that can contribute to speech driven gesture synthesis and animation.

To increase the usability of the JESTKOD database, selected upper body angles collected Kinect system are adapted with Motion capture system. Since JESTKOD database is

recorded using Motion Capture system this project will help integrating Kinect with our database for further machine learning applications.

This thesis is organized as follows. In Chapter 2, multimodal databases that contain continuous affect annotations are investigated. In Chapter 3, the data collection efforts for JESTKOD are explained. In Chapter 4, the statistical analysis of the annotations are described. Chapter 5 presents the efforts for mapping joint angles that are collected from both Kinect and Motion Capture system. Finally Chapter 6 presents the conclusions.

Chapter 2

LITERATURE REVIEW

There are a variety of multimodal databases that contain continuous affect annotations, which are publicly available for research purposes.

The HUMAINE database includes a large collection of multimodal naturalistic and induced recordings [12]. The SEMAINE database consists of audio-visual data in the form of conversations between participants and a number of virtual characters with particular personalities [13].

The acted audio-visual MSP-IMPROV database investigates emotional behaviors during conversational dyadic improvisations [14]. Although motion capture technologies are becoming widely available, there exist only a limited number of audio-visual databases which also include 3D motion data for modeling bodily gestures [15], explore technical setups, scenarios and challenges in building a motion capture database for virtual human animation. Busso [7] present their interactive emotional dyadic motion capture (the USC IEMOCAP) database, which is a multimodal and multi-speaker database of improvised and scripted dyadic interactions. USC IEMOCAP [7] database consists of 5 sessions of dyadic interactions which couples sit in front of a camera and their facial gestures are recorded. Only one of the participants hand gestures are recorded in each session. All recordings are annotated in activation dominance and valence emotion space by at least two people. This database is a very nice tool for examining facial gestures however it lacks the recordings of the body gestures since only one participant's body gestures are recorded.

The USC CreativeIT database [8] contains full-body motion capture information in the context of expressive theatrical improvisations [8]. It consists of 3 hours of recordings. The USC CreativeIT database [8] is annotated using the valence, activation and dominance attributes by at least 3 people. The theater performance ratings such as interest and naturalness are also rated. Since interaction performances of the USC CreativeIT database [8] are theatrical, speech and body gestures are rather amplified and exaggerated in their settings.

Chapter 3

JESTKOD DATABASE

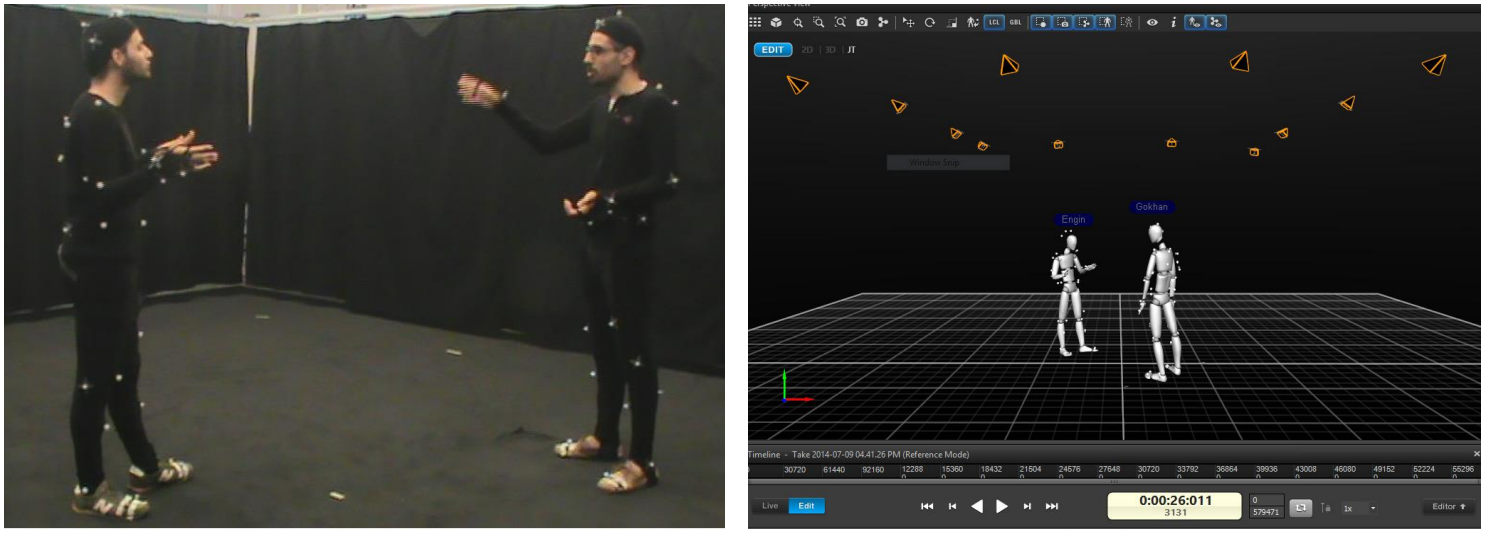
The main goal for the creation of the JESTKOD database is to put together a multimodal body gesture and speech data in a most naturally possible dyadic interactions. This database can be used to investigate body gestures in the context of spoken interactions.

3.1 Data Collection

The JESTKOD database consists of dyadic interaction recordings of 10 participants, 4 female and 6 male, ages from 20 to 25. For data collection, each participant was paired with the same interlocutor for both agreement and disagreement scenarios. Hence 10 participants were grouped in 5 pairings. Recordings for each pairing were collected in 5 sessions, all in Turkish. In each session, there are 19-23 clip recordings of 2-4 minutes, where in each clip participants discuss on a topic that they agree or disagree in dyadic interaction. The total duration of the recordings is 259~minutes. Recordings were captured by a high-definition video camera, a full body 3D motion capture system and Sony condenser tie-pin microphones. We used the *OptiTrack Flex 13*[1] system and the *Motive* [2] software for full body motion capture, which consists of 12 infrared cameras capturing 21 body joints at 120~fps with a resolution of 1280x1024. A sample scene from the video recordings and a screen shot from the motion capture software are given in Figure 3.1.

[1] *Flex 13 system - http://www.optitrack.com/products/flex-13/*

[2] *Motive Optical motion capture software - http://www.optitrack.com/products/motive/*



(a) Video Recordings (b) Motion – Capture

Figure 3.1 Video recording session (a) and motion-capture system sample (b) of the JESTKOD Database

Topics of the dyadic interactions were set by the moderator of the session using a preliminary information form, which was filled by all participants before the recordings. In the preliminary information form, participants were asked to state their favorite and disliked soccer teams, foods, restaurants, world cuisines, computer games, movies, operating systems, game consoles, etc. Using these forms we compiled a list of topics and paired up the participants with proper topics to create agreement/disagreement interactions during the recordings. The moderator instructed participants to engage in the conversation in their natural daily manners. Distributions of the topics in each session are summarized in Table 3.1. Note that, the same topic may appear in both agreement and disagreement interactions. For example, the soccer topic can initiate a strong disagreement interaction between participants supporting different teams. On the other hand, participants supporting the same team can engage in a congruent conversation. In the JESTKOD database we have 56 dyadic interactions in agreement and 42 dyadic interactions in disagreement with total durations of 154 and 105 minutes, respectively.

Session	Agreement Scenario	Disagreement Scenario
1	Movies, world cuisine, holiday resorts, TV series	Soccer, mathematics, game consoles, PC games
2	Soccer, world cuisine, movies, literature	Geography, holiday resorts, theater, dance
3	Movies, sports, PC games, music, world cuisine	Movies, history, animals, education
4	World cuisine, holiday resorts, science-fiction, history, theater, cities	Soccer, movies, PC games, TV series, literature, physics
5	Movies, languages, PC games, cities, game consoles	Movies, sports, holiday resorts, nutrition, musicals

Table 3.1 Summary of topics in the JESTKOD database recording sessions

3.2 Annotations

The annotations for the JESTKOD database are performed according to continuous time representation of the AVD modelling space which are activation, valence and dominance. Activation space describes how dynamic are the emotions of the person. As an example an excited person has high activation level and a distressed person has a low activation level. Valence space defines the positive and the negative emotion state of a person. As an example a happy person has high valence level and a furious person has low valence level. Third affect modelling space is the domination and it defines a person's authority among the incidents he/she encounters and also gives information about how successful that person handles these incidents. A detailed overview on the continuous representation of affect can be found in [16].

The aim of the JESTKOD is to represent natural variations of the affect in dyadic interactions. For data collection, participants were asked to engage in a conversation on a given topic, which triggered agreement or disagreement in an open-ended and continuous interaction.

Dimensional attributes of activation, valence and dominance (AVD) were used to assess affective variability of dyadic interactions.

Annotators rated each participant in the recordings using dimensional attributes of AVD in [0, 1] range. Ratings were collected using the general trace program (GTrace) from Queen's University [17]. Annotation effort was carried out separately for each participant in the recordings and for each dimensional attribute. A joystick interface was used with the GTrace software to deliver continuous-time ratings of the activation, valence and dominance attributes. A total of six annotators contributed to collect three sets of ratings for valence and dominance, and four sets of ratings for activation attribute. The annotators were selected among undergraduate engineering students who took a pre-training program on the annotation task and performed well in this task.

In the pre-training program candidates were instructed on definitions and characteristics of the attributes, and they practiced annotation task on three sample videos for each attribute. These three sample videos are annotated previously by psychology students. They received feedback on their annotation performance according to these previous annotations and repeated the task several times. In Figure 3.2 you can see an interface sample of the training program. The average duration of the pre-training program was around one hour.

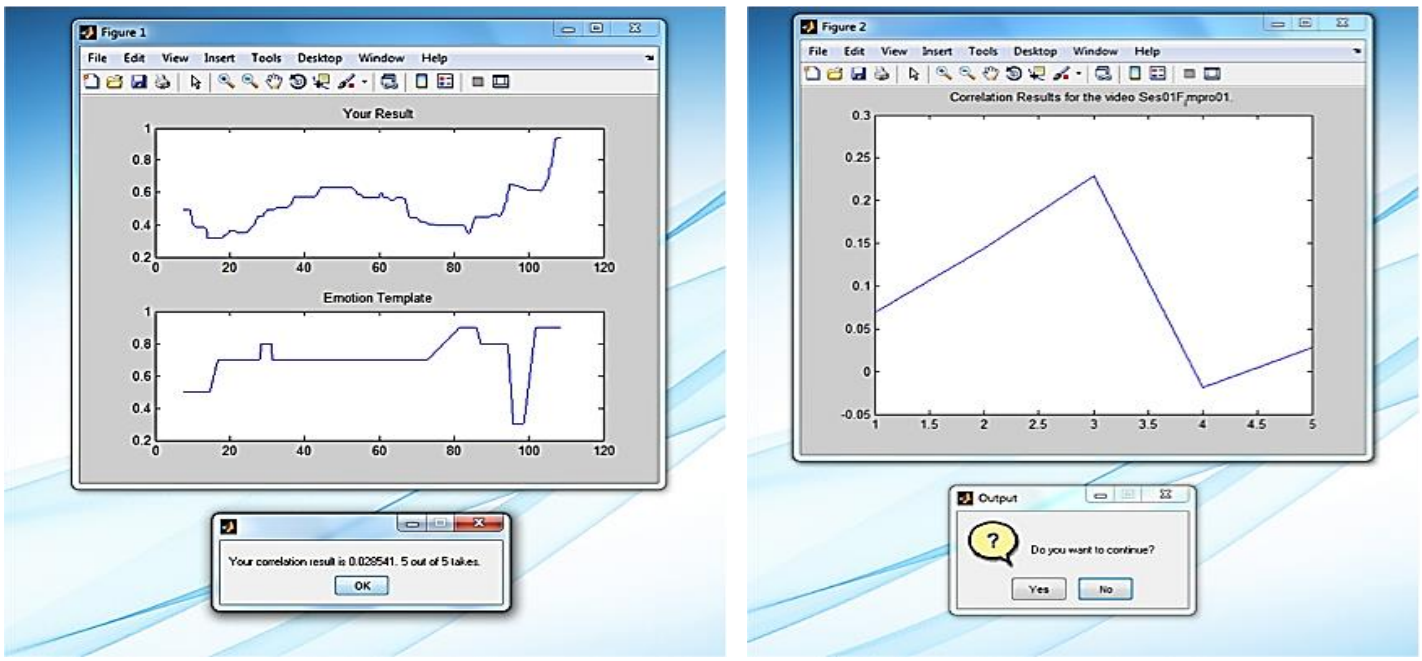


Figure 3.2 Interface sample of the annotation training program. On the left program gives feedback about the correlation result with the template. On the right the program gives feedback about overall performed after couple of takes.

Chapter 4

CONSENSUS RATING AND STATISTICAL ANALYSIS OF ANNOTATIONS

4.1 Consensus Rating

As mentioned in Chapter 3 we have completed the annotation process using a program which is called GTrace [17]. The outputs we get from GTrace are continuous time annotations in AVD space for each selected video. We have annotations that are completed by 3 people for dominance and valence and 4 people for activation space. A decision process needs to be performed using these annotations to come up with a final annotation for each video which we call “consensus rating”.

Consensus rating is extracted using correlation as a basis of concordance between annotators as in [18]. Each recording of a participant can then be represented with the consensus rating in AVD attributes. The consensus rating is extracted over temporal windows of duration 15 seconds, which are temporally 50% overlapping. 15 second window size is picked as a reasonable duration to correlate annotations. Shorter windows improve pruning quality however it is harder to expect high correlation results within a short window. On the other hand longer windows give better correlation results but pruning quality might degrade since if the elimination will be done by throwing longer segments of data which might have certain reliably annotated parts. The window-level ratings are linearly combined over the overlapping segments to set the final consensus rating for each recording of a participant. Linear combination over overlapping segments performs a smoothing over the consensus ratings to eliminate sharp transitions at the window boundaries. Window-level consensus rating computation is carried out to prune ratings of annotators in low agreement with the resulting consensus rating. For pruning of unreliable ratings, correlation coefficients between the consensus rating and individual ratings of the annotators are calculated for each window. Similarly to what is done in two recent related studies [18] [19].

In detail “consensus rating” extraction method is performed like this: For dominance and valence annotations are completed by 3 people. For these two emotion space the process starts with partitioning the annotations into windows. An appropriate window length is decided (15sec). If we concentrate only one window the method goes on like this: The mean of the annotations of all annotators is taken. In this particular window correlation coefficients between the mean and each annotation are checked. A threshold of 0.4 is set and if at least 2 of the annotations’ correlation results are above the threshold their mean is taken as the consensus rating for this particular window. If less than 2 annotations are below the threshold the mean of the two annotations with the highest correlation coefficients is taken as the consensus rating. You can check Figure 4.1 for visual explanation of this method.

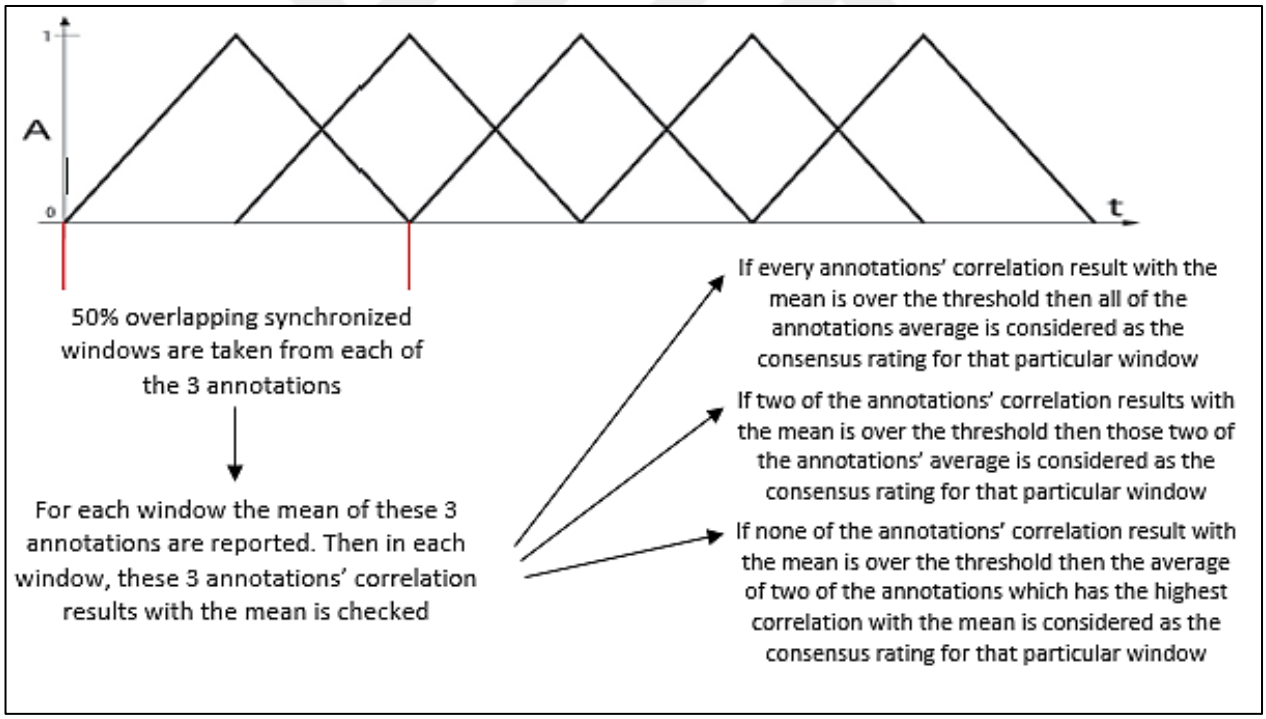


Figure 4.1 Consensus rating extraction method

For activation since we have 4 annotators a slightly different method is applied. First again the correlation with the mean is checked and the average of annotations is calculated as consensus rating if there are at least two correlations above the threshold. However if more than two are below the threshold we repeat the same process by defining a secondary mean with 3 of the annotations that has the highest correlation coefficient with the previous mean and remove the fourth one. When we go through all of the windows for each annotation and concatenate the results we end up with an overall consensus rating for that particular clip.

Figure 4.2 represents presents a sample plot with three individual ratings for dominance affect space and the extracted consensus rating.

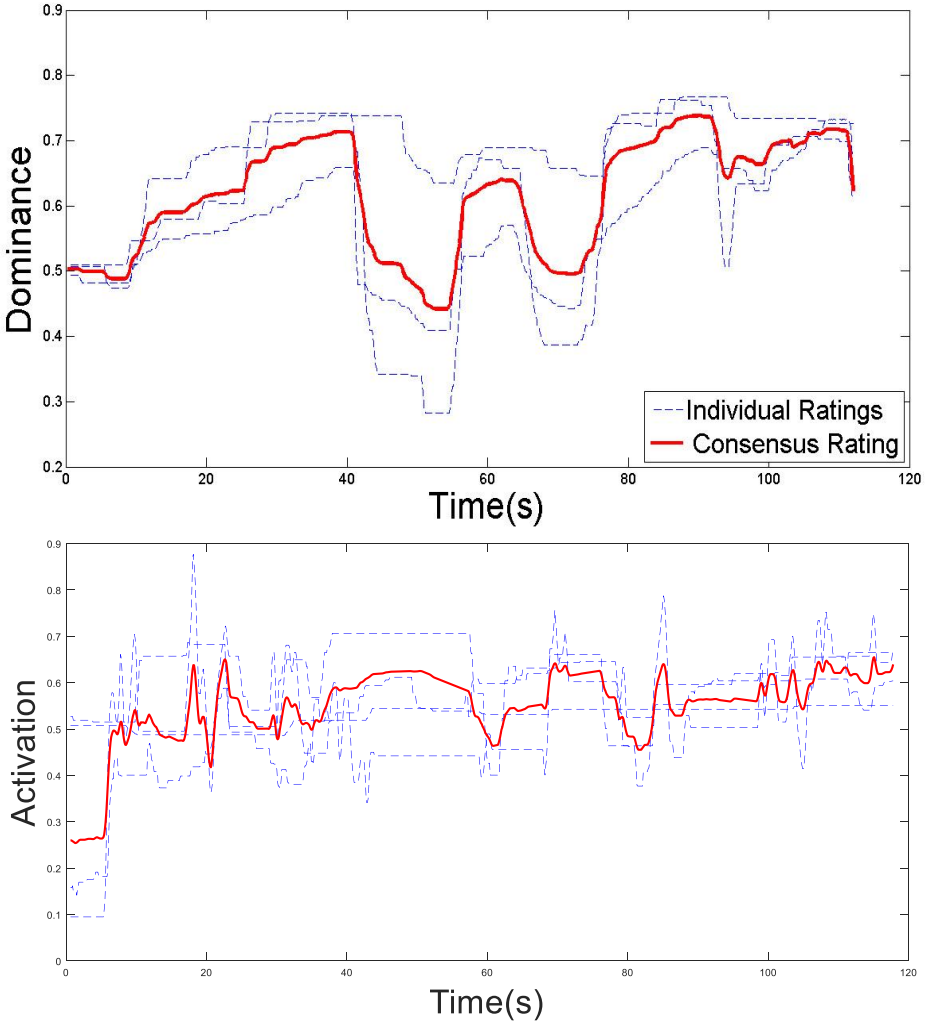


Figure 4.2 Sample ratings of annotators and the extracted consensus rating. Upper example calculated from 3 annotations since it is dominance and lower example calculated from 4 annotations since is activation.

4.2 Statistical Analysis of Annotations

In order to assess the quality of the consensus rating obtained by the procedure described in Section 4.1 the Pearson's correlation coefficients are calculated between the consensus rating and individual ratings of the annotators over the temporal windows. The mean and standard deviation values of correlation coefficients are reported in Table 4.1 for each of the AVD attributes. Note that the mean values of the correlation are all higher than 0.5. While the correlations under disagreement interactions are slightly higher for the activation attribute, they are lower than the correlations under agreement interactions for the valence and dominance attributes. Similarly, under the agreement interactions, concordance is the least with 0.56 mean correlation value for the activation attribute and it is higher with mean correlation values 0.61 and 0.76 respectively for the valence and domi nance attributes.

The maximum mean correlation values over all data are reported for the dominance as 0.75 and for activation and dominance it is 0.58.

	Activation		Valence		Dominance	
	MEAN	SD	MEAN	SD	MEAN	SD
Agreement	0.56	0.28	0.61	0.27	0.76	0.23
Disagreement	0.59	0.26	0.54	0.32	0.74	0.24
All	0.58	0.27	0.58	0.30	0.75	0.24

Table 4.1 Mean and standard deviation of the Pearson's correlation between the consensus rating and individual ratings for the activation, valence and dominance attribute under different interaction type

The ratio of windows with high correlation values, i.e. higher than the threshold value 0.4, is also considered as a statistical metric to evaluate the consensus rating. In Table 4.2 the percentages of windows are given that are over the threshold and are used to construct consensus rating. If we take a look the Table 4.2 for activation since there are 4 annotators there can be consensus ratings calculated with 4 annotators in each window. At 9.20% of all windows consensus rating is extracted by using 4 annotators. At 26.87% of all windows' consensus rating is extracted by using 3 annotators and at 39.98% of all windows' consensus rating is extracted by using 2 annotators. By saying 2 annotators at the last sentence let me remind you that these are the number of annotators that pass the threshold of correlation with the mean for that particular window. If we take a look at the total section of the graph for activation it is 76.05%, which means 23.95% of all windows' consensus ratings are constructed by cases that less than 2 annotations pass the threshold for those particular windows at the end of the consensus rating building process. Note that every window's consensus rating is constructed by at least 2 annotations that has highest correlation with the mean even if less than two annotation passes the threshold. For the cases of dominance and valence there are 3 annotators so the upper section is empty for that reason.

Annotation count to make consensus for particular window	Activation	Valence	Dominance
4	9.20	–	–
3	26.87	15.08	35.61
2	39.98	50.76	47.64
Total	76.05	65.84	83.25

Table 4.2 Ratio of windows of consensus rating that is calculated with annotations that has higher correlations than 0.4 with the mean annotation

The results in Table 4.2 can be interpreted as the level of concordance between annotators since we construct the consensus rating according to a correlation based approach. Note that dominance has the highest level of concordance with 83.25%. Activation has the second highest level of concordance with 76.05%, and finally valence has the lowest level of concordance with 65.84%. Even with the valence attribute, the annotators are in strong concordance with each other on two thirds of the database.

Another interesting investigation is to observe distributions of the consensus AVD ratings for agreement and disagreement interaction types.

AVD histograms over the consensus ratings are computed for each attribute separately for agreement and disagreement. Figure 4.3 plots the computed histograms. Activation and dominance do not convey significant distribution differences between agreement and disagreement. However, histograms of the consensus valence ratings differ significantly under agreement and disagreement interactions. Note that, as expected, the histogram in the case of agreement is shifted along the positive valence axis when compared to disagreement. The reason for this is that valence is directly related to positive and negative emotions however high or low active and dominant behavior can all include positive and negative emotions together. As an example in a disagreement case a person could be really furious and oppressive towards the other person that he is talking to which makes him highly active. Also on an agreement case for example a person do not have any control on the events which makes him less dominant. The point is that we expect a homogeneous looking histogram in activation and dominance cases because high and low attributes of these emotion spaces can be both found in agreement and disagreement scenarios.

Similarly, Figure 4.4 plots the joint histograms of the consensus AVD rating pairs. Note that activation-valence and dominance-valence consensus rating distributions significantly differ with the interaction type. Again in this joint histograms we can observe the influence of valence as the probability on the agreement scenarios is higher towards higher agreement levels.

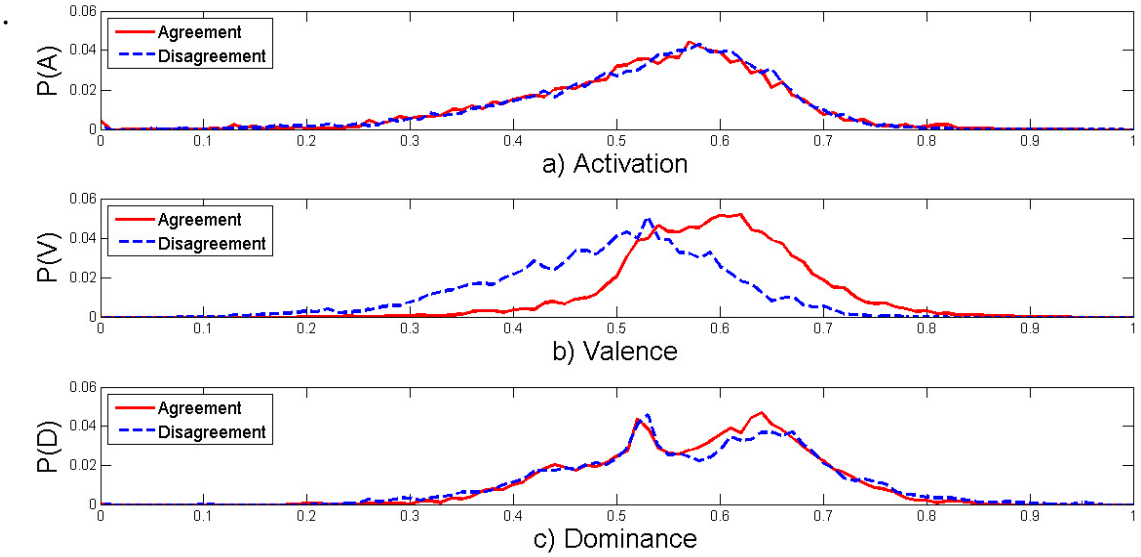


Figure 4.3 Histograms of the consensus ratings for activation, valence and dominance attributes under agreement and disagreement interactions

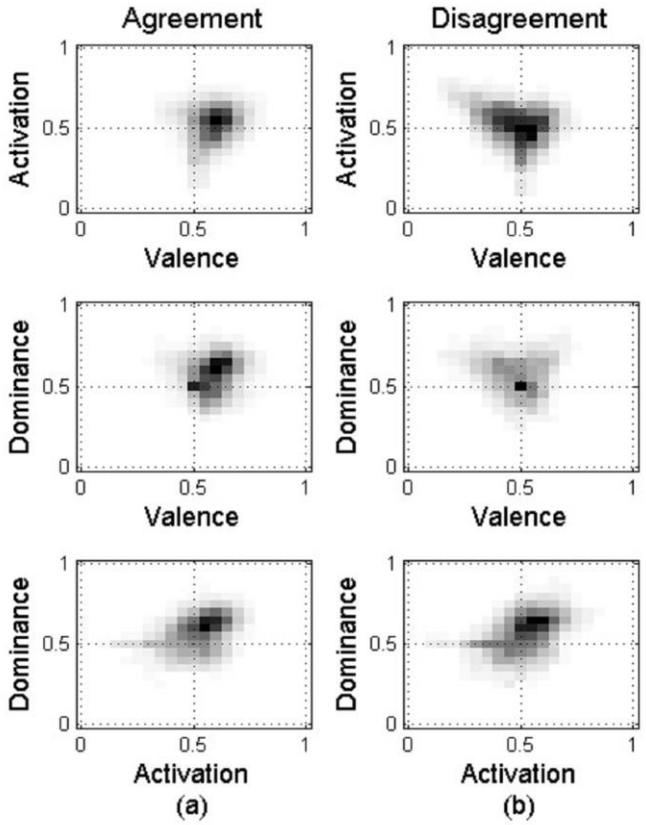


Figure 4.4 Joint histograms of the consensus rating for (a) agreement and (b) disagreement interactions over activation-valence, dominance-valence and dominance-activation attribute spaces

Kullback-Leibler divergence (KLD) is utilized to quantify statistical difference between agreement and disagreement distributions. The KLD and symmetric KLD can be defined respectively as,

$$D_{KL}(P_A||P_D) = \sum_k P_A(k) \log \frac{P_A(k)}{P_D(k)}, \quad (1)$$

And

$$D_{KL}(P_A, P_D) = \frac{1}{2}(D_{KL}(P_A||P_D) + D_{KL}(P_D||P_A)), \quad (2)$$

where $P_A(\ )$ and $P_D(\ )$ are probability distributions respectively over agreement and disagreement scenarios, and *k* runs over the sample space of consensus ratings. Table 4.3 presents on the diagonal the symmetric KLD distances between the distributions over agreement and disagreement scenarios for activation, valence and dominance, whereas the off-diagonal values represent the distance between joint distributions. All KLD distances are computed using the same number of bins for each attribute, hence they are comparable with each other in magnitude. The agreement and disagreement distributions have the strongest difference for the valence attribute with a KLD distance of 1.184. The KLD distance for the activation attribute is 0.033, which does not indicate significant statistical difference between agreement and disagreement, whereas dominance presents a KLD distance of 0.208. As for the joint distributions, dominance-valence yields the highest KLD distance as 4.697 while the activation-dominance distribution attains a distance of 1.157, which is lower than the KLD distance over the marginal valence distribution.

	$D_{KL}(P_A, P_D)$		
	Activation	Valence	Dominance
A	0.033	3.042	1.157
V	3.042	1.184	4.697
D	1.157	4.697	0.208

Table 4.3 Symmetric KLD distances between agreement/disagreement interactions for the consensus activation, valence and dominance ratings: diagonals are over the marginal distributions and off-diagonals are over the joint distribution

4.3 Conclusion

In this chapter the analysis on the JESTKOD database annotations are given. Some of the important findings include the ratio of windows having correlations higher than 0.4 between the mean annotation and individual ratings are 76.05% for activation, 65.84% for valence and 83.25% for dominance. This can be interpreted as even with the valence attribute, the annotators are in strong concordance with each other on two thirds of the database. Also another important point is that on the histograms of the annotations, at the valence space we can observe a higher probability of annotations leaning to positive values for agreement cases and negative for disagreement cases. This output show us that the annotators are represented the valence well and the recordings also well-defined and separation is clear in agreement and disagreement scenarios.

Chapter 5

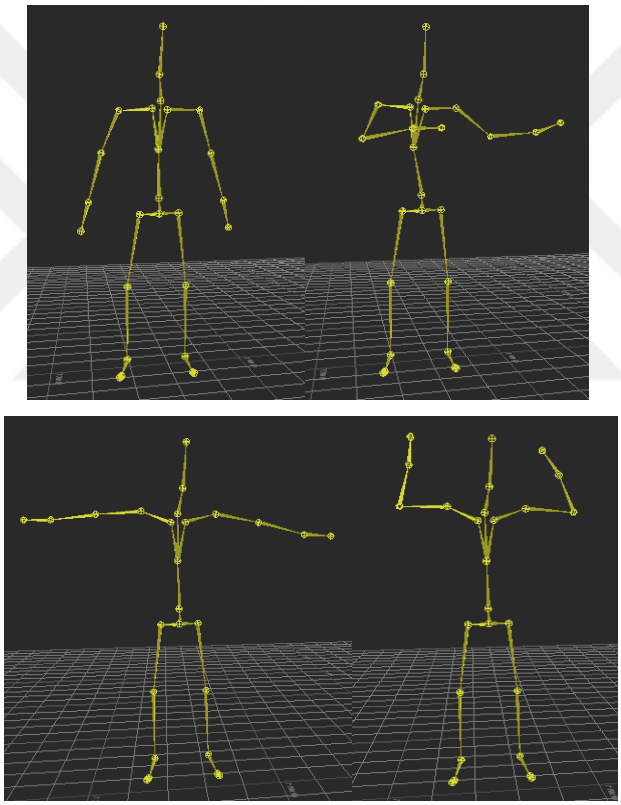
MAPPING JOINT ANGLES OF KINECT AND MOTION CAPTURE

In this chapter the efforts for mapping Kinect and Motion Capture systems are explained. For increasing the usability of our database JESTKOD, we decided that it will be a good idea to adapt the two systems in a way that recorded information from Kinect can also be used together with Motion Capture system for machine learning applications. Note that JESTKOD is recorded by using Motion Capture system. For this reason it seemed a reasonable idea to map body joint angles with each system. For this project only the upper body joint angles are considered. A new database is created. This database consists of gestures recorded together with Kinect and Motion Capture systems. The body angles that are calculated by using the coordinates collected from both systems are matched with linear regression method. To test the performance of the matching, "goodness of fit" measure is used and also gesture recognition performance is tested. This gesture recognition system is tested and trained both by Kinect and Motion Capture systems.

5.1 Database creation for mapping

For creation of this database Kinect version 2 and OptiTrack Motion Capture systems are used. 5 different gestures are recorded in 5 different sessions. In each session 20 gestures are recorded. The total recorded gesture count is 100, however 19 of these gestures are eliminated since they had did not have a proper recording. Since the synchronization of these gestures are very important for better results, after every gesture the actor clapped his hands for reference. The distance between right and left hands are calculated to find these reference points. Also during the recordings, attention is paid that the hands of the actor do not come as close as a clap during the recordings of the gestures.

The gestures are selected by keeping in the mind that these gestures must span regular movements of a natural conversation because this linear mapping of two systems will be used eventually for regular human interactions. These gestures must also give a good understanding of the behavior of the change in body angles. As an example if you take a look at Figure 5.1 you can see two of the gestures recorded with Motion Capture system's layout. Another point is that Kinect records data at a rate of 30 fps and Motion Capture system records at a rate of 120 fps. For the synchronization of the time samples interpolation is used and they are both set to 120 fps.

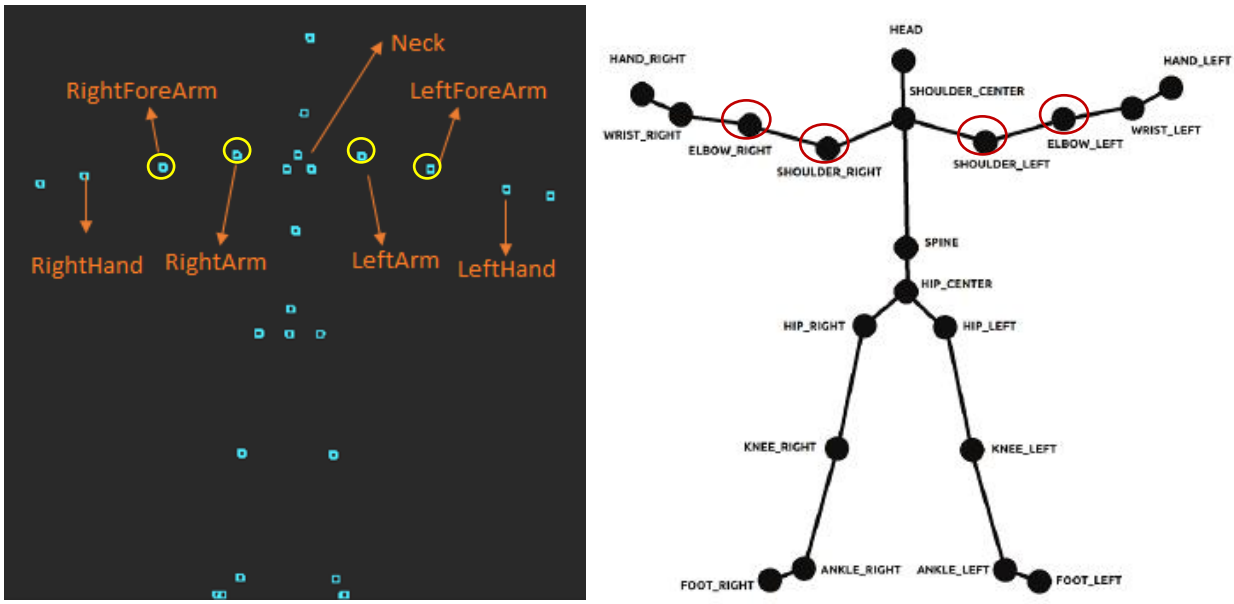


(a) (b)

Figure 5.1 Two examples of gestures recorded. These gestures are also recorded with Kinect systems simultaneously. Here we see the examples with the skeletal structure of the Motion Capture system the gestures are recorded to give a good example of regular conversation movements. In each figure section left figure (a) represent the beginning of the gesture and right figure (b) represent the end of the gesture movement recorded.

5.2 Calculation of the body joint angles

Many machine learning and animation applications use body joint angles for defining features. In this subchapter the calculation of the body joint angles will be explained. First of all, the joint angles for the mapping procedure is decided since there are two different skeletal structures, one from Kinect and one from Motion Capture system. In Figure 5.2 you can observe skeletal outputs from each system with corresponding joint names.



(a) (b)

Figure 5.2 Skeletal structure and joint names of Motion Capture on the left (a) and skeletal structure and joint names of Kinect on the Right (b)

Below you can find the definition of the calculated angles by joint names. The shoulder and elbow angles are decided for mapping. These angles are calculated from 3 dimensional joint information which is collected from both systems.

Motion Capture Right Shoulder Angle = ∠ (Neck, RightArm, RightForeArm)

Motion Capture Left Shoulder Angle = ∠ (Neck, LeftArm, LeftForeArm)

Motion Capture Right Elbow Angle = ∠ (RightArm, RightForeArm, RightHand)

Motion Capture Left Elbow Angle = ∠ (LeftArm, LeftForeArm, LeftHand)

Kinect Right Shoulder Angle = ∠ (shoulder_center, shoulder_right, elbow_right)

Kinect Left Shoulder Angle = ∠ (shoulder_center, shoulder_left, elbow_left)

Kinect Right Elbow Angle = ∠ (shoulder_right, elbow_right, wrist_right)

Kinect Left Elbow Angle = ∠ (shoulder_left, elbow_left, wrist_left)

These angles are calculated from 3 dimensional coordinates of the joints. Below the process of finding the angles from 3 points in space is as follows:

2 vectors are created from 3 points,

$$\overrightarrow{AB} = B - A$$
$$\overrightarrow{BC} = C - B$$

The dot product of these vectors are taken,

$$\overrightarrow{AB} \cdot \overrightarrow{BC} = \|\overrightarrow{AB}\| \, \|\overrightarrow{BC}\| \cos\theta$$

And finally by rearranging this formula the angle between 3 points in space can be found.

$$\theta = \arccos\left(\frac{\overrightarrow{AB} \cdot \overrightarrow{BC}}{\|\overrightarrow{AB}\| \, \|\overrightarrow{BC}\|} \right)$$

5.2 Linear Regression

It is logical to assume the relation between the angle data that is recorded simultaneously with both systems are linearly related. Therefore for mapping these data linear regression method is used. For linear regression to be successful the two quantities must be correlated. At this point it is very important that the data collected from Kinect and Motion Capture is very well sync together for maximum correlation. As mentioned in Chapter 5.1, for synchronization a hand clap is performed after each recorded gesture. After deciding the beginning and end points by this method, time samples are fixed on two datasets by interpolation since Kinect data is recorded in 30 frames per second and Motion Capture is recorded at 120 fps. At the end of this process we have perfectly synchronized datasets consists of angle vs frame graphs collected

from both systems. In Figure 5.3 the left elbow joint angle and left shoulder angle representation for one of the gestures recorded after synchronization.

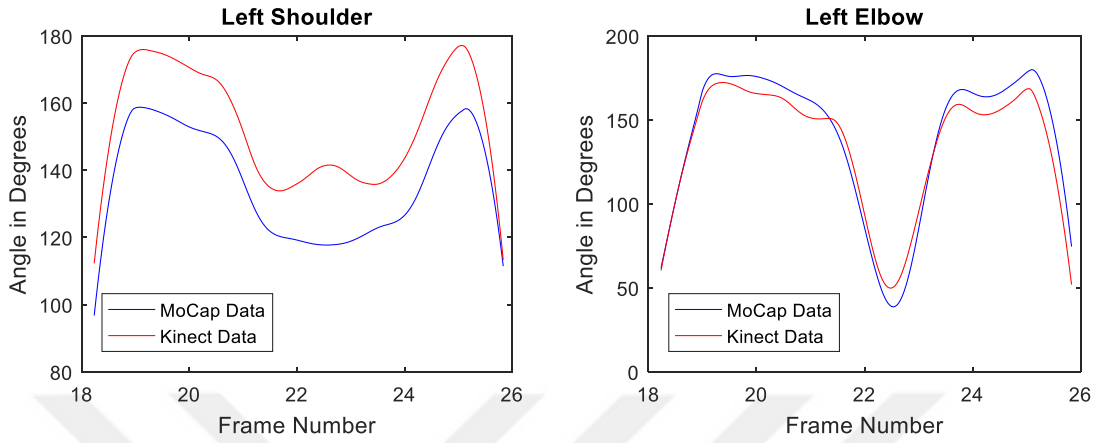


Figure 5.3 Left shoulder data on left and left elbow data on right one of the recorded gestures after synchronization

Linear regression is a method that is used fit a linear model to the data. Axis X consist of one set of observations and axis Y consist of the other set of observations. After the modeling when a value of X is then given without its matching value of y, the fitted model can be used to predict the value of y.

The relationship between the between the dependent variable y_i and regressors x_i can be given as:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Where $\boldsymbol{\beta}$ is the regression coefficient and ε_i is the error term. When n equations put together it looks like this:

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$ and the components are as follows:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

from this equation $\boldsymbol{\beta}$ can be found arranging the equation as $\mathbf{X^T y = (X^T X)\,\beta}$ and then the solution is $\boldsymbol{\beta} = \mathbf{(X^T X)^{-1}\, X^T y}$.

In our case the number of equations to solve is predicted by the degree of the polynomial which we want use to model our data. For example for a second order polynomial model the regressors for the first set of observations are taken as;

$$\mathbf{X = (x_{11}, x_{12}) = (x, x^2).}$$

For the degree of the polynomial decision of the linear model, the "goodness of fit" measure is used. The coefficient of determination is a statistical measure and it gives information about the goodness of the fit of a model. In regression, the coefficient R^2 gives information about how successful does the regression line approximate the real data points. An R^2 of "1" points out that the regression line perfectly fits the data and "0" means it is an unsuccessful fit.

The y vector which is $y = [y_1 ... y_n]$ is the dataset which each of its elements have a predicted value $f = [f_1 ... f_n]$ which comes from the modeling of the regression.

$\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

total sum of squares which is related to variance of the data can be found by:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

the residual sum of squares can be calculated by:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

and the coefficient of the determination is:

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

Since the database created for this mapping project includes 5 different gestures, a 5 fold training and testing system can applied for each body joint angle. The highest "goodness of fit" point taken into account for choosing the best fitting equation.

For each angle:

One gesture's data taken as a test from both Kinect and Motion Capture systems and the other 4 gesture's data is used for training. For the training part, Kinect data is concatenated to create the **X** data and Motion Capture Data is concatenated to create the **Y** data. The relationship with these **X** and **Y** data is found by linear regression considering different degree of fitting polynomial. For each gesture this calculated polynomial is applied to test data and the "goodness of fit measure" is checked. After 5 fold of operations, an overall equation for the particular angle is decided by the average of the weights of polynomials taken from each of the 5 training sections. The Table 5.1 presents the best fitting polynomial equations for each angle.

Angle	Degree of polynomial	Goodness of Fit	Equation
Right Shoulder	1	0.73	y = 0.88*x + 14.9
	2	0.72	y = -6.18e-04*x^2 + 1.06*x +1.7
	3	0.72	y = 4.5e-05*x^3+ 0.02*x^2+3.9*x-136
Left Shoulder	1	0.75	y = 0.83*x + 13.9
	2	0.74	y = -0.0014*x^2 + 1.23*x -15
	3	0.74	y = -8.9e-05*x^3 + 0.03*x^2 -4.45*x+255
Right Elbow	1	0.81	y = 0.94*x + 16.9
	2	0.82	y = -0.0031e-04*x^2 + 1.63*x -10
	3	0.83	y = -6.72e-05*x^3 + 0.02*x^2 -0.7*x+63
Left Elbow	1	0.8	y = 0.93*x + 25.4
	2	0.8	y = -8.9e-04*x^2 + 1.15*x +5.84
	3	0.8	y = -4.5e-05*x^3 + 0.01*x^2 -0.5*x + 63

Table 5.1 Linear Regression Results

If we take a look at Table 5.1 the most efficient choice of equations are highlighted. Selection of a degree choice of "1" maximizes the "goodness of fit" result for both right and left shoulder angles, although the "goodness of fit" (R^2 coefficient) does not fall immediately for right and left elbow angles they max around the same level of degree "1". Their goodness of fit results do not pass 0.83. Considering the computational costs the most logical choice for the fitting polynomial degree in this case is "1". Below in Figure 5.4 there are two figures that help giving idea of how well this mapping is completed.

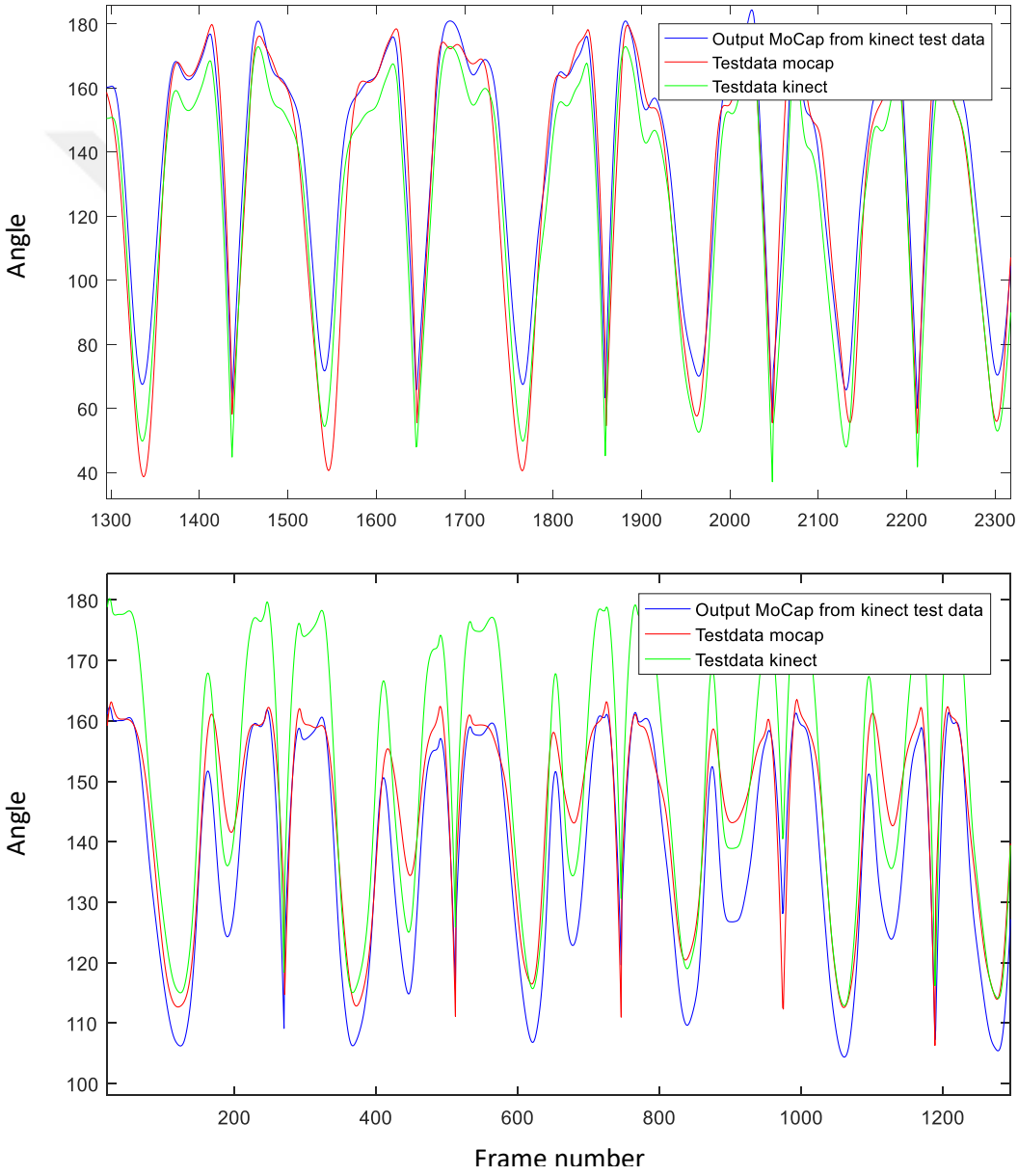


Figure 5.4 Angle vs Frame output samples from linear regression. Upper one is left elbow gesture and lower one is left shoulder angle for a gesture

5.3 Gesture Recognition with Dynamic Time Warping

In this section the implementation of a basic gesture recognition system is explained. For this gesture recognition the gesture database for linear regression is used. The data collected from Kinect and Motion Capture is mapped before the gesture recognition process by the results from Chapter 5.2. This gesture recognition system can be also considered as a performance test of the angle mapping process that is explained in the Chapter 5.2. In this gesture recognition system dynamic time warping technique is used to compare the similarity between the gesture angles.

Dynamic time warping is a technique which is usually used for calculating the similarity between different time series. The main reason that it is chosen for this application is that dynamic time warping can measure the similarity of signals with different time lengths. Since the gestures compared for recognition have different lengths of time this technique is chosen for recognition. The dynamic time warping algorithm that is performed is as follows:

We have two different signals to compare with different lengths n and m

$\mathbf{X = \{ x_1, x_2, \dots x_n\},\ \ Y = \{ y_1, y_2, \dots y_m\},}$

Create an n by m matrix "D" with all the Euclidean distances with each pair of data points,

$\mathbf{D(i, j) = d(x_i, y_j)}$ where $\mathbf{d(i, j) = | x_i - y_j | ,}$

create an empty cache matrix g with the size of n by m and apply the initial condition as **g(1,1) = D(1,1)** then complete the matrix as,

$$\mathbf{g(i , j) = \min \begin{Bmatrix} g(i, j - 1) + D(i, j) \\ g(i - 1, j - 1) + D(i, j) \\ g(I - 1, j) + D(i, j) \end{Bmatrix}}$$

finally the distance between two vectors is equal to **g(n,m).**

There are 5 gestures recorded in 5 session in the database created for mapping. For the training data of the recognition system, one of the gestures from each session which has the highest mapping correlation rate between Kinect and Motion Capture systems is chosen. These gestures are recorded and mapped at the finest and represents the gestures of the sessions in the best way. The other gesture are left for gesture recognition testing.

DTW distance is calculated between the angles of the training data joints and the angles of the test data joints and these distances are added for each test gesture. For a single test data this total distance of angles is calculated with each of the training gestures. The gesture name for a test data is decided by selecting the training data gesture which has the minimum total distance with that particular test data. A single distance measurement between one gesture's training data and one test data looks like this:

Total Distance between one gesture's Training Data and Single Test Data =
dtw(shoulderRightAngle_Training, shoulderRightAngle_Test) +
dtw(shoulderLeftAngle_Training, shoulderLeftAngle_Test,) +
dtw(elbowRightAngle_Training, elbowRightAngle_Test)+
dtw(elbowLeftAngle_Training, elbowLeftAngle_Test)

Table 5.2 gives the gesture recognition rate for each experimental setup including the recognition results without linear mapping between two systems. The highest recognition is achieved between Motion Capture test data and Motion Capture Training data and the lowest recognition is reported between Kinect test data and Motion Capture training data. If we investigate the improvement of the recognition with the linear mapping we can see that there is an increase from 69.74% to 92.11% in the experimental setup that includes Kinect test data and Motion Capture training data. This improvement is important because this scenario is more related to our purpose which is using Kinect motion data with our already existing JESTKOD database. The reason for this improvement is that the Kinect data varies a lot more in the recordings than the Motion Capture system. Motion Capture system is more stable due to the joint markers on the recorded actor and 360 degree recording space however Kinect records the coordinates from one angle and without markers on the actor during recordings. The linear mapping generates a Kinect Data that is better for comparison with the more stable Motion Capture data. In the experimental setup that includes Kinect training data and Motion Capture

Test Data there is a less improvement in recognition which is from 96.05% to 97.37%. This is an expected result and it is due to the fact that Kinect training data is selected from the gestures that are already recorded and synchronized very well with the Motion Capture data. Since the variance between the recordings in the Motion Capture test data is small between gestures because it is a very stable recording system we observe less improvement in this experimental setup.

	Kinect Test Data	Motion Capture Test Data
Kinect Training Data	96.05%	97.37% (96.05% without mapping)
Motion Capture Training Data	92.11% (69.74% without mapping)	98.68%

Table 5.2 Gesture recognition rate using DTW for each experimental setup.

5.4 Conclusion

This linear mapping method could be very important for increasing the usability of our JESTKOD database. Since the gesture recognition rates are high, linear mapping show that it is the right method to be used for this task. However using this technique requires a good synchronization of the recorded data. Notable experimental results for gesture recognition include 97.37% between Kinect training data and Motion Capture test data and 92.11% between Motion Capture training data and Kinect test data. Also the improvement of the recognition with the linear mapping between Motion Capture training data and Kinect test data from 69.74% to 92.11% is also important since it is related to the scenario which is using recorded gestures from Kinect with already existing JESTKOD Database.

Chapter 6

CONCLUSION

In this study, the multimodal database JESTKOD is introduced. This database consists of speech, motion capture data and video recordings of continuously annotated affective dyadic interactions under agreement and disagreement scenarios. A detailed description of the data collection process and the annotation of recordings in the continuous affect domains are presented. The analysis of the annotations assures that the annotations have completed in a quality that it is ready to be used for machine learning applications. Also the joint angle mapping application of the two systems could open new areas for applications involving JESTKOD database.

Some of notable success of this research include during the calculation of consensus rating, the ratio of windows that have correlations higher than 0.4 between the mean annotation and individual ratings are 76.05% for activation, 65.84% for valence and 83.25% for dominance. Also the mean of the Pearson's correlation between the consensus rating and individual ratings are for the activation and valence 0.58 and for dominance 0.75. This results all contribute to the reliability of this database.

Another point is that this is a Turkish database. Since there are not many Turkish resources that investigates multimodal analysis of gestures this database could be a very valuable tool for Turkish researchers. This database will be available as a package on the internet for everybody.

The outputs of linear mapping of Kinect and Motion Capture systems showed that it is possible to use Kinect information together with JESTKOD database which could create new applications of research for machine learning. As a future work this linear mapping could be applied to rotational joint angles as well. This application would make creation of the animations easier.

BIBLIOGRAPHY

[1] Adam Kendon, “Gesticulation and speech: Two aspects of the process of utterance,” in The Relationship of Verbal and Nonverbal Communication, Mary Ritchie Key, Ed., pp. 207–227. Mounton Publishers, The Hague, the Netherlands, 1980.

[2] David McNeill, Hand and Mind: What Gestures Reveal about Thought, University Of Chicago Press, 1992.

[3] “Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents„” p. 1–27, 2000.

[4] Gebhard P. Lockelt M. Ndiaye A. Pfleger N. Klesen M. Reithinger, N., “Virtualhuman – dialogic and emotional interaction with virtual characters,” in Proceedings of the Eighth International Conference on Multimodal Interfaces, 2006.

[5] Kipp M. Albrecht I. Seidel H.P. Neff, M., “Gesture modeling and animation based on a probabilistic re-creation of speaker style,” ACM Trans on Graphic, vol. 27, no. 1, 2008.

[6] Krahenbuhl P. Thrun S. Levine, S. and V. Koltun, “Gesture controllers,” ACM SIGGRAPH, p. 124:1–124:11, 2010.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, Nov. 2008.

[8] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth S. Narayanan, “The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation,” in Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC), May 2010.

[9] Bavelas JB, Chovil N, Coates L, Roe L (1995) Gestures specialized for dialogue. Personality and social psychology bulletin 21(4):394-405

[10] Yang Z, Metallinou A, Erzin E, Narayanan S (2014a) Analysis of interaction attitudes using data-driven hand gesture phrases. In: Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)

[11] Mehu M, Maaten L (2014) Multimodal integration of dynamic audio {visual cues in the communication of agreement and disagreement. Journal of Non-verbal Behavior 38(4):569-597

[12] Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin JC, Devillers L, Abrilian S, Batliner A, Amir N, Karpouzis K (2007) The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In: Paiva

A, Prada R, Picard R (eds) A_ective Computing and Intelligent Interaction, Lecture Notes in Computer Science, vol 4738, Springer Berlin Heidelberg, pp 488-500

[13] McKeown G, Valstar M, Cowie R, Pantic M, Schroder M (2012) The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on A_ective Computing 3(1):5-17

[14] Busso C, Parthasarathy S, Burmania A, Abdel-Wahab M, Sadoughi N, Provost EM (2016) MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. IEEE Transactions on A_ective Computing

[15] Heloir A, Ne_ M, Kipp M (2010) Exploiting Motion Capture for Virtual Human Animation: Data Collection and Annotation Visualization. In: Proc. Of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality

[16] Gunes H, Schuller B, Pantic M, Cowie R (2011) Emotion representation, analysis and synthesis in continuous space: A survey. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, pp 827–834

[17] Cowie R, Cox C, Martin JC, Batliner A, Heylen D, Karpouzis K (2011) Issues in Data Labelling. In: Petta P, Pelachaud C, Cowie R (eds) Emotion-Oriented Systems: The Humaine Handbook, Springer-Verlag, Berlin Heidel- berg, pp 215–244

[18] Metallinou A, Yang Z, Lee Cc, Busso C, Carnicke S, Narayanan SS (2015) The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. Lanuage Resources and Evaluation

[19] Devillers L, Cowie R, Martin J, Douglas-Cowie E, Abrilian S, McRorie M (2006) Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches. In: Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC), Genoa, Italy