Towards Modelling Human Interaction

Jingjing Deng

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Master of Science, by Research

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

2012

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

In this thesis, we examine whether 3D pose and face features can be used to both learn and recognise different conversational interactions. For example, can we distinguish from these cues whether two people are most likely discussing work or a recent holiday experience? We believe this to be the first work devoted to this subject and show that this task is indeed possible with a promising degree of accuracy using both features derived from pose and the face. To extract 3D pose we use the Kinect Sensor and to extract face features we use a texture based model. Whilst both of these contain noise we show that they can still be used to build classifiers to perform this task. The Gaussian mixture model (GMM) is applied to the pose features to investigate the data structure and its distribution characteristics. Based on the clustering result using GMM, Random Forest classifier is used to classify conversational scenarios, globally. Then, the Hidden Markov Models (HMM) and Coupled Hidden Markov Models (CHMM) are employed to model the interactions, by which the temporal characteristics of the dynamics process during the conversational communication can be learned. The CHMM using two Markov chains is shown to be effective in capturing the dynamics of interaction. Moreover, the multi-modal features are also considered in this work, i.e. face and pose. It is observed that modest performance can still be achieved whilst observing only a single participant of the interaction. Though this performance is significantly improved when combining cues provided by multiple people. The interactions in our data set are extremely subtle, we do not employ actors or ask our participants to exaggerate any motions or responses. It is also worth noting that the interactions we are concerned with are not micro or instant events, such as hugging and high-five, but rather conversational interactions over a period of time that consists rather sim-

ilar individual, micro actions and interactions. That is actions and interactions at a short time window are shared commonly across all the scenarios in our dataset. We compare using both HMM and CHMM and show that the HMM is outperformed by the CHMM on both face and pose features. Finally we demonstrate that face and pose features can be combined to improve overall performance.

Table of Contents

Chapter 1

Introduction

Contents

1.1 Motivation

Automatic action recognition from video still remains an open and challenging problem. One of the principal reasons is that direct pose estimation in the first instance is problematic and inaccurate, particularly in 3D. There is however already a body of work interested in the detection and recognition of social interaction between multiple people, which is particularly difficult since the actions of multiple people must be inferred and understood. However, the dependence on low-level features has meant that the class of social interactions examined thus far typically have been limited to those that can be readily identified and most easily described by a particular set of motions or poses, e.g., a handshake or high-five. Alternatively, observation is made at a coarse level to recognise interactions which are only dependent on high-level tracking of entire individuals, e.g., in a surveillance setting.

This work is in part motivated by recent work that showed that features derived from

3D human pose are much more discriminative than their low-level image based counterparts [AYG11]. Therefore, we believe that having access to these features provides the capacity of detecting and classifying much more subtle interactions than currently possible. Often the differences between the interactions examined in this work are not themselves intuitive and we are therefore unable to build an intermediate representation based on our perception of what we think may or may not be important features.

1.2 Overview and Contributions

In this thesis, we propose to leverage recent advances in technology in extracting 3D pose using the Microsoft Kinect and its SDK to examine the feasibility of detecting much more high-level behavioural interactions between two people. This work is only possible due to the availability of the consumer sensor (Microsoft Kinect), as it permits a cheap and convenient means to both collect a large amount of 3D pose data. Rather than recognizing just key social events, we attempt to analyze and detect different conversational interactions. We investigate whether just by observing the 3D pose of two interacting people we can recognise the type of conversation they are conducting. For example are they talking about a recent vacation or a new idea for a work project? We do not as in previous work studying interaction examine strongly differentiable interactions, such as high-tempered arguments or disputes. Neither do we employ the use of actors. Subjects are not told the purpose of our work and they are simply asked to discuss different topics, such as tell one another a story, debate a topic or discuss a question. They are not asked to exaggerate any aspect of their behaviour. We collected over 15 hours video including full body image and face image, and 7.5 hours Kinect data including 3D skeleton sequences and depth maps. All the sequences are synchronised manually, and ready for publishing.

The unsupervised learning approach, GMM (Gaussian Mixture Model) is applied to the features which are extracted from Kinect sequences in order to study the characteristics of the feature data. Firstly, we perform clustering across whole dimension of the features and constructed histogram of clusters for each scenarios to investigate the cross correlation among

them. Secondly, the GMM is applied to each dimension of features space, then the discriminative algorithm, RF (random forest) is applied to classify different scenarios by frames or clips using the features constructed from the clustering using GMM.

The generative model HMM is then applied to model the dynamics during the conversational communication. In this part of work, in addition to examining the use of pose features, we also examine the benefit of using facial features. Meanwhile, both visual cues which simultaneously provided by two conversation participants are considered and modelled by coupled HMM with two Markov chain conditional dependent to each other, each of which corresponds to one participant. We believe this to be the first work devoted to conversational interactions where we are interested in identifying the content of a conversation using pose and facial features. These have not previously been applied to distinguishing between types of conversation and in this context it is unclear whether pose and facial features are indeed complementary, although multiple-modality cues have been demonstrated to be of benefit in related tasks [NGP11]. We conduct an empirical study to examine the extent of the visual cues provided by humans in recognizing conversational interactions. In addition, we also investigate the importance of modelling interaction, i.e., can we still recognise the social interaction just by observing a single person? Moreover, we also explore the use of facial features in interaction modelling. Whilst in itself novel, the benefits of this are two fold: Firstly, we can examine how pose features compare to facial features for conversational analysis. Secondly, we examine whether face and pose features are complementary and can be combined to build stronger classifiers.

1.3 Thesis Layout

The rest of the thesis are organized as follows:

- **Chapter 2**, - *Background*: This chapter provides a brief review of related works, which includes some representative works of action, activity and facial expression recognition. We have a particular focus on overview of HMM based approaches as they have been

found effective in handling sequential data and modelling the conditional dependence between the subjects by introducing more Markov chains. Also, some of related work on facial expression and affect recognition are discussed in this chapter.

- **Chapter 3**, - *Data Acquisition and Preparation*: This chapter describes the experiment set up for data collection, and provides examples of video and Kinect sequences for each conversational scenarios. All the sequences including Kinect kinematic chain and video on face and full body are synchronised manually. Furthermore, three fiducial points for locating the interesting region of the face image are labelled.

- **Chapter 4**, - *Features Extraction*: In this chapter, we mainly discuss the methods of feature extraction which include the upper body motion, head orientation, and similarity transformations of the interesting region of face.

- **Chapter 5**, - *Data Analysis using GMM and Random Forest*: The GMM algorithm is applied to the pose features Histograms of clusters for 7 scenarios are constructed from the results of the clustering, and the cross correlation is investigated. Based on the results of clustering, RF is applied in order to classify different scenarios.

- **Chapter 6**, - *Interaction Modelling using HMM*: The single HMM and coupled HMM are applied to model the conversational interaction. Different schemes of feature coupling and combination are investigated in order to compare the performance of the single HMM and the coupled HMM. Whether or not the face and pose features are complementary to each others is examined as well.

- **Chapter 7**, - *Conclusions and Future Work*: This chapter summaries the research finding, and focus on the discussion of interaction modeling and the advantages and disadvantages of use of multi-modality features. It also provides suggestions for future work.

- *Appendices A-C*: The consent form, the instructions for the experiment which we provided to the participants and the questionnaire which the participants were asked to fill in after data recording are provided.

Chapter 2

Background

Contents

2.1 Overview

Action recognition systems can often be built on relatively easy to extract low-level features, such as temporal SIFT features [SAS07] or temporal Harris corner features [Lap05]. Whilst typically the actions being recognized may be easily distinguishable from a visual perception point of view (e.g., waving, jumping jacks), these approaches can also be used to describe surprisingly subtle behaviours. For example when applied to mice, behaviours such as grooming, drinking and eating can be distinguished [DRCB05] and the intentions of shoppers inferred from surveillance cameras [HCL+09]. Low-level features, such as Local Binary Patterns [SGM09], can also be used to recognise basic facial expressions [SGM09, KP07]. However, social interactions are more complex and difficult to recognise since the actions, motions and motivations of multiple people must be understood. For example, as well as placing

a bounding box around each participant, higher-level information such as head orientation or body pose may first be needed. Despite these limitations social interactions such as high-fives, hugs and kisses have been shown to be detectable in unconstrained scenes [PPMZR10] and the ability to detect key social interactions, such as eye contact in real-world scenes demonstrated [MJZF11]. Whilst these works are impressive, the accuracy of the features extracted are likely to limit the ability to discriminate between more complex or subtle interactions.

There has been some success in using features extracted from high-level information such as body pose, e.g., automatically learning sign language [BEZ09]. However, typically assumption about the subjects in the scenes, such as body orientation, are made to first constrain the solution. A further problem with studying social interaction is that there will often be occlusion since usually participants would face one another, meaning observations are often incomplete. For this reason, often the interactions examined are less intimate and can be viewed at a coarser resolution. For example Zhang *et al.* [ZGPBM06] studied group interactions in a work meeting between multiple people, detecting events such as presenting to the group, conducting a group discussion or note taking etc. This is achieved by first estimating the state of each participant and then using this information to infer the group action.

Decomposing the group interaction into a two level process of firstly inferring what each person is doing, and then from this deducing the group action is a common approach [ZGPBM06, ORP00, AR11]. Probabilistic models such as Hidden Markov Models (HMM) can be employed to overcome noisy observations, both at the image level and on the person dependent action classification level. However, for this approach to be effective there needs to be an understanding of which motions, poses or gestures that an individual performs is likely to be an important building block. Often this is dependent on the granularity of the actions being observed. For example, detecting a stretched out hand may be intuitively of importance for detecting a hand-shake compared to a hug, however, for detecting the difference between whether a subject was nervous or confident whilst performing a handshake this representation is likely to be too coarse. It is unlikely to expose the affective state of the person being observed.

2.2 Action and Activity Recognition

With the progress of visual motion and action recognition[AC99, Gav99, MHK06], recently, various methods for modelling interaction process have been proposed, all of which are widely used in intra-group and inter-group activities analysis, such as Human-Human interaction, Human-Object interaction and Group-Group interaction. Two comprehensive reviews can be found in [AR11, TCSU08], which have covered the most of related works before 2010. In term of the way of representing the elements of interaction, essentially, the methods fall into two categories, state-based approach and rule-based approach. In this section, we briefly review those previous works which emphasize on interaction modelling, and compare their contributions with our proposed work, see Table 2.1.

Table 2.1: An incomplete summary of action and activity recognition work

Approach	Representative work	Model	Hierarchical	Observation	Targets
State-based	Nuria Oliver et al.[ORP00]	CHMMs		Agent Trajectories	Human Activity
	Nuria Oliver et al.[OGH04]	LHMMs	✓	Video, Keyboard and Mouse Activity	Human Activity
	Zhang et al.[ZGPBM06]	LHMMs	✓	Video and Speech	Group Activity
	Natarajan et al.[NN07]	CHSMM	✓	Video	Human Activity
	Dai et al.[DDD+09]	EDM-DBN	✓	Video and Environmental Configuration	Group Activity
	Loy et al.[LXG10, LXG12]	xCCA and TD-PGM	✓	Video	Group Activity
Rule-based	Ivanov et al.[IB00]	SCFGs	✓	Video	Human Activity
	Joo et al.[JC06]	Attribute Grammar	✓	Video	Human Activity
	Ryoo et al.[RA09]	CFG and HMM	✓	Video	Human Activity
	Ryoo et al.[RA11]	CFG and MCMC	✓	Video	Group Activity

2.2.1 State-based Approaches

Oliver et al.[ORP00] modelled the interaction by constructing a coupled HMM which derived from the thought of that each chain correspond to individual subject, and at time t, all states depends on the states at time $t-1$ not only in the chain itself but also in other chains. It is obvious that the interaction in the temporal scale can be learned by maximum a posterior (MAP) state estimation. In their paper, the interaction model were tested on a visual surveillance dataset, and CHMMs with two dependent chains successfully classify three different

behaviours. Meanwhile, compared with the results given by basic HMMs, CHMMs performed much better.

However, the single-layer HMMs would generate a huge parameter space which could only be learned with amount of train data, while the accuracy and performance could not meet the need of real application. Oliver et al.[OGH04] presented Layered HMMs (LHMMs), a hierarchical statistical method for modelling activities happening in the office. In LHMMs, the complex activities are split up into various levels, each of which is represented by one HMMs. The upper layer HMMs take the inferential results given by lower neighbouring layer HMMs as its observation. By its nature, both the logical and temporal structure of the activities are represented in the model. The interactions between humans in a meeting environment have been detected and recognized using the system built based on the LHMMs. As well as Zhang et al.[ZGPBM06] demonstrated the ability of multi-layered HMMs in recognizing group activities under a meeting environment. In their two-layered HMMs system, taking advantage of the visual and audio features extracted from cameras and microphones respectively, the bottom layer HMMs gives the results of recognizing atomic actions, such as speaking, writing and idle, in terms of which the top layer HMMs can represented group activities, such as discussion, monologue, presentation and so on.

Subsequently, variation of HMMs are investigated by many research groups. To address the limitation that traditional HMM with first order Markov assumption would lead to exponential decay in duration probability of the state, Natarajan et al.[NN07] proposed coupled hidden semi Markov models (CHSMM) for learning human interactions in a sign language recognition task and a simulated visual surveillance task. Compared with parallel hidden Markov models (PaHMM) and CHMM, CHSMM significantly improve the accuracy at the expense of efficiency. Dai et al.[DDD$^+$09] proposed a event-driven multilevel dynamics Bayesian network (EDM-DBN) for modelling dynamic context in group interactions. Being similar to LHMMs, the state nodes which correspond to the events are well organized in terms of different abstract levels of group interactions.

In addition, human activity understanding with multiple cameras also have been studied.

Loy et al. presented a new cross canonical correlation analysis (xCCA) to modelling time delayed correlation within the activities captured by multiple non-overlapping cameras[LXG10]. A time delayed probabilistic model (TD-PGM) was constructed for learning visual context changes over time, where different nodes correspond to activities in different areas from different cameras, and the time delayed dependencies are encoded by the directed links between nodes[LXG12]. The methods are successfully applied to detect the abnormal events occurred at a busy underground station using the surveillance camera network.

2.2.2 Rule-based Approaches

A hierarchical methods taking advantage of HMMs for recognizing atomic actions and stochastic context-free grammars (SCFGs) for parsing complex activities was proposed by Ivanov and Bobick[IB00]. Given a large number of stochastic productions rules, all activity possibilities in their environment can be well encoded, so the high-level activities recognition task can be simply carried out by parsing these strings consisting of the low-level atomic actions. The limitation of the method based on purely syntactic grammar is that it is incapable to encode the features by finite symbols. In order to deal with the problem, Joo and Chellappa[JC06] introduced an attribute grammar extended from SCFG for event recognition and anomaly detection. In their model, the features and the additional attributions are attached and associated with primitive events, the event can be successfully recognized when both syntax rules of specified event and feature constrains are satisfied.

Different from these syntax-based approaches which have been mentioned above, description-based approaches adopt the CFG as a formal method to describe the features of activities. The role of CFG is changed from recognizing activity to representing activity. Ryoo et al.[RA09] proposed a probabilistic semantic-level recognition algorithm marrying up the HMMs and CFG based representation scheme for recognizing composite activities. The system they constructed can recognise both simple interactions like 'pushing' and recursive interactions such as 'fighting'. Ryoo et al.[RA11] extend their works to recognition of group activity by introducing a Markov chan Monte Carlo (MCMC) based probability distribution sampling method. In the system, the group activities are described in terms of CFG hierarchically, a stochastic searching

for the individuals satisfying the representation of the group activities with highest probability is carried out. Some complex activities such as 'group stealing' and 'group assaulting' can be recognized.

Whilst we have been unable to include all work on both action recognition and interactions in this section we refer interested readers to the following comprehensive reviews [AC99, Gav99, MHK06, AR11, TCSU08] The methods which have been mentioned above make great contributions to the visual activity recognition task especially to the interaction modelling. For state-based approaches, HMMs is the first choice for modelling sequential signal, where the interaction features of activities are encoded by the transition probability of each state in one or more Markov chains. The task can be performed by learning the transition trends between states along time scale. On the other hand, rule-based approaches used to utilize the HMMs to recognise the atomic events or actions which are elementary component of activity descriptor. The inner characters of interaction and relations between each subjects involved in the activities are represented by the rules of formal language which is given before hands. As result, detecting the specified activity can be achieved by searching in the bag of sentences which are formed by the atomic descriptor in terms of the constraint rules.

2.3 Facial Expression Recognition

For facial expression and behavioural analysis, mid-level descriptors defined by the Facial Action Units System (FACS) [EF78] have been introduced to build detectors to provide high-level annotations. This system consists of forty six basic action units for modelling all types of facial movements and have been proposed to analyze facial behaviours, such as recognizing facial expression. A variety of frameworks are designed for this task. Tian *et al.* [TKC01] extracted shape features and transient features from multistate face components to train a three-layer neural network to recognise a set of action units. In [TLJ07], a Dynamic Bayesian Network was used to learn the relationship among action units on Gabor features.In [PP06], Pantic and Patras tracked facial salient points and used a rule-based system to classify the action units.

Our work is also related to affect recognition [ZPRH09, NGP11], where the task is to gain insight into the mood or emotional state of the person being observed by indirect means such as facial expression or pose. Although facial expressions are treated as the most important cues to identify emotions in traditional research [PR00], multiple cues and modalities have been investigated to learn whether the extra information helps to improve affect recognition performance [NGP11]. We also investigate multiple modalities in this work, examining the performance of facial and pose features. However, in contrast to affect recognition, where a single observation can typically be used to identify the affective state (e.g., smile implies happiness), there is not a direct connection between a single observation and the type of the conversation being performed; rather it is the sequence of observations as an interaction is in progress that is of importance.

Expressive facial animation is another related task which has wide range of applications in computer graphics, including video games and virtual communications. Cao *et al.* [CTFP05] builds a generative model of expressive motion driven by related speeches. Deng *et al.* [DUL$^+$06] proposes to combine synthesized neutral visual speech motion by speech co-articulation models and expression changes from an expression subspace, called Phoneme-Independent Expression Eigenspace to generate natural facial animation from motion capture data. We believe further investigation on social interactions enable to facilitate the improvement of animation performance.

It is worth noting that the emphasis of this thesis is on pose feature which includes both bodily motion and head orientation, although we do use the face feature in modelling conversational interaction.

2.4 Discussion

There has long been an interest in modelling and learning complex social interactions between multiple people. From what we have mention above , HMMs are a popular choice for modelling social interactions since they are well suited for detecting actions with varying tem-

poral duration. A first order HMM has a simple structure containing a set of hidden nodes to represent the internal state of the process and attached to each of these an observed state. This allows the likelihood of a particular node being in a given state to be estimated from extracted features. There are many variations of HMM's used in the context of recognizing and understanding social interactions. Other Hierarchical methods use HMMs for recognizing lower-level atomic actions and other methods for high-level integration [IB00, RA09]. Ivanov and Bobick use Stochastic Context-Free Grammars (SCFGs) for parsing complex activities [IB00, JC06], which allows high-level activities to be performed by parsing these strings consisting of low-level atomic actions. The limitation of the method based on purely syntactic grammar is that it is incapable of encoding the features by finite symbols. Joo and Chellappa [JC06] hence introduced an attribute grammar extended from SCFG for event recognition and anomaly detection. Ryoo *et al.* [RA09] proposed a probabilistic semantic-level recognition algorithm marrying up the HMMs and Context Free Grammar based representation scheme for recognizing composite activities.

One problem with these hierarchical approaches is that they require the lower-level semantic actions to be defined manually. For the interactions explored in this work no prior knowledge is available about which features are most indicative for each conversation observed. Therefore, we are not in a position to exploit these methods. Instead we use a single layer HMM, so we are not imposing our own design features onto it which if wrong, could mask any information contained within the observations. We also experiment with using coupled HMMs, which allows the state spaces to be learned independently for each subject, rather than simply concatenating the features together, therefore doubling the dimensionality of the training space.

In addition, examining the use of facial features is another highlight of this work. We consider that, from a psychological perspective, both upper body language and facial expression play important role in the social activities. There is few work studying the interaction process occurred in activity micro-cosmically using multi-modality cues [NGP11]. In this work, we are going to study whether the pose feature and face feature are complementary to each other by

comparing the performance of coupled or concatenated features with the pose and face features alone.

Chapter 3

Data Acquisition and Preparation

Contents

3.1 Experiment Setting

Data was collected using a multi-camera set-up, each person was recorded using a Kinect Sensor, which captured pose at 30fps. The face was captured using a High Definition Camera operating at 25 fps and the entire body by another camera at PAL 25fps that was used primarily to keep a visual record of the participants motions. The participants were positioned about a meter apart facing one another. So that the participants did not occlude one another; each of the cameras was slightly offset from a direct frontal view. They were then asked to stand approximately on a marker on the ground to remain in view of the cameras and Kinect sensors.

The data was captured in a lecture theatre and the participants monitored through a camera so that no one was required to be in the room with the participants to allow natural interaction. The sound was not monitored so the participants did not feel conscious about what they were saying. They were also told that sound would be removed from the final dataset. The capture

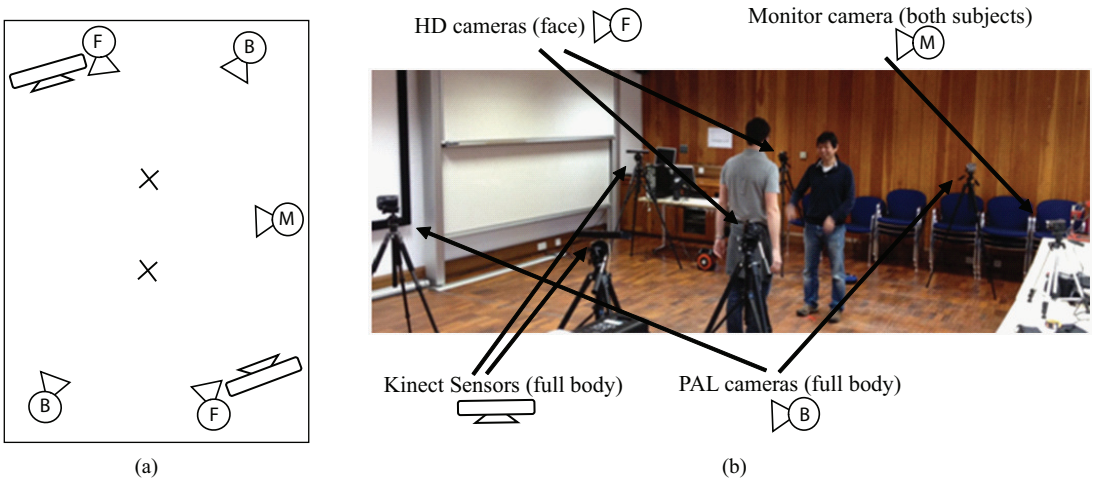

Figure 3.1: Overview of the data collection. (a) shows a top view of the capture area and the positions of the cameras and Kinect sensors. The participants were asked to stand on the 'X' marks. (b) shows an image of the actual environment used with the cameras and Kinect sensors labeled.

area used for data collection is depicted in Figure 3.1.

The participants were given seven tasks to complete, five of which they were asked to prepare before arriving. The first task was to discuss an area of their current work. The second task was to prepare an interesting story to tell their partner, such as a holiday experience or an experience of a friend. The third task was to jointly find the answer to a problem. The fourth task was a debate, where the participants where asked to prepare arguments for a particular point of view on an issue we gave to them. The fifth task was a discussion where they were asked to discuss between them the issues surrounding a statement and come to agreement whether they believe the statement is true or not. The sixth task was to answer a subjective question and the seventh task was to take it in turn telling jokes to one another. A full description of the different tasks are provided in Table 3.1, and the consent form, joke and question materials, and the questionnaire can be found in Appendix.

Each set of tasks took about 50 minutes for a pair to complete. They were told roughly how long we expected each task to take as a guide, however, they were not being timed. Before

Table 3.1: Description of each of the tasks given to the participants to perform. The rightmost column describes whether the participants were told about the task and asked to prepare before attending.

Task Name	Description	Prepared
Describing Work	Each participant was asked to describe to their partner their current work or a project they have involved with. Following this each participant then repeated it back so as to confirm they had understood.	yes
Story Telling	Each participant was asked to think of an interesting story they could tell their partner, such as a holiday experience or an experience of a friend.	yes
Problem Solving	The participants were given a problem they were asked to think of the solution of together. The problem was "Do candles burn in space and if so what shape and direction?".	no
Debate	The participants were asked to prepare arguments for a given point of view on the topic "Should University education be free?" and then debate this between them.	yes
Discussion	The participants were asked to jointly discuss the issues surrounding a statement and come to agreement whether they believe the statement is true or not. The statement was "Social Networks have made the world a better place"	yes
Subjective Question	The participants were asked to discuss a subjective question which was "If you could be any animal, what animal and why?".	no
Jokes	The participants were asked to take it turn telling jokes to one another, each participant was provided with three different jokes to learn before attending.	yes

each task, there were given the opportunity to reread any associated material with the task that they may have forgotten. At the end of the session, participants were generally surprised by how much time had passed. A sample of the data collected for each conversational interaction is presented in Fig 3.2.

3.2 Data Pre-Processing

The length of raw data collected for this thesis is shown in Table 3.2. The cameras and Kinect were synchronised manually, this is likely to have a small error attached to it, though as we intend to use features derived over a longer temporal window than typically used we would expect the effect of this to be negligible. Fig 3.3 shows examples of both the full body image sequence, face image sequence and Kinect skeleton sequences after synchronisation.

In addition to synchronising the face video sequences, the ground truth of the region of interest for face analysis is also hand labelled. Some samples of region of interest are illustrated in Fig 3.4. A standard 2D template of face which including the positions of eyes' centre and mouth centre and the region of interest spanned by these points is prepared. The geometrical similarity transform of these three points from the template to face image is then computed, based on which the region of interest within the template can be wrapped to the face image.

Figure 3.2: Examples of observations made for each pair during different conversational interactions. The time difference between each consecutive frame shown is two seconds.

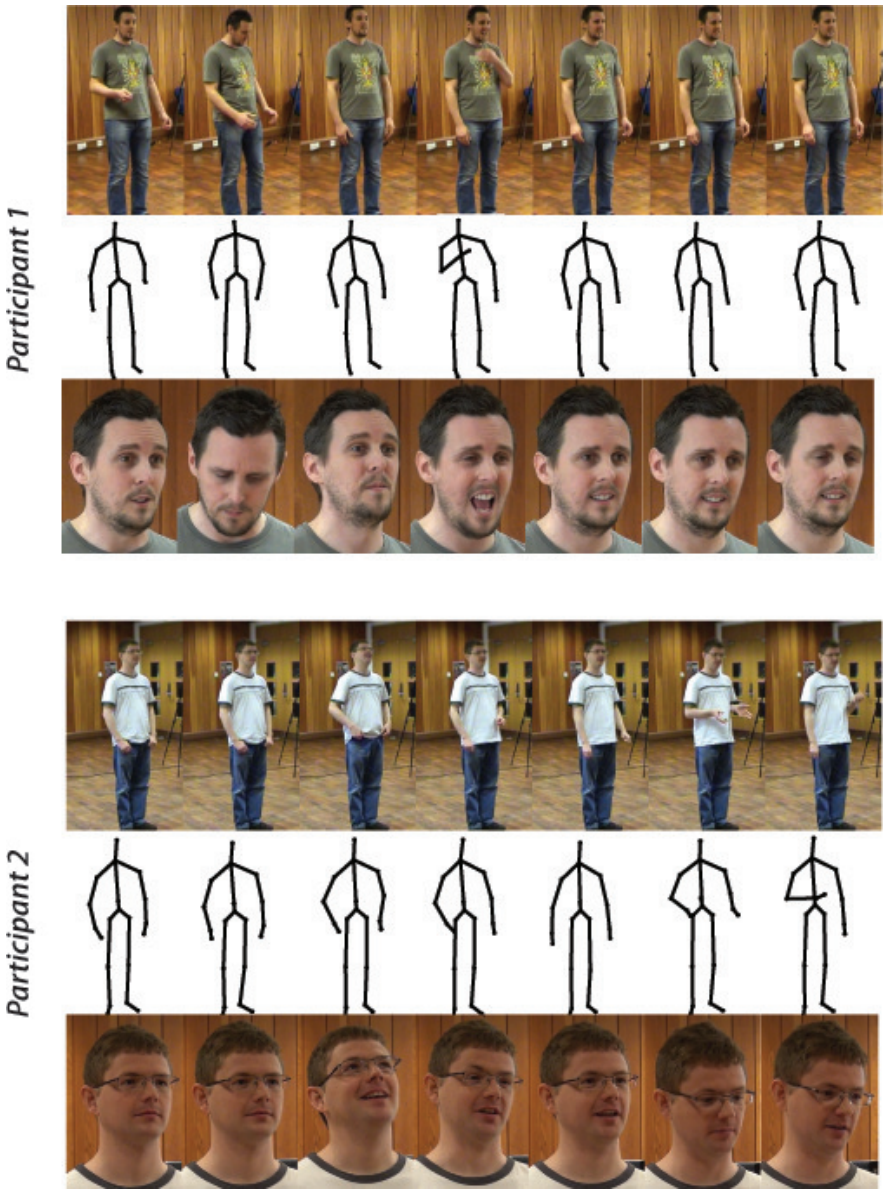

Figure 3.3: Examples of observations from full body camera, Kinect sensor and face camera, made for the pair G2 during the "Joke Telling" scenario. The time difference between each consecutive frame shown is two seconds.

Table 3.2: The length of the raw data collected for this work under the experiment setting. The length of the video including face and full body video is counted in mm:ss:ff, where mm stands for minutes, ss seconds and ff frames. The length of the Kinect sequence is counted in frames.

Group	Participant	Face Video	Full Body Video	Kinect Sequence
G1	P1	33:51:03	33:52:18	60483
	P2	34:25:24	35:18:17	61411
G2	P1	32:48:18	31:23:09	55223
	P2	34:10:09	33:48:11	55656
G3	P1	53:47:06	59:37:12	115653
	P2	45:14:10	60:15:17	115429
G4	P1	53:47:07	41:12:10	73617
	P2	45:14:10	41:28:18	74497
G5	P1	53:46:10	45:46:16	82947
	P2	53:46:07	47:06:09	81027

So, the face region can be located by tracking three fiducial points on itself. In the data pre-processing stage, we labelled these three fiducial points every 10 frames manually, in order to provide the ground-truth for semi-automatic tracking algorithm to evaluate the performance and re-initialise when it fails. The main purpose of manually locating the face area are two folds. Firstly, they are used to evaluate the performance of face tracking algorithm. Secondly, as we shown in Fig 4.4, the face images have contained occlusions such as occluded by hands or glasses, which causes great difficulties in the fiducial points tracking. An elaborate approach may be employed to reduce manual re-initialisation. However, the emphasis of this work is to investigate how to model conversational interactions given reasonable localisation of both facial and bodily features.

3.3 Summary

In this chapter, we describes the process of data collection and necessary data pre-processing. Each pair of participants were asked to carry out 7 different conversational tasks as follows: describing work, story telling, problem solving, debate, discussion, subjective question and telling joke. For each participant, the 3D pose data were directly captured by the Kinect sensor, and two cameras are used to recode the views of full body and face close-ups. The video sequences and Kinect output were synchronised manually, and the ground-truth of fiducial points were labelled by hands as well. The next chapter will discuss the techniques of feature extraction based on these well prepared data.

Figure 3.4: Examples of manually labelled interesting region (blue squares) of face image using three fiducial points (red crosses).

Chapter 4

Features Extraction

Contents

In this chapter, we describe the features used to represent both pose and face. The pose features are extracted from motion capture data collected using the Microsoft Kinect Sensor and represent the motions and distances between different parts over time. In addition to these features we also extract the head orientation using the geometry of features tracked on the face. We use these additional features as it is likely head orientation and motion is a strong cue during conversational interactions. To represent the face we extract texture based features to implicitly represent shape and expression.

4.1 Pose Feature Extraction

3D pose features have been shown to be useful in motion capture data retrieval and action recognition. Motivated by the works, such as [KG04, MRC05, AYG11], we extract three types of features to depict the pose and motion of upper body. These geometry features extracted

from a kinematic chain are simple but powerful for representing human gesture and motion over time. Fig 4.1 shows the kinematic chain model with 20 joints used by Kinect sensor. The first feature we use measures the distance between two joints at different time intervals and is depicted in Fig. 4.2(a). The second feature measures the distance between a joint and a reference plane defined using different parts of the body (see Fig. 4.2(b,c)). The third feature measures the velocity of a joint (see Fig 4.2(d)).

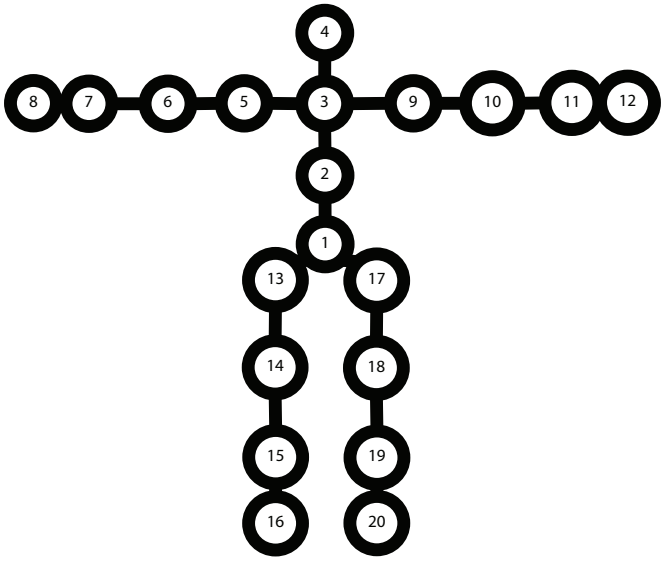

Figure 4.1: The kinematic chain model with 20 joints introduced by Kinect sensor.

The 3D location of a joint at time slice t is denoted as $\omega_{i,t} \in R^3$ and the vector defined by two joints by$\pi_{ij,t} \in R^3$, where i and j indicates the identity of the joints. We define two types of plane $\phi_{ijk,t}$ which is spanned by the joints $\omega_{i,t}, \omega_{j,t}, \omega_{k,t}$, and the plane $\psi_{ijk,t}$ passing through $\omega_{k,t}$ and whose normal vector is aligned with $\pi_{ij,t}$. The normal vector of the plane $\phi_{ijk,t}$ can also be represented by $\pi_{ijk,t}$.

The feature F_d representing the Euclidean distance between joints over Δt is defined as follows:

$$F_d = Distance\{(\omega_{i,t}), (\omega_{j,t+\Delta t})\} \tag{4.1}$$

If $i = j$, then the feature measures the distance of movement of the joint over time Δt, otherwise,

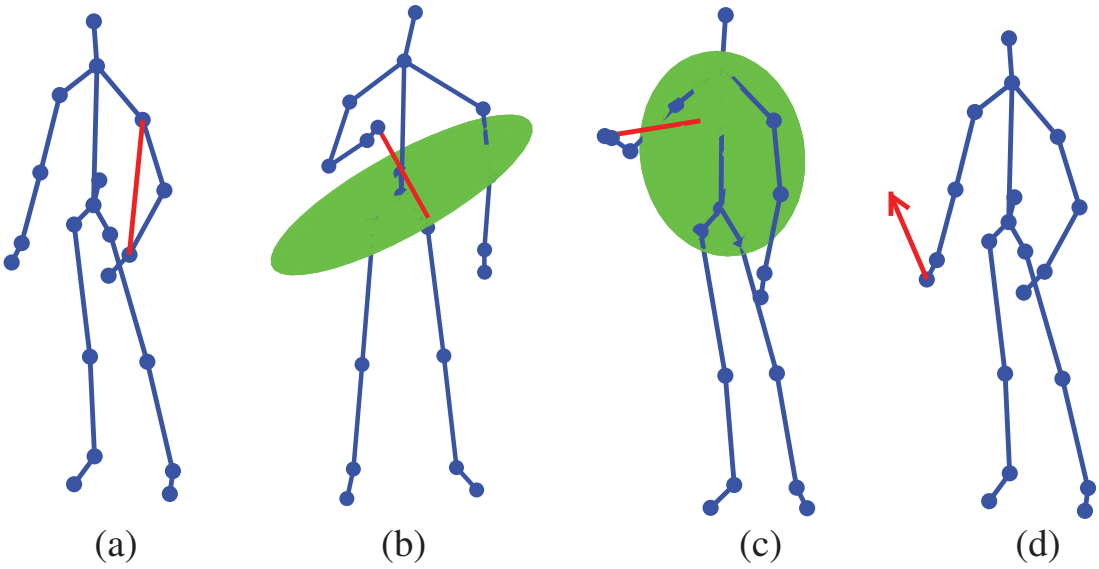

Figure 4.2: Visualization of the pose-based features. (a) Illustrates the distance between joints feature. (b) & (c) The distance of a joint to a reference plane. (d) The joint velocity

it measures the distance between two different joints separated by time.

The features F_{pd1} and F_{pd2} measure the shortest distance from joint $\omega_{n,t}$ to the plane $\phi_{ijk,t+\Delta t}$ and the plane $\psi_{ijk,t+\Delta t}$ respectively.

$$F_{pd1} = Distance\{(\omega_{n,t}),(\phi_{ijk,t+\Delta t})\} \tag{4.2}$$

$$F_{pd2} = Distance\{(\omega_{n,t}),(\psi_{ijk,t+\Delta t})\} \tag{4.3}$$

We also extract F_{jv}, F_{pv}, the component of the joints' velocity along the direction of the vector $\pi_{ij,t+\Delta t}$ and vector $\pi_{ijk,t+\Delta t}$ respectively.

$$F_{jv} = Velocity\{(\omega_{n,t}),(\pi_{ij,t+\Delta t})\} \tag{4.4}$$

$$F_{pv} = Velocity\{(\omega_{n,t}),(\pi_{ijk,t+\Delta t})\} \tag{4.5}$$

The 34 different features listed in Table 4.1 are extracted from the kinematic chain, based on Kinect output. Since we only focus on the upper body motion, both left and right hands, wrist and elbow are selected out of the total of 20 joints. The features given by hand and wrist

from the same side may strongly correlated to each other, however, as the 3D position of the hands tracked by Kinect sensor is much more noisy than the wrist, both of them are used. In addition, there are three reference planes defined by certain joints in the kinematic chain as follows:

- $PlaneR = \psi_{921,\tau+\Delta t}$ and $PlaneL = \psi_{521,\tau+\Delta t}$ are the planes whose normal is the vector connecting the initial joint spine and terminal joint right and left shoulder respectively, and passing thought the center of the hips. Measuring these features with respect to the $PlaneR$ and $PlaneL$ represent the motion leftwards and rightwards, at the same time, the upwards and downwards motions are also encoded. (See Fig 4.2 (b))

- $PlaneFB = \phi_{951,t+\Delta t}$ is spanned by the left shoulder, right shoulder and the center of the hips. The distance and velocity with respect to the $PlaneFB$ represent the motion of upper body frontwards and backwards. (See Fig 4.2 (c))

Table 4.1: Pose features

No	Definition	Description
01	$Distance\{(\omega_{8,t}),(\omega_{8,t+\Delta t})$	Distance of left hand movement during Δt
02	$Distance\{(\omega_{7,t}),(\omega_{7,t+\Delta t})$	Distance of left wrist movement during Δt
03	$Distance\{(\omega_{6,t}),(\omega_{6,t+\Delta t})$	Distance of left elbow movement during Δt
04	$Distance\{(\omega_{5,t}),(\omega_{8,t+\Delta t})$	Displacement distance between left shoulder and left hand at time t and $t+\Delta t$ respectively
05	$Distance\{(\omega_{5,t}),(\omega_{7,t+\Delta t})$	Displacement distance between left shoulder and left wrist at time t and $t+\Delta t$ respectively
06	$Distance\{(\omega_{12,t}),(\omega_{12,t+\Delta t})$	Distance of right hand movement during Δt
07	$Distance\{(\omega_{11,t}),(\omega_{11,t+\Delta t})$	Distance of right wrist movement during Δt
08	$Distance\{(\omega_{10,t}),(\omega_{10,t+\Delta t})$	Distance of right elbow movement during Δt
09	$Distance\{(\omega_{9,t}),(\omega_{12,t+\Delta t})$	Displacement distance between right shoulder and right hand at time t and $t+\Delta t$ respectively
10	$Distance\{(\omega_{9,t}),(\omega_{11,t+\Delta t})$	Displacement distance between right shoulder and right wrist at time t and $t+\Delta t$ respectively
11	$Distance\{(\omega_{8,t}),(\phi_{951,t+\Delta t})\}$	Displacement distance from left hand at time t to the plane which spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
12	$Distance\{(\omega_{7,t}),(\phi_{951,t+\Delta t})\}$	Displacement distance from left wrist at time t to the plane which spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$

13	$Distance\{(\omega_{6,t}),(\phi_{951,t+\Delta t})\}$	Displacement distance from left elbow at time t to the plane which spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
14	$Distance\{(\omega_{12,t}),(\phi_{951,t+\Delta t})\}$	Displacement distance from right hand at time t to the plane which spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
15	$Distance\{(\omega_{11,t}),(\phi_{951,t+\Delta t})\}$	Displacement distance from right wrist at time t to the plane which spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
16	$Distance\{(\omega_{10,t}),(\phi_{951,t+\Delta t})\}$	Displacement distance from right elbow at time t to the plane which spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
17	$Distance\{(\omega_{8,t}),(\psi_{921,t+\Delta t})\}$	Displacement distance from left hand at time t to the plane with normal vector connecting right shoulder and spine, and passing through hip center at time $t+\Delta t$
18	$Distance\{(\omega_{7,t}),(\psi_{921,t+\Delta t})\}$	Displacement distance from left wrist at time t to the plane with normal vector connecting right shoulder and spine, and passing through hip center at time $t+\Delta t$
19	$Distance\{(\omega_{6,t}),(\psi_{921,t+\Delta t})\}$	Displacement distance from left elbow at time t to the plane with normal vector connecting right shoulder and spine, and passing through hip center at time $t+\Delta t$
20	$Distance\{(\omega_{12,t}),(\psi_{521,t+\Delta t})\}$	Displacement distance from right hand at time t to the plane with normal vector connecting left shoulder and spine, and passing through hip center at time $t+\Delta t$
21	$Distance\{(\omega_{11,t}),(\psi_{521,t+\Delta t})\}$	Displacement distance from right wrist at time t to the plane with normal vector connecting left shoulder and spine, and passing through hip center at time $t+\Delta t$
22	$Distance\{(\omega_{10,t}),(\psi_{521,t+\Delta t})\}$	Displacement distance from right elbow at time t to the plane with normal vector connecting left shoulder and spine, and passing through hip center at time $t+\Delta t$
23	$Velocity\{(\omega_{8,t}),(\pi_{92,t+\Delta t})\}$	Velocity of the left hand at time t along the direction of vector connecting right shoulder and spine at time $t+\Delta t$
24	$Velocity\{(\omega_{7,t}),(\pi_{92,t+\Delta t})\}$	Velocity of the left wrist at time t along the direction of vector connecting right shoulder and spine at time $t+\Delta t$
25	$Velocity\{(\omega_{6,t}),(\pi_{92,t+\Delta t})\}$	Velocity of the left wrist at time t along the direction of vector connecting right shoulder and spine at time $t+\Delta t$
26	$Velocity\{(\omega_{12,t}),(\pi_{52,t+\Delta t})\}$	Velocity of the right hand at time t along the direction of vector connecting left shoulder and spine at time $t+\Delta t$
27	$Velocity\{(\omega_{11,t}),(\pi_{52,t+\Delta t})\}$	Velocity of the right wrist at time t along the direction of vector connecting left shoulder and spine at time $t+\Delta t$
28	$Velocity\{(\omega_{10,t}),(\pi_{52,t+\Delta t})\}$	Velocity of the right wrist at time t along the direction of vector connecting left shoulder to spine at time $t+\Delta t$

29	$Velocity\{(\omega_{8,t}),(\pi_{951,t+\Delta t})\}$	Normal velocity of the left hand at time t along the direction of normal vector of the plane spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
30	$Velocity\{(\omega_{7,t}),(\pi_{951,t+\Delta t})\}$	Normal velocity of the left wrist at time t along the direction of normal vector of the plane spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
31	$Velocity\{(\omega_{6,t}),(\pi_{951,t+\Delta t})\}$	Normal velocity of the left elbow at time t along the direction of normal vector of the plane spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
32	$Velocity\{(\omega_{12,t}),(\pi_{951,t+\Delta t})\}$	Normal velocity of the right hand at time t along the direction of normal vector of the plane spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
33	$Velocity\{(\omega_{11,t}),(\pi_{951,t+\Delta t})\}$	Normal velocity of the right wrist at time t along the direction of normal vector of the plane spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$
34	$Velocity\{(\omega_{10,t}),(\pi_{951,t+\Delta t})\}$	Normal velocity of the right elbow at time t along the direction of normal vector of the plane spanned by left shoulder, right shoulder and hip center at time $t+\Delta t$

4.2 Head Orientation Estimation

The orientation of the head and its motion is a useful visual cue to identify different conversational scenarios, as nodding and shaking head are meaningful body language which are frequently used in conversational interactions. Currently the Kinect sensor lacks the ability to estimate the full orientation of individual joints, only their 3D positions. Hence, we perform head orientation estimation by extending the results from the facial feature extraction, which is detailed in the next section. As part of facial feature extraction, we obtain a set of five fiducial points for each face image: two external eye corners, two mouth corners, and nose tip. We follow the work by Gee and Cipolla [GC94] to estimate the head orientation from a single image using these fiducial points.

It is assumed that the ratio of the distance between mouth and nose and the distance between mouth and eyes is fixed, and the ratio of the height of nose and the distance between mouth and eyes is also fixed. It is also assumed that the eye corners, mouth corners and nose

base lie in a single plane, defined as the face plane. Following the notation described in [GC94], let L_f denote the distance between eyes and mouth, L_m the distance between mouth and nose base, L_n the height of nose that is measured as the distance from nose tip to the face plane. Thus, the two ratios are defined as: $R_m = L_m/L_f$ and $R_n = L_n/L_f$. The values for R_m and R_n are set as 0.6 and 0.4 respectively as suggested in [GC94]. R_m is used to locate the nose base in the image, measured along the symmetric axis of the face. Figure 4.3 illustrates these distances projected onto the image plane, denoted using lower case letters. The fiducial points can be used to directly estimate these distances in image units, which can then be used to estimate the head orientation. In addition, τ is the angle between the projected normal of the face and the x-axis of the image, θ is the angle between the projected normal and the symmetry axis of the face, and σ is the slant angle of the facial plane. The normal of the face plane (and therefore head orientation) is defined as $\overrightarrow{n} = [\sin\sigma\cos\tau, \sin\sigma\sin\tau, -\cos\sigma]$, where $\sigma = \cos^{-1}|d_Z|$ and d_Z is the component of the normal along the z axis, which is calculated as:

$$
\begin{aligned}
{d_Z}^2 &= \begin{cases} \frac{{R_n}^2}{m_1+{R_n}^2} & for \quad m_2 = 1 \\ \frac{t1+t2}{2(1-m_2){R_n}^2} & for \quad m_2 \neq 1 \end{cases} \\
t1 &= {R_n}^2 - m_1 - 2m_2{R_n}^2 \\
t2 &= \sqrt{(m_1 - {R_n}^2)^2 + 4m_1m_2{R_n}^2} \\
m_1 &= \left(\frac{l_n}{l_f}\right)^2 \quad and \quad m_2 = \cos^2\theta.
\end{aligned}
\tag{4.6}
$$

The derived head orientation is then considered as part of the 3D pose for the person.

4.3 Facial Feature Extraction*

The face images acquired have varied poses and sometimes contain occlusions (e.g., glasses and hand movement). In Figure 4.4 we show typical examples of occlusion, many face instances are self-occluded or occluded by the hand or glasses. Consequently, holistic models, such as active appearance models [CET01], have been found not robust enough to track the

*This work is mainly carried out by Dr. Hui Fang as part of a collaborative work

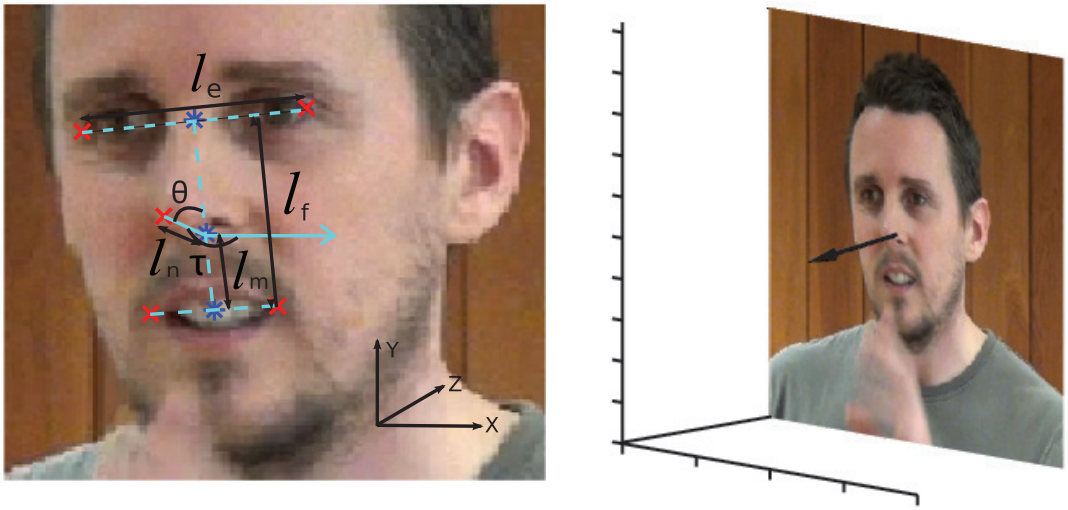

Figure 4.3: Left: the five fiducial points are overlaid on an example face image, and the required face plane and other measurements are derived from these localized fiducial points. Right: an illustration of the computed head orientation which is normal to the face plane.

faces beyond a few dozens of frames. Part-based models have been shown to perform better in localizing the fiducial points for head pose estimation and face tracking [LWSC10, TBA+09, CC08]. Therefore, we employ Constrained Local Models (CLM) since these have been demonstrated to be more robust under partial occlusion [CC08]. However, each of the conversational scenarios contains several minutes of image frames, which proved to be very challenging for even CLMs. Since our goal is to study the role of facial features in recognizing human interactions, we allow periodic reinitialization to ensure quality localization of fiducial points, face tracking, and consistency in extracted facial features.

At the bootstrapping stage, we define our region of interest, i.e., face patch, in the first frame of each sequence. SIFT features [Low04] in individual frames are detected, and it is assumed that majority of the SIFT features within the face patch in each sequence are reasonably stable. By matching the SIFT features following the first frame, the face patch locations in the subsequent frames are determined and their transformations relative to the first frame can then be computed. It is inevitable that some feature points can be occluded or missing due to, for example, facial movement. We use the RANSAC algorithm to minimize the impact of such

Figure 4.4: Examples face images that exhibt large pose variations, hand occlusion, and occlusion from glasses.

effects, as well as scale variations and on-plane rotations. These similarity transformations are then used to map all the faces across sequences to a common face patch so that feature extraction can be performed.

In the mapped common space, where face regions have been transformed as described above, face patches are then vectorized, followed by a principal component analysis (PCA). Lower dimensional feature descriptors are then obtained from the subspace of this PCA by eliminating principal components with small eigenvalues. The PCA coefficients extracted from the subspace are representative, as the reconstructed patch obtained by re-projecting the coefficients to the image space is usually similar to the original patch, while reducing the noise interference. Figure 4.5 provides some typical examples where dominant facial features are retained after PCA re-projection, whilst background and insignificant features are suppressed.

Notably, we do not explicitly extract shape representation as part of our facial features. This is because in conversational interactions, the lips are the most dynamic region and it is well understood that this is particularly problematic for shape localization, i.e., it is very likely that the extracted shape features are corrupted by inaccuracy due to poor localization of lip contour. The textural descriptors, however, implicitly encapsulate such dynamics in a form of changes in appearance in the mouth region. Furthermore, shape localization and subsequent shape descriptors are more likely to suffer from out-of-plane rotations.

Figure 4.5: Example frames projected onto the low dimensional PCA space followed by reconstration in the image domain.

In order to accurately localize the fiducial points, i.e. two external eye corners, two mouth corners and the nose tip, for head pose estimation, we build a constrained local model by applying a two-stage PCA. A training set consisting of randomly selected 20 images per sequence is used to build a shape and texture model. Manual labelling of those five fiducial points is carried out on the training set. Local patches of size 30×30 are extracted and vectorized to form the feature space for texture, and the location of those points across sequences are also vectorized to form the feature space for shape. Independent PCA is performed on these two feature spaces, and the reduced PCA coefficients are concatenated to form a combined feature for shape and texture. The mean location of the five points is readily available to be used as initialization. Note, all these steps are carried out in the mapped common face space derived at the feature extraction stage described above. The second-stage PCA is then performed on the combined feature space to produce the combined model.

Next, the remainder of the images are initialized with those five fiducial points using the

mean locations derived from the training set. Local patches from those regions are then extracted and combined with shape information, using the same method as described in the previous paragraph. These features are then projected to the combined subspace, and their coefficients are then re-projected back onto the image domain to generate local patch templates.

A cost function (Equation 4.7), is then used to update the positions of the five feature points. This is achieved by considering both the similarity of the texture in the regions around the updated locations to the generated templates and shape constraints.

$$f(s) = \omega \sum_{i=1}^{n=5} C_i(X_i, T_i) - \sum_{j=1}^{k} \frac{b_j^2}{\lambda_j}, \tag{4.7}$$

where $C_i(X_i, T_i)$ is the normalized correlation between a texture patch (e.g., nose or eye corner) and the generated template, b represents one of k shape coefficients, λ_j is the eigenvalue of the jth dimension and ω is a weight which controls the balance between texture and shape information. This processes is then iterated until it converges to a position which has similar texture generated from the model. The Simplex Downhill algorithm can be used to minimize the cost function.

4.4 Summary

In this chapter, we presented the methods used to extract features in order to represent the upper body pose motion, head orientation and the facial motion. The pose features were extracted as velocity and displacement measurements from Kinect output which can be noisy. The head orientations were estimated by tracking fiducial points on the face images and assuming that the face structure defined by those fiducial points only undergoes rigid transformations. Constrained local model was used to localise and track faces. Facial features were derived using principal component analysis of face image patches. It captures appearance changes due to both facial movement and face orientation changes head motion. In the following chapter, we employ GMM technique to analysis these features, pose features in particular.

Chapter 5

Data Analysis using GMM and Random Forest

Contents

To investigate the distribution of the features, especially for the pose features, Gaussian Mixture Model (GMM) is applied to generalise and cluster features. The Expectation and Maximisation (EM) algorithm is used to estimate the GMM parameters. Different approaches are employed to study the data distribution and their discriminative nature. Firstly, before applying statistical analysis, visual assessment is employed to inspect the correlation among the feature components. Secondly, one single mixture model is fitted to features obtained across all scenarios and all participants, which give us some insight of the general distribution of the feature points. We also look into the distribution between different scenarios, based on which we examine the correlation among them. Thirdly, we applied the GMM to each dimension of the features to generate "bag of words" type of higher level features, which are then used in Random Forest

based classification to distinguish different conversational scenarios.

5.1 Gaussian Mixture Model

Gaussian Mixture Model is a common and powerful method in parametrizing complex, often multi-modal distributions. A Gaussian mixture distribution can be formulated as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where K is the number of the components, π_k is the mixing coefficients and $\mathcal{N}$ denotes the normal distribution with mean μ_k and covariance Σ_k. The mixing coefficients π_k must satisfy the constrains $\sum_{k=1}^{K} \pi_k = 1$ and $0 < \pi_k < 1$.

Given a set of samples $X = \{x_1, x_2, ..., x_n, ...x_N\}$, the parameters of the GMM, π, μ and Σ are estimated by maximizing the *log* likelihood function given by:

$$ln\ p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \tag{5.1}$$

EM algorithm is the most popular algorithm for finding maximum likelihood solution to the Equation 5.1. More details can be found in [DLR77, MK08, Bis06].

5.2 Random Forest

Random Forest was first introduced by Leo Breiman [Bre01]. It is an ensemble classifier consisting of a set of decision trees, which significantly improves the generalisation ability of the classifier compared to single decision tree. At the Bootstrap aggregating stage (bagging), assuming that the data sample is independent and identically distributed, new training sets are generated by randomly sampling with replacement from the complete training set. For each new training set, one decision tree is constructed which consists of a set of split nodes and linking edges. Each non-leaf node stores a random test function which is applied to the input data, and leads to the leaf node. In the leaf nodes, the final predictor is stored. The gain of information is introduced to evaluate the performance of random test function, which can be

computed as:

$$I = H(S) - \sum_{i \in \{1,2\}} \frac{|S^i|}{|S|} H(S^i)$$

where the Shannon entropy is defined as: $H(S) = -\sum_{c \in C} p(c) log(p(c))$ in discrete case. In continuous case, for example, the entropy of Gaussian distribution with d multi-variable can be computed following the equation, $H(S) = \frac{1}{2} log\left((2\pi e)^d |\Lambda(S)|\right)$. In the prediction stage, all the trees classify the incoming data independently, the most voted class given by the trees is considered as the final classification of the forest.

5.3 Experiments Evaluation

The main purpose of this part of work is to investigate the distribution of features using clustering and tree based classification techniques. The pose features are extracted following the approach described in Chapter 4 with parameter $\Delta\tau = 30 frames$. The features extracted from different pairs, participants and conversation types are denotes as a subset of complete feature space:$\chi_{i,j,k,d}$ where $i \in \{1,2,3,4,5\}$ and $j \in \{1,2\}$ correspond to different pairs and participants within the pair, respectively, $k \in \{1,2,3,4,5,6,7\}$ indicates the specified conversation type, the order of which is defined in Table 3.1, and d is a subset of $D = \{1,2,...,34\}$ which represents the dimension of the feature space. Three different approaches are applied to the pose features as follows. Firstly, we visually evaluate the features correlation. Secondly, the features are clustered using GMM algorithm, and the histograms of clusters are constructed to examine the distributions of the features. Finally, the random forest is applied to classify different conversation types based on the high level features using histograms of clusters.

5.3.1 Visual Assessment

We randomly sample a subset from the feature space $\sum_{i=1}^{5} \sum_{k=1}^{2} \sum_{k=1}^{7} \chi_{i,j,k,d}$, and projected onto 2D spaces by arbitrary selecting two dimensions of the features. Some of the visualization results are shown in Figs 5.1 and 5.2. Different colours indicate different pairs, where $\circ$ and $*$ represent the participant 1 and participant 2 within the pairs. Fig 5.1 shows that the features with the same type extracted from geometrically adjacent joints are closely correlated. Taking

the hand and twist for example, compared to the motion of the forearm, the movement of hand to the wrist is subtle. The former which involves simultaneous movement of both hand and wrist leads to the correlation on the feature components. This suggests that clustering in the feature space is likely to be useful.

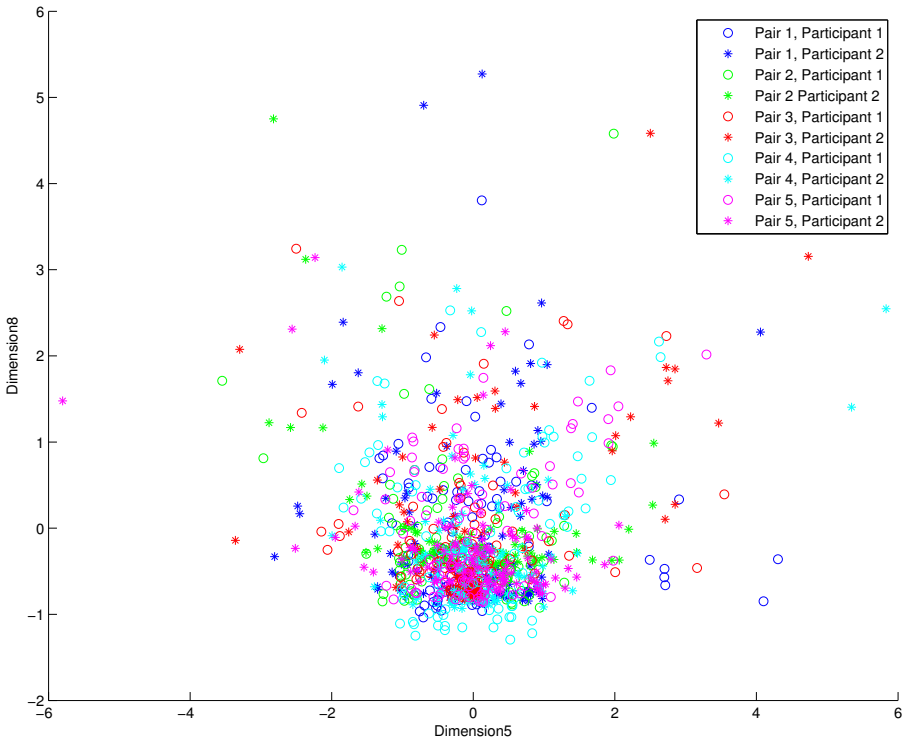

Figure 5.1: The distribution of the samples, projected on the Dimension5 and Dimension8 , which are the displacement distance between left shoulder and left wrist, and the distance of right elbow movement over time.

5.3.2 Clustering

The EM algorithm is applied to the complete features space $\sum_{i=1}^{5}\sum_{k=1}^{2}\sum_{k=1}^{7}\chi_{i,j,k,d}, d = D$ to fit a GMM with 10 Gaussian components, each of which correspond to one cluster. Based on the model, the clusters of each sample are then obtained. For each pair, we construct the histograms of clusters by scenarios (See Figs 5.3, 5.4, 5.5, 5.6, 5.7). Apart from the histograms

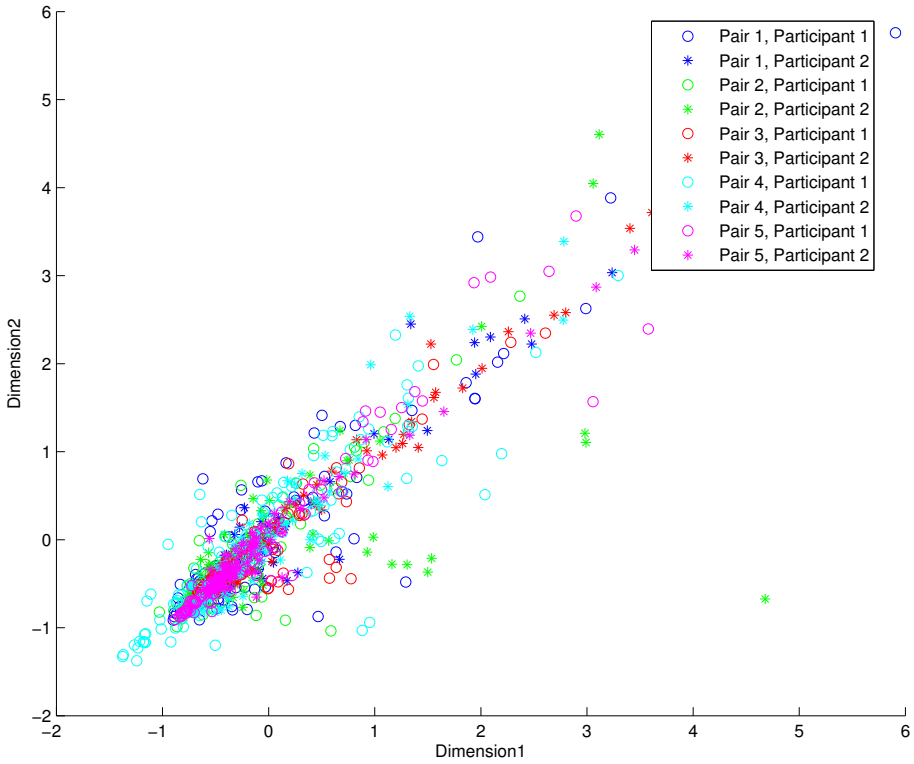

Figure 5.2: The distribution of the samples, projected on the Dimension1 and Dimension2, which are the distance of left hand and wrist movement over time.

of "Pair4, Participant 1" seems similar among different scenarios, which may to some extent, imply that the identity information is somewhat encoded in the statistical descriptors, in general the features are distributed randomly across different participants and different conversational scenarios.

The histograms for each different scenarios are constructed as well, which are obtained by summing the histograms with the same conversation type across different pairs that we have computed previously (see Fig 5.8). Then the cross correlation between 7 conversation types is computed and listed in Table 5.1, which shows the similarities among those scenario histograms. From the cross correlation matrix, it is necessary to mention that "Describing Work"

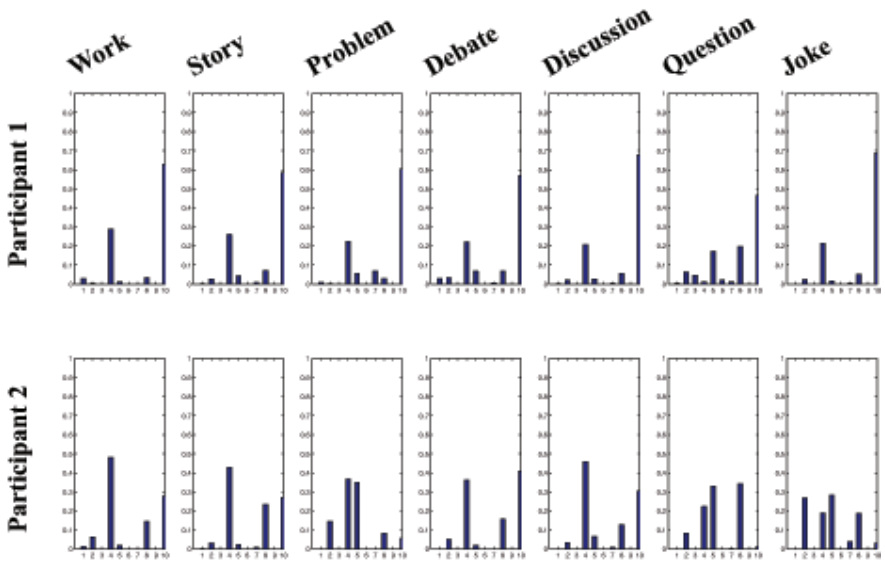

Figure 5.3: The histograms of clusters of pair 1

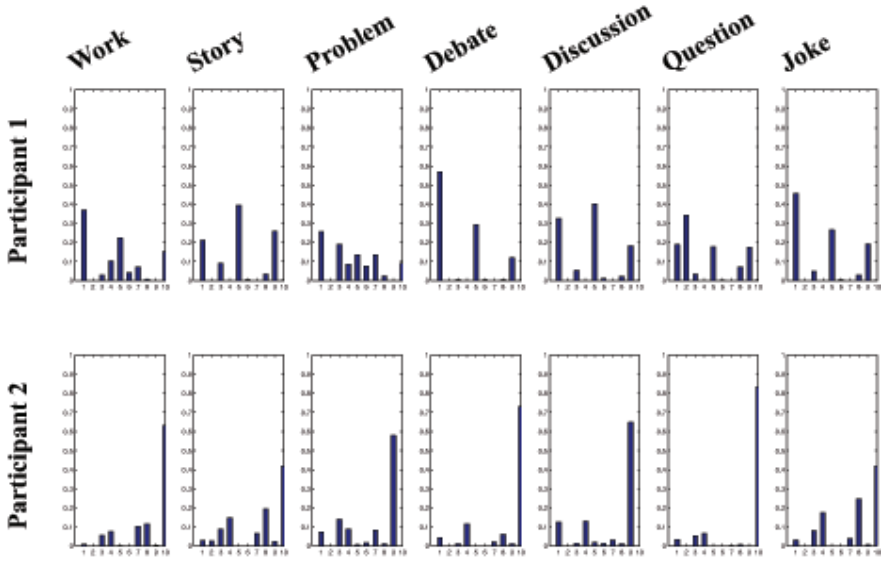

Figure 5.4: The histograms of clusters of pair 2

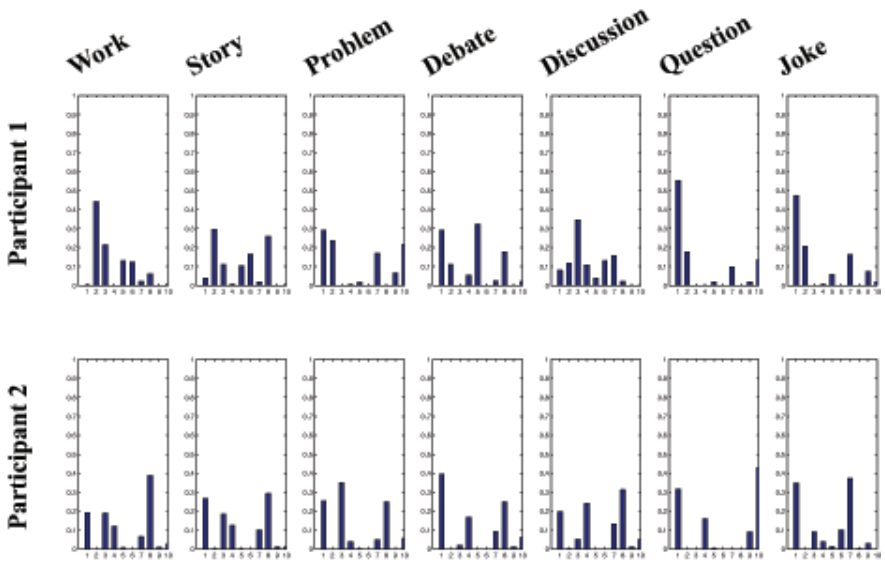

Figure 5.5: The histograms of clusters of pair 3

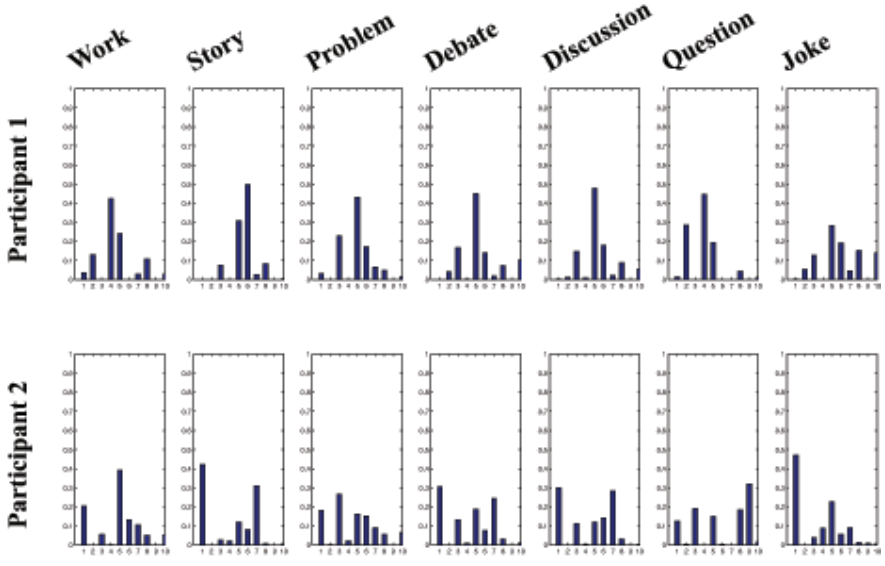

Figure 5.6: The histograms of clusters of pair 4

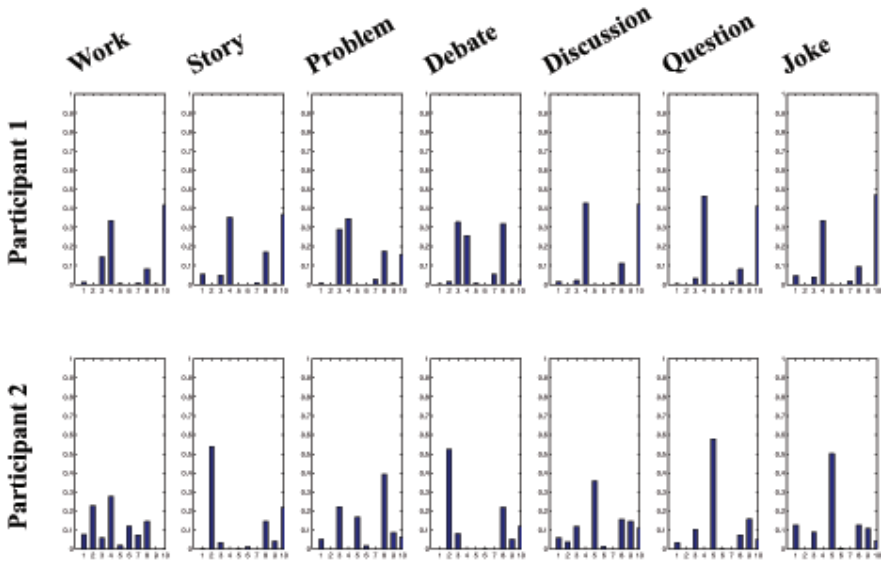

Figure 5.7: The histograms of clusters of pair 5

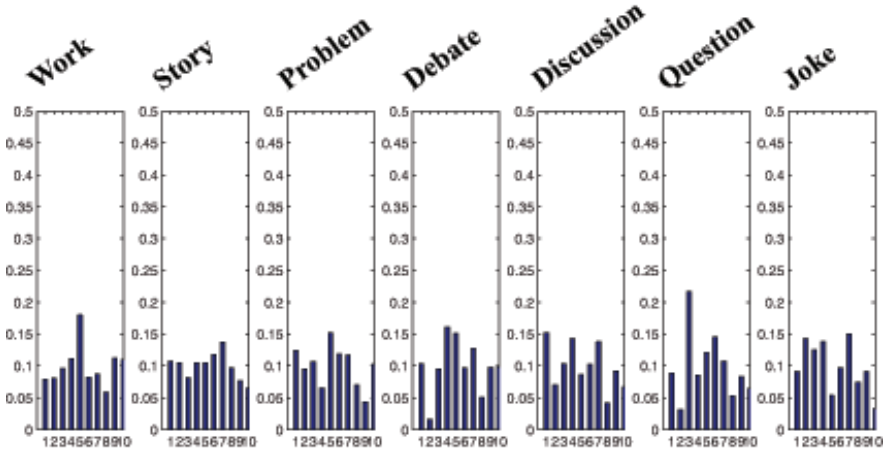

Figure 5.8: The histograms of clusters of different conversation types

Table 5.1: Cross correlation between the histograms of clusters

	Work	Story	Problem	Debate	Discussion	Question	Joke
Work	1	-0.164	0.202	0.353	0.634	-0.367	0.035
Story	-0.164	1	0.033	0.403	0.179	0.550	0.487
Problem	0.202	0.033	1	0.369	0.385	0.133	0.313
Debate	0.353	0.403	0.369	1	0.243	0.189	0.194
Discussion	0.634	0.179	0.385	0.243	1	0.069	0.617
Question	-0.367	0.550	0.133	0.189	0.069	1	0.493
Joke	0.035	0.487	0.313	0.194	0.617	0.493	1

is strongly correlate with “Discussion” and uncorrelated with “Story telling” and “Subjective question”. Other interesting finding is that “Discussion” and “Joke” are strongly correlate as well. These suggest that analysing the features globally is unlikely to distinguish different scenarios.

5.3.3 Classification using Random Forest

We also applied the clustering to each dimension of the features $\sum_{i=1}^{5}\sum_{k=1}^{2}\sum_{k=1}^{7}\chi_{i,j,k,d}, \forall d \in D$ independently, then 34 GMMs were fitted. Sliding windows with fixed step width $\Delta S = 5 frames$ and fixed window size $S = 100$, is introduced. Within the window the histograms of clusters for each dimension of features can be constructed, and concatenated into one vector η_w with 34×10 components. Repeatedly, for each conversation types, the η_w are computed using sliding window across all participants, based on which the random forest classifier is learned and used to classify conversation types. 500 decision trees were built on 4 pairs to form the forest classifier, and 1 pair is left out as testing set. The training performance including out of bag (OOB) error rate and mean decrease in accuracy and Gini index are shown in Fig 5.9. We performed classification on the testing set, and the confusion matrix is listed in Table 5.2, the average recognition rate of “Pair4, P1” and “P2” are 32.9% and 30.5%. It is obvious that “Story telling”, “Problem solving” and “Subjective question” are not discriminative in terms of the descriptors using the histograms of GMM clusters. The possible explanation is that the temporal information during the communications which plays important role in identifying conversation type was neglected and not encoded in the descriptors.

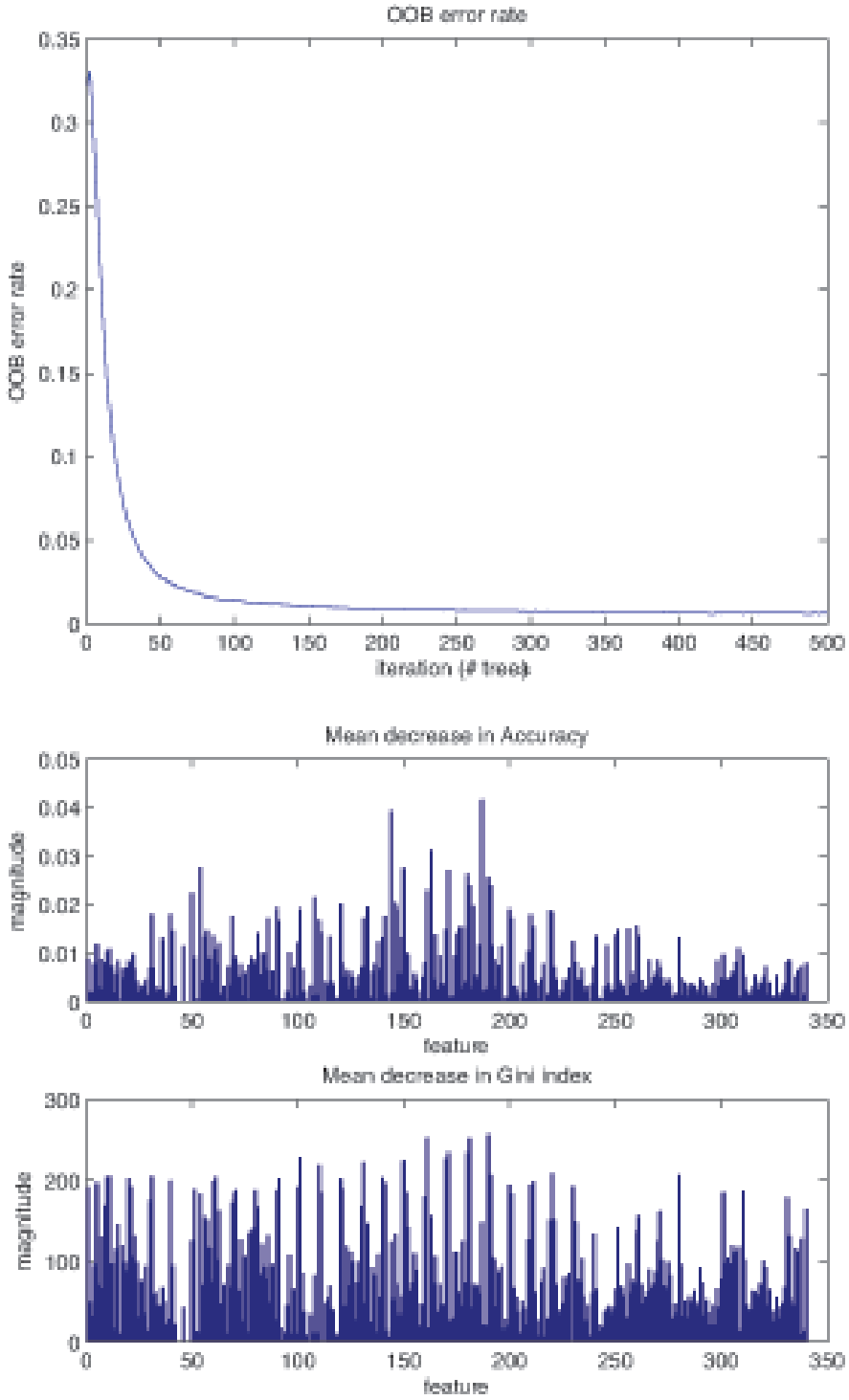

Figure 5.9: Performance of training using random forest

Table 5.2: Confusion Matrix based on pose descriptors

		Work	Story	Problem	Debate	Discussion	Question	Joke
Pair 4, P1	Work	0.359	0.072	0.122	0.172	0.085	0.064	0.127
	Story	0.418	0.090	0.111	0.209	0.051	0.049	0.072
	Problem	0.159	0.032	0.157	0.087	0.287	0.054	0.224
	Debate	0.297	0.066	0.102	0.398	0.027	0.057	0.052
	Discussion	0.001	0.000	0.005	0.000	0.455	0.005	0.534
	Question	0.282	0.035	0.167	0.100	0.119	0.139	0.154
	Joke	0.001	0.000	0.014	0.000	0.270	0.009	0.705

Average = 0.329

Pair 4, P2	Work	0.304	0.072	0.090	0.387	0.040	0.041	0.062
	Story	0.181	0.060	0.078	0.560	0.043	0.031	0.043
	Problem	0.330	0.078	0.068	0.422	0.031	0.031	0.037
	Debate	0.234	0.076	0.069	0.564	0.007	0.031	0.014
	Discussion	0.000	0.000	0.033	0.000	0.646	0.001	0.319
	Question	0.360	0.061	0.157	0.095	0.104	0.073	0.157
	Joke	0.002	0.001	0.219	0.000	0.348	0.010	0.418

Average = 0.305

5.4 Summary

We employ GMM clustering and supervised classification techniques to explore the distribution of the pose features. The pose features were described in Chapter 4. Then the GMMs were constructed by applying parameter estimation to the whole features space, also to each dimension of the features independently. Based on clustering results using GMM and classification results provided by random forest, it is obvious that the differences between these types of conversation are extremely subtle and temporal dynamics need to be taken into account. In the next chapter, the single HMM with one Markov chain and coupled HMM with two conditional dependent Markov chain are used to model the interaction and classify different conversational types.

Chapter 6

Interaction Modeling using HMM

Contents

To classify each observed conversation we conduct experiments using both Hidden Markov Models (HMM) and Coupled Hidden Markov Models (CHMM) following [ORP00]. The benefit of using these models is that they are suitable for classifying sequences with different temporal lengths. We briefly describe the HMMs used in this work though refer the reader to [Rab90] for a full description. A first order Hidden Markov Model represents a directed graph composed of a discrete set of t hidden nodes, each of which is connected to an observed node. The hidden nodes represent random variables that cannot be directly observed, but can be inferred from both the observed node connected to it and information passed from it's neighbouring hidden nodes. Each hidden node is constrained to be one of a fixed set of m discrete states and each observed node z is modelled over a continuous n-dimensional space, $z \in R^n$.

6.1 Hidden Markov Model

A HMM is defined by three components $\Lambda = \{\pi, \mathbf{A}, \mathbf{B}\}$, a prior distribution π, an m by m state transitional matrix $\mathbf{A}$ and a likelihood function $\mathbf{B}$. The graphical model representing a HMM can be seen in Figure 6.1 (a). The likelihood function represents the edges between the observed and hidden nodes, representing the distribution $p(z_t|y_t)$. The state transitional matrix represents an edge between connected hidden nodes $p(y_{t+1}|y_t)$. The prior represents the initial distribution over the states in the first time instance $p(y_1)$.

The likelihood function is calculated by using a different model for each state, where each model represents the probability distribution $p(z_t|y_t = n)$, where n represents the class id. This distribution is approximated using a Gaussian Mixture Model, thus $p(z_t|y_t = n) = \sum_{k=1}^{K} \lambda_n^k \mathcal{N}(z_t; \mu_n^k, \Sigma_n^k)$, where K is the number of components in the model, and λ_n^k, μ_n^k and Σ_n^k represent the kth component's weight, mean and covariance respectively. All model parameters can be learnt automatically using the Expectation Maximization algorithm and a set of training data. Each feature used in the training set is covariance normalized so as to balance the influence of each when training the GMMs.

A separate HMM is learnt for each of the seven classes, $\{\Lambda_1, .., \Lambda_7\}$. Given a set of T observations $Z_T = \{z_T, z_{T-1}, .., z_1\}$ from an unknown class we classify it to the model that maximizes the distribution $p(\Lambda_n|Z_T)$, where n represents the class id. This is calculated in two stages, firstly the forward-backward algorithm is used to calculate the distribution $p(Z_T|\Lambda_n)$. This is achieved by recursively calculating

$$p(y_t = j|z_{t-1}, .., z_1, \Lambda_n) = \sum_{i=1}^{m} \mathbf{A}_{ij} p(z_{t-1}|y_{t-1} = i) p(y_{t-1} = i|z_{t-2}, .., z_1, \Lambda_n) \tag{6.1}$$

and then summing the probabilities over all states in the final time instance, i.e.

$$p(Z_T|\Lambda_n) = \sum_{i=1}^{m} p(z_T|y_T = i) p(y_T = i|z_{T-1}, .., z_1, \Lambda_n)$$

Following which the distribution $p(\Lambda_n|Z_T)$ can be calculated using Bayes' law assuming a flat prior across all classes.

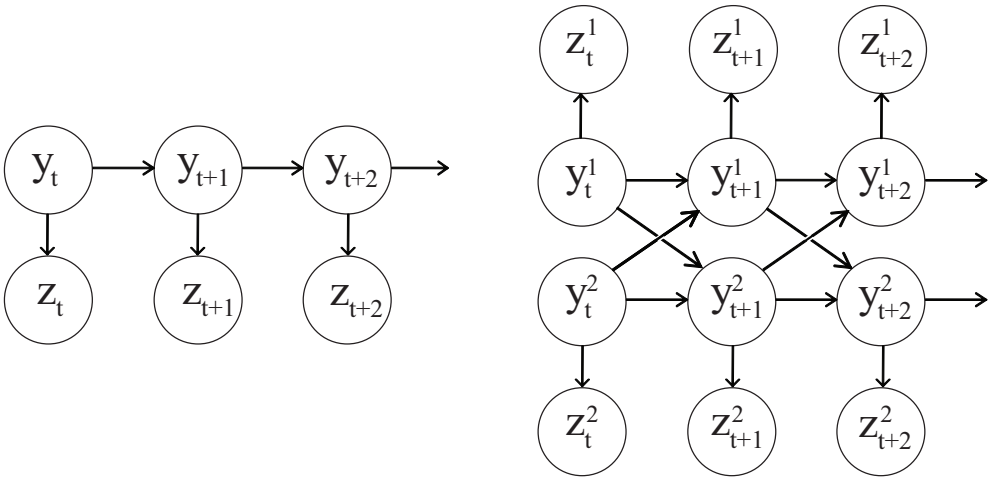

Figure 6.1: Graphical model used to represent single HMM (a) and coupled HMM (b). The y represent the hidden nodes and the z represents the observed nodes.

The above model describes a single random process, however, an obvious question is how to proceed when two different observations can be made at each time frame corresponding to a different person, $\{z_t^1, z_t^2\}$. A naive approach is to concatenate the features, $z_t^{12} = (z_t^1, z_t^2)$, and use the model described previously. However, one of the principal limitations with this approach is that the dimensionality of the state space over which the likelihood function must be learnt would double. This would mean that a much larger amount of training data would be required, though this could implicitly capture correlations between the individuals through the likelihood function. However, an alternative is to explicitly model the dependence between the two subjects by using a HMM to represent each person and then adding an edge between the subjects across time. This is represented in Figure 6.1 (b). This additional edge is implemented through the transitional matrix, A, however, assuming both models have the same number of states, now A is an m by m by m cube and represents the distribution $p(y_t^1 | y_{t-1}^1, y_{t-1}^2)$. However, now an extra transitional matrix is required to represent the distribution $p(y_t^2 | y_{t-1}^2, y_{t-1}^1)$. Again these can be learnt using the EM-algorithm and the model evaluated using a forward backward algorithm, performing an extra summation over the states of the coupled chain. The benefit of

this is that learning the likelihood function can still be performed in R^n and instead the added complexity of the model is appended to the transition matrix. Provided the number of possible states is not too large, this is a preferable arrangement.

It would be expected that the Coupled HMM can better capture the interactions between two people, hence, we use this to examine the effect of the interaction on classifying conversational interactions.

6.2 Experiments Evaluation

The approach described in Chapter 3 was used to collect the data set used in the presented experiments. In total all tasks were completed by 5 different pairs of people. A sample of the data collected for each conversational interaction is presented in Figure 3.2 in Chapter 3. As can be seen each class is not obviously distinct from the others and though there are some representative poses of each class, it would be extremely difficult to determine the class using only pose from a single frame. Another observation is that often as people were standing they opted to place their hands in their pockets, whist initially this raised some concern, we observed that when they started to talk in general the hands would be removed from the pockets and actually this provides a cue as to who is talking. However, one of the biggest challenges of the data set is the sheer variation in the types of motion, gestures and expressions performed by each participant during the task, even the neutral pose of each participant as they are listening is very different.

The PCA patch model, used to represent the face, was learnt using seven hundred randomly selected frames of data from the camera observing the face across all participants. Learning a model from this data was particularly difficult since the faces were generally not fronto-parallel. The resulting PCA patch model was 20 dimensional. Observed image patches of the face were projected through this PCA model to use as both training features and for classification. Pose features were extracted from the Kinect data as outlined in chapter 4 with $\Delta t = 1.0s$. To reduce the dimensionality of the pose features a PCA model was also learnt and used to project them

to just 20 dimensions so that they had the same dimensionality as the face features. Examples of extracted 3D pose and tracked facial features are shown in Figure 6.2.

To train and test the classifiers each recorded sequence was split into 20 second sections. Each section was labeled as the task from which it was extracted and used as a single example, both for training and testing. Whilst HMMs are well suited to classifying sequences of different lengths we did not want to inadvertently introduce any biases into the results as because of differing temporal lengths. The frame rate of the features used was down sampled to $2.5fps$, so that each 20 second sequence was composed of observations from 50 time instances. In all experiments 10 fold cross validation was used to test each method. All HMMs had three hidden states and the observational likelihood function for each state was represented by a three component GMM.

Initially we experimented using pose features from only a single person, this is to understand how much information can be extracted by observing one participant. For this a standard HMM was used. The normalized log likelihood for each classifier is depicted in Figure 6.3. As can be seen in the initial frames there is uncertainty as to which class of interaction is being observed. However, as more observations become available the model is able to recover the correct class. This demonstrates the large temporal window that is needed to reliably estimate which scenario is being observed. This is different to previous approaches where actions are recognized using very short temporal windows since they are far more easy to discriminate. The confusion matrix for this experiment is presented in the top row of Table 6.1, where an average of 0.51 is achieved. It can be seen that telling jokes is the most difficult task to classify and that describing work the easiest.

For the next experiment we still used pose features, however, we experimented with combining features from both participants. We compared two approaches, the first used a standard HMM and naively concatenated the features together and the second treated them separately and train a Coupled Hidden Markov Model (CHMM). The confusion matrix from each experiment are labelled as “Concat. HMMs” and “Coupled HMMs” respectively in Table 6.1. The accuracy of both schemes are significantly higher than using observations from a single person.

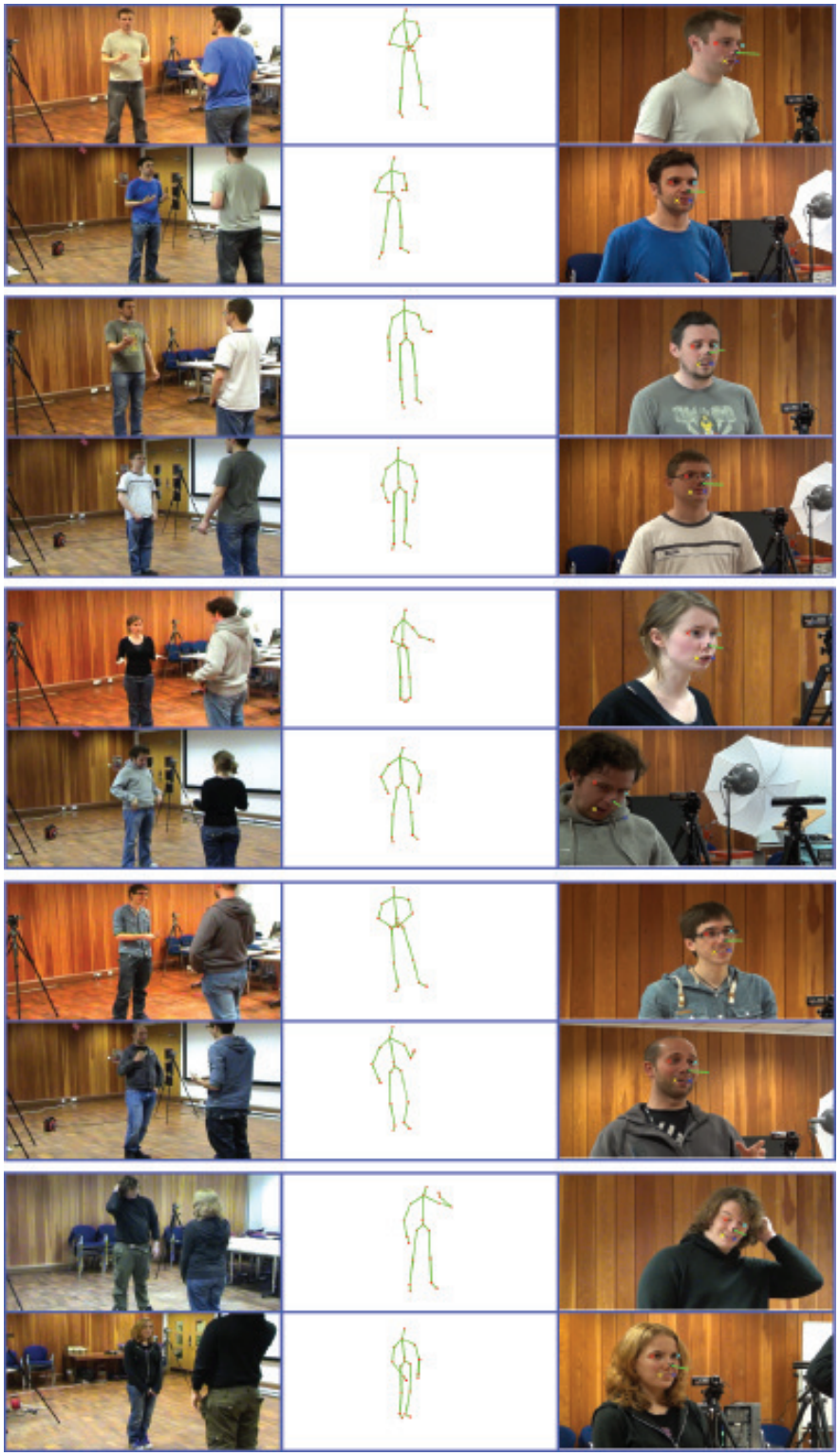

Figure 6.2: Examples of pose estimated using the Kinect and tracked facial features.

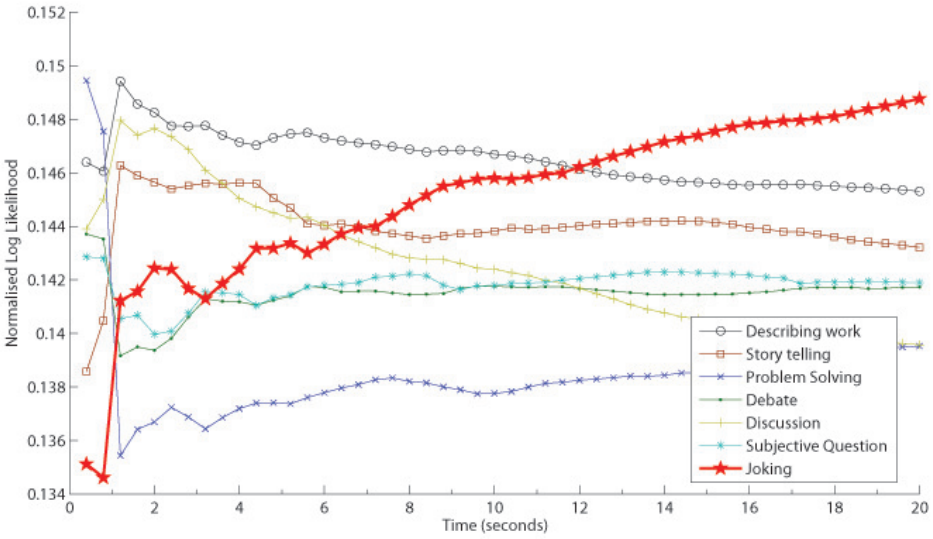

Figure 6.3: Visualization of the relative classifier likelihoods for each scenario as a function of time. The correct class is joking.

This highlights the benefit of having multiple streams of information when observing people during an interaction as they can be used to better discriminate the task being performed. Interestingly, a drop in the True Positive (TP) rate can be observed when using a HMM to model two people compared to just one person for "problem solving". This suggests when observing two people the class becomes less distinguishable and more easily confused with the "Describing Work" task and "Debate Task", though we note that the performance is still well above chance.

The CHMM significantly outperformed the HMM, where we see consistently higher TP rates for all classes compared to using observations from just a single participant. We also note that although there is still confusion between "problem solving" and "Work" for the CHMM, the confusion between "problem solving" and "discussion" is significantly reduced as are many of the false positives for many other classes compared to the classifiers using only a HMM. This is likely to be because of a CHMM, where the interaction is modelled between the hidden states, is better suited to modelling interactions compared to a standard HMM where the inter-

Table 6.1: Confusion Matrix based on pose descriptors

		Work	Story	Problem	Debate	Discussion	Question	Joke
Single HMMs	Work	0.627	0.104	0.059	0.071	0.093	0.008	0.038
	Story	0.125	0.599	0.026	0.112	0.098	0.013	0.026
	Problem	0.170	0.070	0.410	0.103	0.120	0.057	0.070
	Debate	0.114	0.107	0.008	0.516	0.180	0.000	0.074
	Discussion	0.135	0.115	0.058	0.121	0.455	0.033	0.084
	Question	0.100	0.080	0.017	0.040	0.093	0.593	0.076
	Joke	0.057	0.214	0.029	0.100	0.214	0.014	0.371

Average = 0.510

Concat. HMMs	Work	0.895	0.046	0.000	0.023	0.030	0.000	0.007
	Story	0.134	0.718	0.000	0.070	0.064	0.000	0.014
	Problem	0.267	0.067	0.250	0.233	0.133	0.000	0.050
	Debate	0.112	0.000	0.000	0.660	0.229	0.000	0.013
	Discussion	0.050	0.000	0.000	0.104	0.834	0.000	0.033
	Question	0.250	0.000	0.000	0.083	0.100	0.533	0.033
	Joke	0.192	0.217	0.000	0.083	0.108	0.000	0.400

Average = 0.613

Coupled HMMs	Work	0.837	0.126	0.007	0.008	0.015	0.000	0.007
	Story	0.089	0.804	0.000	0.029	0.038	0.000	0.041
	Problem	0.217	0.033	0.583	0.100	0.067	0.000	0.000
	Debate	0.033	0.033	0.000	0.743	0.131	0.000	0.060
	Discussion	0.013	0.050	0.000	0.064	0.873	0.000	0.000
	Question	0.033	0.117	0.067	0.050	0.067	0.633	0.033
	Joke	0.133	0.058	0.000	0.050	0.108	0.000	0.650

Average = 0.732

action is modelled in the state space of the observed node. The CHMM at the observational node level, can treat each observation independently allowing the model to better represent the possible combinations of interchanging states.

Next we investigated using only face descriptors. Again we experimented using a standard HMM with just the observations obtained from one participant, a HMM with the features from both participants concatenated together and finally a CHMM. The average for each approach is 0.561, 0.667 and 0.722 respectively and we see again that combining the features from both subjects significantly improves performance. We also observe that the CHMM outperforms

Table 6.2: Confusion Matrix based on face descriptors

		Work	Story	Problem	Debate	Discussion	Question	Joke
Single HMMs	Work	0.623	0.168	0.011	0.093	0.068	0.015	0.022
	Story	0.072	0.638	0.013	0.092	0.100	0.000	0.086
	Problem	0.050	0.070	0.430	0.120	0.120	0.120	0.090
	Debate	0.049	0.114	0.033	0.638	0.140	0.000	0.025
	Discussion	0.050	0.101	0.033	0.199	0.534	0.006	0.076
	Question	0.020	0.107	0.140	0.000	0.040	0.577	0.117
	Joke	0.029	0.129	0.086	0.029	0.157	0.086	0.486
								Average = 0.561
Concat. HMMs	Work	0.909	0.076	0.000	0.007	0.007	0.000	0.000
	Story	0.052	0.893	0.000	0.029	0.013	0.000	0.014
	Problem	0.133	0.067	0.350	0.067	0.217	0.067	0.100
	Debate	0.000	0.143	0.000	0.724	0.133	0.000	0.000
	Discussion	0.000	0.038	0.013	0.079	0.846	0.000	0.025
	Question	0.033	0.133	0.150	0.050	0.033	0.500	0.100
	Joke	0.033	0.292	0.117	0.025	0.058	0.025	0.450
								Average = 0.667
Coupled HMMs	Work	0.890	0.087	0.000	0.015	0.007	0.000	0.000
	Story	0.013	0.904	0.000	0.055	0.014	0.000	0.014
	Problem	0.033	0.067	0.517	0.050	0.200	0.100	0.033
	Debate	0.000	0.129	0.000	0.745	0.126	0.000	0.000
	Discussion	0.013	0.038	0.000	0.091	0.859	0.000	0.000
	Question	0.000	0.067	0.150	0.033	0.033	0.583	0.133
	Joke	0.000	0.258	0.050	0.000	0.108	0.025	0.558
								Average = 0.722

the HMM using the concatenated features. Compared to the averages for pose, 0.510, 0.613 and 0.732, the face features are more discriminative than pose using both a single observation and concatenating the features. However, for the CHMM the face and pose features achieve similar accuracy. The confusion matrices for each model are shown in Table 6.2. Interestingly, the TP-rate for the concatenated features is lower than that of using just a single participant for "problem solving" as observed when using pose features. The most improved classes are Work, Story and Discussion.

Finally, we examined coupling together pose and face features. This was achieved by

Table 6.3: Confusion Matrix based on pose and face descriptors

Coupled HMMs	Work	Story	Problem	Debate	Discussion	Question	Joke
Work	0.895	0.084	0.000	0.007	0.015	0.000	0.000
Story	0.041	0.848	0.000	0.055	0.027	0.000	0.029
Problem	0.067	0.000	0.567	0.000	0.233	0.033	0.100
Debate	0.000	0.083	0.000	0.807	0.110	0.000	0.000
Discussion	0.013	0.025	0.000	0.089	0.873	0.000	0.000
Question	0.000	0.033	0.033	0.000	0.033	0.683	0.217
Joke	0.000	0.183	0.058	0.000	0.117	0.058	0.583

Average = 0.751

concatenating the pose and face features together for each subject and then using a coupled HMM to represent the interaction between the participants. The confusion matrix for this experiment is presented in Table 6.3 where we see the average is slightly improved over using coupled face features alone. Though we note that for those interaction classes with TP-rate less than 0.8 when using coupled face features, the average improvement is 0.059 when using face combined with pose features, suggesting that the margin of improvement could be greater than the results suggest. This is since it's likely that for the other classes, the TP-rate maybe approaching its ceiling as it is unlikely that all examples can be correctly classified using just a $20s$ temporal window.

Whilst the Kinect sensor permits direct estimation of 3D pose that is currently more robust and accurate than RGB camera methods, the accuracy of the data collected still contains some noise, as does the face features used in this work. However, despite this we have shown that recognition of conversational interactions can still be achieved. However, we note that in regards to the comparison of facial versus upper-body features, a different set of features could achieve different comparative levels of accuracy. So it is difficult to conclusively measure the relative importance of face versus upper body cues during conversational interactions.

6.3 Summary

In this chapter, we have examined whether 3D pose extracted from a consumer device can be used to recognize different conversational interactions using generative model, HMM. We have also investigate facial dynamics in recognizing different types of conversation. Two major findings can be summarized based on the experimental results. Firstly, human interaction, as a key factor of communications, can be exploited, and it simultaneously is a promising way to understanding high-level semantic events. Secondly, it is found that combining temporal features from both pose and face brings further benefits to the task.

Chapter 7

Conclusions and Future Work

Contents

7.1 Conclusions

In this thesis, we presented a comprehensive study on gesture and facial cues in understanding human interaction, in particular conversational interactions. The data sets used in the experiment were carefully collected. Bodily movements were captured by two Kinect sensors, and face features were extracted from face tracking and localisation in video footage. Head pose orientation was also computed based on face tracking. Sequences including video and Kinect output were manually synchronised, and necessary data pre-processing such as providing the ground-truth of fiducial points were carried out. The difference among the seven conversational scenarios are very subtle, and the primitive actions and interactions are commonly exhibited across different scenarios. Rather than trying to recognize individual primitive action or interaction, we attempted to recognise different types of conversation categories given by topic.

We firstly employed GMM to analyse the distributions and correlations of 3D pose features among different interaction scenarios. At the same time, a discriminative based on random

forest was applied to explore the global differences among seven conversational scenarios. The experiments suggested that, in terms of the distribution of the features, the differences between these types of conversation were extremely subtle and the global features in pose was not discriminative enough on its own. The temporal dynamics are important in identifying the differences in these conversational interactions.

Inspired by the results of feature analysis using GMM and random forest, we applied both HMM and CHMM to classify the conversational interactions. The experimental results suggested that although it was possible to recognise conversational interactions simply based on a single subject, the recognition rates were significantly better when taking into account interactions. Both the 3D pose cue and the face cue provided promising and similar recognition rates. The combination of the two under the current framework, however, showed some but limited benefit. We suspect this may due to the subtlety in the conversational interactions, although a more sophisticated integrating model may improve the results. The experimental results also showed that the CHMM outperformed HMM.

Overall, this work provides a different perspective in analysing human interactions. Instead of focusing on the effort of extracting and classifying individual micro actions or interactions, we put an emphasis on classifying interactions that consist of similar micro actions and interactions. The established dataset is also valuable for the researchers in this area.

7.2 Future work

We have shown in this thesis that it is possible to classify different conversation types using pose and facial features, and we believe this to be the first work devoted to conversational interaction modelling. The followings are worth further investigation.

Currently, five pairs of participants data were collected. This provided us a large data set, in terms of number of image frames and Kinect sequences, and it was very time-consuming to prepare the data set. However, in order to further extract the action and interaction patterns across subjects, it is necessary to collect even more participant data, i.e., in order to more

effectively generalise universal interaction patterns. Given sufficiently large amount of data, it is then possible or more realistic to carry out more challenging recognition or classification task, such as leave-one-pair-out. The conversation types investigated in this work have subtle differences among them. However, if more participants can be recruited, more conversation scenarios can then be considered, which will further increase its scientific value.

Secondly, other types of feature can be examined. For example, the spatio-temporal analysis using wavelets can be applied to extract features. Its coefficients may be used to construct feature descriptors in order to capture temporal dynamics efficiently. It is also worth employing state-of-the-art feature selection algorithms to further improve its performance. Combining face features and pose features has been found useful in recognition. More work can be carried out by investigating the role of face geometrical dynamics. Also, more advanced face tracking and localisation techniques will reduce or eliminate the manual involvement and improve overall performance, e.g. resulting more stable head pose orientation estimation.

Thirdly, it is also worth considering to explicitly extracting micro action and interactions before modelling conversational interactions as a whole. This may be equally revealing in terms of understanding the dynamics in human interactions.

Finally, it will be useful and scientifically interesting to carry out interaction synthesis. This can provide additional potential applications, for example simulating natural human behaviour in humanoid robotics.

Appendix A

Participants Consent Form

Swansea University – Computer Science Department
Research Consent Form

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

Research Project Title
Understanding Human Emotion

Researchers
Dr. X. Xie
Dr. P. W. Grant
Dr. H. Fang
Dr. B. Daubney

Experiment Purpose

This work addresses two issues: (1) Using computing techniques to extract, generalise and synthesise human facial expression and bodily movement; (2) Understanding the role of face expression and body gesture in human emotion.

The experiment involves video recording pairs of participants conversing to each other in a set of given scenarios, including telling jokes, explaining new concepts, and debating and discussing. Multiple camcorders and a pair of Microsoft Kinect sensors are used for data capture. The materials used, such as jokes and debate topics, will be given to you before the video recording. None of those materials is intended to cause any personal stress. However, if you feel in any shape or form uncomfortable with any of the material, you have every right not to participate the experiment at any stage, including before and during video recording, without giving us any reason.

Participant Recruitment and Selection
Staff personnel, Undergraduate and graduate students from Swansea University as well as participants from the public are being recruited for this experiment.

Procedure
This session will require about 30-40 minutes of your time. You will be asked to discuss some topics with each other, some of which are listed below

(i) Telling jokes
You will be given some joke material so that you will tell such jokes to each other during the experiment.

(ii) Explain a concept
You will be asked to explain to each other your own research or work.

(iii) Debate and discussion
You will be given a list of topics you will discuss with your partner.

(iv) Telling stories

*A copy of this consent form has been given to you to keep for your records and reference.

We would like you to tell an interesting story, which may from your own experience, to your partner. Your partner will then tell you an interesting story.

None of the above is a test – our objective is to record facial dynamics and your body gesture to be added into a database for processing.

Data Collection

Your facial expression and body gesture will be recorded by camcorders and Kinect sensors and added into a database. The recorded audio on the camcorder will not be used in this research, published, or permanently stored – they will be removed after we processed video footages.

Data Archiving

The video sequences will be kept securely in an internal database. However, we would like to have this database (after pre-processing and removing audio) available to external researchers who are working in the field of Computer Vision so that this work can be carried out in a wider research community. Your contribution to the research hence is extremely valuable.

Please note the database will be only available to registered research institutions and for research purpose only. The data (video footages and Kinect data) will NOT be freely downloadable online. The following restrictions will be applied:

1. Any individual researcher or research group who wish to access the database has to fill in an online registration form, which will be *manually* checked by us for approval.
2. External bodies are NOT allowed to redistribute the database.
3. The database can ONLY be used for research purpose.
4. Copyright and non-redistribution forms have to be signed before releasing any data.
5. Only a very few short sample footage and video frames are to be published on the database website.

Confidentiality

Confidentiality and participant anonymity will be strictly maintained. All information gathered will be used for statistical analysis only and no names or other identifying characteristics will be stated in the final or any other reports.

Likelihood of Discomfort

There is no likelihood of discomfort or risk associated with participation.

Researchers

Dr. X. Xie is working as a lecturer (RCUK Academic Fellow) in the Computer Science Department at Swansea University. Dr. Xie can be contacted in Talbot 56, Swansea University. Contact number is 01792 602916 and email address is x.xie@swansea.ac.uk

Dr. P. W. Grant is working as a senior lecturer in the Computer Science Department at Swansea University. Dr. Grant can be contacted in 312 Faraday Tower, Swansea University. His phone number is 01792 2955392 and email address is p.grant@swansea.ac.uk

Dr. Hui Fang is working as Research Assistant in the Computer Science Department at Swansea University. Dr Fang can be contacted in room 401 Faraday Tower, Swansea University. His phone number is 01792 295393 (or internal extension 4534) and his email address is h.fang@swansea.ac.uk

Dr. Daubney can be contacted in room 401 Faraday Tower, Swansea University. His phone number is 01792 295393 (or internal extension 4534) and his email address is b.daubney@swansea.ac.uk

*A copy of this consent form has been given to you to keep for your records and reference.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to take part as a participant. In no way does this waive you legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to not answer specific items or questions in interviews or on questionnaires. You are free to withdraw from the study at any time without penalty. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact the researcher.

Participant	Date
Investigator/Witness	Date

*A copy of this consent form has been given to you to keep for your records and reference.

Appendix B

Instructions For Participants

Data Collection Instruction

Studying facial and body behaviors of interactions

Dear Participant:

Thank you for agreeing to take part in our data collection. The purpose of this work is to study human behaviour during interactions, specifically during one on one conversations. We expect the study to take about 30-40 minutes in total. You will have been assigned a partner with whom you will undertake the experiment with. During the data collection you will both be filmed with multiple cameras and Kinect sensors. To aide you in the direction and topic of conversation with your partner we have set some simple tasks that we will ask you to perform. These do not require any specific preparation on your part, though you may like to think about some of the answers before attending your session. There are no right or wrong answers and we are not interested in the audio content of any captured data, this will be removed during post processing.

It is very important that you do not discuss the tasks with anyone in the lab or your partner *before or after* the study.

For each task we have provided you with a rough idea of how long we envisage each task to last. This is just for guidance and does not represent an upper or lower limit. The tasks we will ask you to perform are the following:

1. **Discussing an area of your current work**

   In this part of the study we would like you to your current work (e.g. project) to your partner. Following this, your partner will then describe back to you your work in their own words. They can ask questions to better understand your work and likewise you can interrupt if you feel they haven't understood the topic. Following this you will reverse roles with your partner and listen to them explaining their work and you will describe it back to them. As well as describing the work, you might want to explain its novelty and why you feel the work is important.

   We expect this task to take about 5 minute for each person including discussion.

2. **Telling stories to your partner**

   For this part of the study we would like you to tell an interesting story to your partner. This could be either your own experience or a story heard from others (e.g. holiday experience). Your partner will then tell you an interesting story.

   This task will take about 5 minutes for both stories.

3. **Debating and discussing selected topics with your partner**

   We have selected two topics. For the first you will be given a specific point of view that we would like you to adopt for the debate, this will be different from your partner. For the second topic you will discuss with your partner the main arguments in support of and against the provided statement. Following which we would like you to decide your own viewpoint and whether you both share the same viewpoint.

   For each topic, we expect it will take about 5 minutes.

4. **Telling jokes with each other**

   You will be provided with three jokes that we will ask you to try to remember. You will then take it turns telling your jokes. Do not worry if you do not find the jokes funny.

NOTE: The time expected to take, as indicated above, for each of the tasks/topics is merely a rough indication. There is NO need to check the time during experiment for this purpose.

Thank you again for you participation. The jokes and debate topics are provided on the following page:

TO be included for subject #1

Debate and Discussion

The debate topic will be "Should University education be free?" - You will support the view point that there should not be tuition fees and higher education is a right.

The discussion topic is "Have social networks have made the world a better place?" - we would like you to provide arguments both in support and against this topic. At the end of the discussion, between you two you may reach an agreement or remain disagreement.

We expect the debate and discussion to last about 5 minutes each.

Jokes

Q. What is the difference between a Ph.D. in mathematics and a large pizza?
A. A large pizza can feed a family of four...

A guy goes to the hardware store to buy some insecticde. He hold up a box and asks the store manager, " Is this stuff good for beetles?" The manager replies, " NO, it'll kill 'em"

While cleaning the attic, Joan and Harry found an old stub for some shoes they left at the repair shop 10 years ago. They thought it would be funny to go to the shop and see if the shoes were still there. So they did. They handed the stub to the repair man who took it and looked in the back. He came out again and said, "They'll be ready on Wednesday."

TO be included for subject #2

Debate and Discussion

The debate topic will be "Should University education be free?" - You will support the view point that university should be paid for by the individual.

The discussion topic is "Have social networks have made the world a better place?" - we would like you to provide arguments both in support and against this topic.

We expect the debate and discussion to last about 5 minutes each.

Jokes

Q. What did one wall say to the other?
A. I'll meet you at the corner.

A reporter was interviewing a 104 year-old woman: "And what do you think is the best thing about being 104?" She simply replied, "No peer pressure."

Two brothers jointly owned a business and both were wise in worldly ways. While dying, one brother instructed his sibling to put half of their combined wealth into the grave with the casket. The brother reluctantly agreed. In time his brother died. At the graveside ceremony the living brother wrote a check for half of their assets and placed it in the casket.

Appendix C

Questionnaire For Participants

QUESTIONNAIRE for Participant 1

1) Are you a native English speaker? ☐ Yes ☐ No

2) Your gender ☐ Male ☐ Female

3) If you are a student, could you answer the following questions

(a) What is your subject _____

(b) What level are you in _____

Jokes:

4) Have you heard of any of the jokes before?

1st : "A large pizza can feed a family of four"

☐ Yes ☐ No

2nd : "Is the insecticide good for beetles, no, it'll kill them"

☐ Yes ☐ No

3 rd : "The shoes will be ready on Wednesday after 10 years"

☐ Yes ☐ No

5) Do you think the jokes worked?

1st ☐ Yes ☐ No

2nd ☐ Yes ☐ No

3 rd ☐ Yes ☐ No

Comment:

_____

_____

_____

4) Debate Topic, “Should University education be free?”

Do you think you have convinced your partner? Give a score from 1 ~ 5

1 – Not Convinced	2 – Still arguable	3 - Neutral	4 – OK to agree	5 - Convinced

Comment:

5) Discussion Topic, “Have social networks made the world a better place?”

a. Do you think you and your partner reached an agreement at last?

☐ Yes ☐ No

Comment:

b. Please tell us how strongly you agree or disagree with each other in discussing this topic?

☐ Strongly agree ☐ Moderately agree ☐ Neutral

☐ Moderately disagree ☐ Strongly disagree

Comment:

Your help is very much appreciated.

Bibliography

[AC99] Jake K. Aggarwal and Quin Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[AR11] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16, 2011.

[AYG11] Gabriele Fanelli Angela Yao, Juergen Gall and Luc Van Gool. Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference*, pages 67.1–67.11. BMVA Press, 2011.

[BEZ09] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[Bis06] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2006.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 10.1023/A:1010933404324.

[CC08] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, (41):3054–3067, 2008.

[CET01] T. Cootes, G. Edward, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[CSK11] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Technical Report MSR-TR-2011-114, Microsoft Research, Oct 2011.

[CTFP05] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24(4):1283–1302, 2005.

[DDD+09] Peng Dai, Huijun Di, Ligeng Dong, Linmi Tao, and Guangyou Xu. Group interaction analysis in dynamic context. *Trans. Sys. Man Cyber. Part B*, 39(1):34–42, February 2009.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[DRCB05] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE international workshop on: Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[DUL+06] Z. Deng, Neumann U., J. Lewis, T. Kim, M. Bulut, and S. Narayanan. Expressive facial animation synthesis by learning speech coariticulation and expression spaces. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1523–1534, 2006.

[EF78] P Ekman and W.V Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[Gav99] Dariu Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[GC94] A. H. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, 12:639–647, 1994.

[HCL+09] Yuxiao Hu, Liangliang Cao, Fengjun Lv, Shuicheng Yan, Yihong Gong, and T.S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *IEEE International Conference on Computer Vision*, 2009.

[IB00] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):852–872, August 2000.

[JC06] Seong-Wook Joo and R. Chellappa. Attribute grammar-based event recognition and anomaly detection. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, june 2006.

[KG04] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, August 2004.

[KP07] I Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Image Processing, IEEE Transactions on*, 16(1):172–187, 2007.

[Lap05] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Interantional Journal of Computer Vision*, 60:91–110, November 2004.

[LWSC10] S Lucy, Y. Wang, J. Saragih, and J. Cohn. Non-rigid face tracking with enforced convexity and local appearance. *Image and Vision Computing*, (28):781 –789, 2010.

[LXG10] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *Int. J. Comput. Vision*, 90(1):106–129, October 2010.

[LXG12] Chen Change Loy, Tao Xiang, and Shaogang Gong. Incremental activity modeling in multiple disjoint cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1799–1813, September 2012.

[MHK06] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[MJZF11] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Here's looking at you, kid. detecting people looking at each other in videos. In *British Machine Vision Conference*, 2011.

[MK08] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, March 2008.

[MRC05] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.*, 24(3):677–685, July 2005.

[NGP11] M. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[NN07] P. Natarajan and R. Nevatia. Coupled hidden semi markov models for activity recognition. In *Motion and Video Computing, 2007. WMVC '07. IEEE Workshop on*, page 10, feb. 2007.

[OGH04] Nuria Oliver, Ashutosh Garg, and Eric Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.*, 96(2):163–180, November 2004.

[ORP00] Nuria Oliver, Barbara Rosario, and Alex Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):831–843, 2000.

[PP06] M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 36(2):433–449, 2006.

[PPMZR10] A. Patron-Perez, M. Marszałek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. In *British Machine Vision Conference*, 2010.

[PR00] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[RA09] M. S. Ryoo and J. K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *Int. J. Comput. Vision*, 82(1):1–24, April 2009.

[RA11] M. S. Ryoo and J. K. Aggarwal. Stochastic representation and recognition of high-level group activities. *Int. J. Comput. Vision*, 93(2):183–200, June 2011.

[Rab90] L. R. Rabiner. A tutorial on hidden markov models and selected apllications in speech recognition. In *Readings in Speech Recognition*, pages 267–296, 1990.

[SAS07] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 357–360, New York, NY, USA, 2007. ACM.

[SGM09] C. Shan, S. Gong, and P. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27:803–816, 2009.

[TBA+09] P. Tresadern, H. Bhaskar, S. Adeshina, C. Taylor, and T. Cootes. Combining local and global shape models for deformable object matching. In *The proceedings of BMVC*, pages 1 – 10, 2009.

[TCSU08] Pavan K. Turaga, Rama Chellappa, V. S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488, 2008.

[TKC01] Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97 –115, 2001.

[TLJ07] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1683–1699, 2007.

[ZGPBM06] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan. Modeling individual and group actions in meetings with layered hmms. *IEEE Transactions on Multimedia*, 8(3):509–520, 2006.

[ZPRH09] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.