THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

UNDERSTANDING HUMAN INTENTION VIA NON-INTRUSIVE CAPTURING OF INTERACTION AND BODY SIGNALS

FU YUJUN

PhD

The Hong Kong Polytechnic University

2019

The Hong Kong Polytechnic University

Department of Computing

Understanding Human Intention via Non-intrusive Capturing of Interaction and Body Signals

Fu Yujun

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

October 2018

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Fu Yujun (Name of student)

Abstract

Computers are commonplace and their ability to understand human intention will open up the possibility to provide potential useful assistance to human in many applications, including human social interactions and daily human-computer interactions. In this thesis, I investigate into these interactions to understand human intention.

For human social interactions, I focus on a common type of social interaction in real life, namely, human fight. Prior studies about fight detection are challenged by some constraints, including the reliance on costly high level feature recognition, like human gestures, actions, or visual words and simulated fight events due to dataset availability. I propose two sets of motion analysis-based features to build human real fight detection models, without recognizing human gestures or visual words. To evaluate, we collect our own human real fight datasets. Experiments demonstrate that my models outperform the state-of-the-art counterparts in human real fight detection. I further extend my investigation to understand fights with real fight intention and simulated fights. The findings suggest that there are fundamental differences between human real fights and simulated fights, and my motion analysis-based features can effectively distinguish them. State-of-the-art data-driven approaches, such as deep learning, are constrained by the limited amount of spontaneous human real fight data available. To address this, I propose an ensemble-based method for cross-species fight detection by adapting knowledge from real animal fight events. Experiments demonstrate that it is feasible to build well-performing human real fight detection models via cross-species learning.

For daily human-computer interaction tasks, I study the user intention

prediction problem. The challenges include limited prescribed tasks, lack of full modalities and the need of expensive intrusive devices to capture interaction and body signals, especially physiological signals. First, I conduct the study on a common but more complex and open-ended daily computer interaction task: web search task, to overcome the limitation of studying on simple prescribed tasks. Second, I propose two feature representations to encode users' interaction and body signals, including mouse, gaze, head and body motion signals. I combine these signal features with historical activity sequences to build effective multimodal user intention prediction models. Experiments indicate that the proposed features can successfully encode these signals and the model can achieve encouraging performance. I further extend the work to the application of detecting user slips. Experiments provide evidence about the feasibility of building useful intention-based user slips detection models. Third, I would like to capture the interaction and body signals with non-intrusive devices, as opposed to contemporary physiological signal measurements with expensive intrusive devices. Since physiological signals could well indicate human emotions and even intentions, measuring them in a non-intrusive and low cost manner would benefit human emotion and intention understanding. I propose a physiological mouse and build a prototype to non-intrusively measure human heart beat and respiratory rate. Experiments illustrate that the mouse can achieve promising performance on measuring the two physiological signals. Further experiments also suggest that it is feasible to correlate the measured signals to human emotions via the physiological mouse prototype.

List of Publications

[1] **Yujun Fu,** Hong Va Leong, Grace Ngai, Michael Xuelin Huang, Stephen C.F. Chan. 2014. "Physiological Mouse: Towards an Emotion-Aware Mouse," in *IEEE International Computer Software and Applications Conference Workshops(COMPSACW)*, 2014, pp. 258–263.

[2] **Eugene Yujun Fu,** Hong Va Leong, Grace Ngai, Stephen C.F. Chan. 2015. "Automatic Fight Detection Based on Motion Analysis," in *IEEE International Symposium on Multimedia (ISM)*, 2015, pp. 57–60.

[3] **Yujun Fu,** Hong Va Leong, Grace Ngai, Stephen C.F. Chan. 2016. "Non-Intrusive Health-Monitoring Devices," in *Encyclopedia of E-Health and Telemedicine, Chapter 55*, 2016, pp. 711–721.

[4] **Eugene Yujun Fu,** Hong Va Leong, Grace Ngai, Stephen C.F. Chan. 2016. "Automatic Fight Detection in Surveillance Videos," in *Proceedings of ACM International Conference on Advance in Mobile Computing and MultiMedia*, 2016, pp. 225-234. **- Best Student Paper Award**

[5] **Yujun Fu,** Hong Va Leong, Grace Ngai, Michael Xuelin Huang, Stephen C.F. Chan. 2017. "Physiological Mouse: Towards an Emotion-Aware Mouse," in *Universal Access in the Information Society*, 2017, pp. 365–379.

[6] **Eugene Yujun Fu,** Hong Va Leong, Grace Ngai, Stephen C.F. Chan. 2017. "Automatic Fight Detection in Surveillance Videos," in *International Journal of Pervasive Computing and Communications*, 2017, pp. 130-156.

[7] **Eugene Yujun Fu**, Tiffany C. K. Kwok, Erin You Wu, Hong Va Leong, and Grace Ngai. 2017 "Your Mouse Revals Your Next Activity : Towards Predicting User Intention from Mouse Interaction," in *Proceedings of the*

IEEE International Computer Software and Applications Conference (COMPSAC), 2017, pp. 869-874.

[8] **Eugene Yujun Fu,** Michael Xuelin Huang, Hong Va Leong, Grace Ngai. 2018. "Cross-Species Learning: A Low-Cost Approach to Learning Human Fight from Animal Fight," in *Proceedings of ACM Multimedia Conference (ACMMM),* 2018, pp. 320-327.

[9] Tiffany C. K. Kwok, **Eugene Yujun Fu**, Erin You Wu, Michael Xuelin Huang, Grace Ngai, and Hong Va Leong. 2018. "Ev'ry Little Movement Has a Meaning of Its Own: Using Past Mouse Movements to Predict the Next Interaction," in *Proceedings of the 23th International Conference on Intelligent User Interfaces (IUI)*, 2018, pp. 397-401.

[10] Jun Wang, **Eugene Yujun Fu,** Grace Ngai, Hong Va Leong. 2019. "Detecting Stress from Mouse-Gaze Attraction". To appear in *Proceedings of ACM/SIGAPP Symposium on Applied Computing*, 2019.

[11] Jun Wang, **Eugene Yujun Fu,** Grace Ngai, Hong Va Leong. 2019. "Investigating Differences in Gaze and Typing Behavior Across Age Groups and Writing Genres". - Submitted. To appear in *Proceedings of the IEEE International Computer Software and Applications Conference (COMPSAC)*, 2019.

Acknowledgements

I would like to express my sincere gratitude to all the people that have assisted me to complete this degree. The following acknowledgments are by no means exhaustive, for which I apologize.

I am extremely grateful and remained indebted to my supervisor, Dr. Hong Va Leong, for his full support and expert guidance. Without the encouragement, constructive criticism and helpful advice from him, my thesis work would have been an overwhelming and frustrating pursuit.

I would also like to thank the professors in my research group: Dr. Grace Ngai, and Dr. Stephen C.F Chan, who patiently support my work through instructional discussions, detailed analyses and continuous suggestions. Their constant sources of knowledge and inspiration provide invaluable guidance through my study.

I have had great pleasure working with members in CHILab: Dr. Michael Xueling Huang, Dr. Jiajia Li, Dr. Yuanyuan Wang, Tiffany Kwok, Jun Wang, You Wu, and Andy Tam. The creativity of all my colleagues has been a constant inspiration throughout my time.

Finally, I would like to acknowledge my parents and my wife, who unconditionally support me in all my decisions.

Table of Contents

Certificate of Originality iii
Abstract iv
List of Publications vi
Acknowledgements viii
Table of Contents ix
List of Figures xiii
List of Tables xvii

Chapter 1 Introduction 1

1.1 Background and Motivation 3

1.1.1 Understanding Human Fight Action and Intention 3

1.1.2 Understanding User Intention in Daily Computer Tasks 6

1.1.3 Understanding Interaction and Body Signals 9

1.2 Study Overview 11

1.2.1 Detecting Human Real Fight Action via Motion Analysis 12

1.2.2 Modeling User Intention in Daily Computer Tasks 14

1.2.3 Monitoring Human Physiological Signals Non-intrusively 15

1.3 Thesis Aims and Outline 16

Chapter 2 Literature Review 19

2.1 Human Fights and Aggressive Behavior Detection 19

2.2 Animal Action Recognition 22

2.3 Transfer Learning 22

2.4 User Interaction Intention Prediction 23

2.4.1 User Intention Detection 23

2.4.2 User Errors Detection 26

2.5 Interaction and Body Signals Analysis 27

2.5.1 Interaction Signals 27

2.5.2 Body Motion Signals 29

2.5.3 Physiological Signals 30
2.6 Summary of Related Works 31
Chapter 3 Automatic Fight Detection via Motion Analysis 33
3.1 Fight Detection by Motion Analysis 36
3.1.1 Optical Flow Images 37
3.1.2 Extracting Motion Signal Features 40
3.1.3 Extracting Local Motion Features 46
3.2 Constructing Human Real Fight Dataset 49
3.3 Evaluating Fight Detection Model in Real Fights 51
3.3.1 Experiments and Results 51
3.3.2 Comparison with the State-of-the-Art Approaches 53
3.4 Discriminating Real Fights from Simulated Fights 55
3.4.1 Evaluating in Simulated Fights 55
3.4.2 Manual Detection 60
3.4.3 Machine Detection 62
3.5 Cross-species Learning in Fight Detection 63
3.5.1 Source Datasets 64
3.5.2 Ensemble-based Adaptation 66
3.5.3 Evaluating Cross-species Learning 70
3.6 Summary 75
Chapter 4 Exploring Multi-modalities User Intention Prediction 78
4.1 User Intention Task 81
4.2 Extracting Features from Multi-modalities 83
4.2.1 Features from Historical Activities 83
4.2.2 Features from Interaction and Body Signals 86
4.3 Constructing User Intention Dataset 95

4.4 Evaluating User Intention Prediction 96
4.4.1 Modeling Historical Activity Sequence 97
4.4.2 Modeling Individual Interaction and Body Signals 100
4.4.3 Going towards Multi-modalities 107
4.5 Towards User Slips Detection 110
4.6 Summary 112
Chapter 5 Physiological Mouse - Non-intrusive Measurement of Physiological Signals 114
5.1 Physiological Mouse Prototype 117
5.2 Measuring Physiological Signals via Physiological Mouse 119
5.2.1 Measuring Heart Beat Rate 119
5.2.2 Measuring Respiratory Rate 121
5.3 Evaluating Physiological Signals Computation 125
5.3.1 Evaluating Heart Beat Rate 125
5.3.2 Evaluating Respiratory Rate 128
5.4 Correlating Physiological Signals with Human Emotions. 134
5.5 Summary 142
Chapter 6 Conclusion and Future Work 143
6.1 Contributions 144
6.1.1 Detecting Human Real Fight 144
6.1.2 Modeling User Interaction Intention 145
6.1.3 Non-intrusively Measuring Physiological Signals 145
6.2 Limitations 146
6.3 Future Work 147
6.3.1 Detecting Intention to Fight 147

6.3.2 Investigating on Diversified Real Tasks 148
6.3.3 Applying Physiological Mouse to User Intention Prediction 149
6.3.4 Integrating with Deep Learning Approaches 149
6.4 Other Relevant Contributions 150
6.4.1 Using LSTM for User Intention Prediction 150
6.4.2 Using LSTM for Fight Detection 151
6.4.3 Modeling Mouse and Gaze Interaction for Stress Detection 154
References 155

List of Figures

Figure 1-1 The flow of this thesis. This thesis studies human intention understanding in two aspects. In human social interactions, this thesis investigates (1) automatic human real fight detection. In human-computer interactions, this thesis explores (2) multi-modalities user intention prediction and (3) non-intrusive measurement of physiological signals for emotion detection. 11

Figure 3-1 Computing optical flow image. (a) original image, (b) optical flow image, (c) noise removal 38

Figure 3-2 Color code scheme (a) and optical flow image (b). 39

Figure 3-3 Color code scheme (a) and light change in optical flow images (b). 40

Figure 3-4 Example of motion signal features extraction. 40

Figure 3-5 Decision tree for classifying motion types. 41

Figure 3-6 Motion attraction computation. $M1$ and $M2$ are the motion magnitudes of motion region 1 and 2 respectively. While, $D12$ is the Euclidean distance between the centroids of the two motion regions. 44

Figure 3-7 Examples of the optical flow image and local motion regions for extracting local motion features. 46

Figure 3-8 Example of fight scenes from human real fights (first row) and simulated fights (second row). 49

Figure 3-9 Example fight scenes from animal fights (first row), hockey fights (second row) and action movies (third row). 65

Figure 3-10 System architecture. Our cross-species learning is achieved through ensemble learning. Taking local motion features as an example, we

extracted features from animal fight videos and a small amount of available human fight videos for ensemble learning. 67
Figure 3-11 Example to illustrate the cross-species learning. 69
Figure 3-12 Performance of adaptation by learning from ensemble classifiers. (a) adaptation from animal fights, (b) adaptation from hockey fights, (c) adaptation from action movies. 72
Figure 3-13 The effect of different feature sets on ensemble learning in human fight detection when adapting from (a) animal fights, (b) hockey fights, and (c) action movies. The proposed *LMF* features generally outperform the state-of-the-art motion features across adaptation from different datasets. 75
Figure 4-1 Web search task and the five types of activities 82
Figure 4-2 Example of capturing gaze and head movement from a webcam with the help of OpenFace toolkit [7]. 88
Figure 4-3 Example of the body based local motion regions. 1 head region, 2 eye region, 3 mouth region, 4 right body region, 5 left body region, 6 right shoulder region, and 7 left shoulder region 89
Figure 4-4 Example of an interaction movement and the attributes 90
Figure 4-5 Mapping interaction movements of (a) mouse, gaze movements, (b) head movements, and (c) body motions, to corresponding histogram bin to generate histogram-based features (d). 93
Figure 4-6 Performance of intention prediction by modeling historical activity sequence only. 98
Figure 4-7 Performance of intention prediction. Features: *SF* within the W most recent seconds. 102
Figure 4-8 Performance of intention prediction. Features: *HF* within the W most

recent seconds. 103
Figure 4-9 Results of predicting users' intention X seconds ahead by using different modalities individually. 105
Figure 4-10 Examples of mouse movements (a), and their corresponding histogram representations (b). Longer radius means more movement magnitudes occur in that direction, deeper color means movement in that direction occur in more recent. The histogram representations can distinguish the two movements. 106
Figure 4-11 Results for predicting users' intention X seconds ahead with multi-modalities. 108
Figure 4-12 Results for predicting users' intention X seconds ahead with different gaze estimation methods. 109
Figure 4-13 Results for user slips detection. 112
Figure 5-1 The prototype of the physiological mouse (a) and its usage (b). ...118
Figure 5-2 Signal smoothing. 120
Figure 5-3 Frequency domain of heart beat (a) and inter-beat interval (b). 120
Figure 5-4 A histogram for respiratory rate candidates. 124
Figure 5-5 iHealth device to measure heart beat rate. 126
Figure 5-6 Average heart beat rate error for all subjects. 127
Figure 5-7 Heart beat rate error for individual subjects. 128
Figure 5-8 Example of candidate respiratory rates in controlled experiment (a good case). 130
Figure 5-9 Example of candidate respiratory rates in controlled experiment (a bad case). 130
Figure 5-10 Average result with error bar for different respiratory rhythms. ..131
Figure 5-11 Average respiratory rate error for all subjects: Natural respiration.

.....133

Figure 5-12 Respiratory rate error for individual subjects: natural respiration. 133

Figure 5-13 Experimental setup for emotion-related experiments by using the physiological mouse.....135

Figure 5-14 Average heart beat rates across subjects in different tasks. 136

Figure 5-15 Average respiratory rates across subjects in different tasks. 137

Figure 5-16 Average heart beat rates across tasks.....139

Figure 5-17 Average respiratory rates across tasks. 139

Figure 5-18 An example of temporal physiological signals captured in playing game. 141

Figure 5-19 An example of temporal physiological signals captured in watching the horror video. 141

Figure 6-1 A dual-stream LSTM for user intention prediction. 151

Figure 6-2 Learning fight detection model from local motion signals using SVM and LSTM. 153

List of Tables

Table 3-1 Motion signals and their statistical features. 46

Table 3-2 Full set of motion signal features generated for a video. 46

Table 3-3 Local motion sequences and statistical features. 48

Table 3-4 Performance of motion signal features on real fights. 52

Table 3-5 Precision and recall of motion signal features on real fights. 52

Table 3-6 Performance of local motion features on real fights. 53

Table 3-7 Precision and recall of local motion features on real fights. 53

Table 3-8 Evaluations on human real fight dataset. 55

Table 3-9 Performance of motion signal features on simulated fights. 57

Table 3-10 Precision and recall of motion signal features on simulated fights. 57

Table 3-11 Performance of local motion features on simulated fights. 57

Table 3-12 Precision and recall of local motion features on simulated fights. ..58

Table 3-13 Evaluations on human simulated fight dataset. 58

Table 3-14 Manual detection on discriminating real fights from simulated fights. 61

Table 3-15 Performance of motion signal features on discriminating real fights from simulated fights. 62

Table 3-16 Performance of local motion features on discriminating real fights from simulated fights. 63

Table 3-17 pseudocode of learning ensemble classifiers. 68

Table 3-18 Evaluations on various fight datasets. 70

Table 4-1 Features from historical activities. 86

Table 4-2 Movement attributes and statistical features. 91

Table 4-3 Statistical features from multi-modalities. 92

Table 4-4 Histogram features from multi-modalities.....95

Table 5-1 Heart beat rate performance.....128

Table 5-2 Respiratory rate performance: controlled experiment. 132

Table 5-3 Respiratory rate performance: natural respiration. 133

Chapter 1 Introduction

Affective computing is an up-surging research area relying on multimodal multimedia information processing techniques to understand human emotion and intention. Recent studies in affective computing have developed beyond simply the recognition of human basic emotions. There are several new directions, such as social signal processing [102] and user intention prediction [50, 90]. Many of these studies investigated on recognizing not only human basic emotions, but also human behaviors and intentions. They focused on the analysis of the communicative or informative patterns to understand the underlying intention behind a human action, in computer interaction tasks, or group and social interactions, etc. The former is more oriented towards human-computer interaction, and the latter more towards human-human interaction.

Once a computer is able to interpret human intention, it can offer assistance to human in advance, in both scenarios of social events and daily human-computer interactions. For instance, when a computer detects the intention of fierce activities inside a bar, it can raise an alarm for warning, which may help to deter the potential violence from taking place. A computer can also help to automatically enlarge a potential target button for a user in a computer interaction task, when it detects the user's interaction intention. We therefore focus our study on understanding human intention by using interaction and body signals in this thesis. We conduct our studies in both of the two directions respectively.

We first investigate techniques for understanding human action and intention in human-human interaction. In our work, we focus on a special kind of social event and human interaction: human fight. Although certain fight detection approaches show promising results in prior studies, not all of them are suitable for

real surveillance applications, due to some limitations, including relying on recognizing high level features, such as human gestures, or visual words, etc. [81, 114] and studying on simulated fight events [10, 29, 66]. However, recognizing high level features may require high computational cost and high quality videos. Meanwhile, the simulated fights may not represent the spontaneous fights in real situations, although they may share some similar gestures. We therefore see the challenges of detecting human fights without recognizing high level features as well as understanding the difference between real fights and simulated fights. As the real fights and simulated fights can be considered as the fight actions with and without real fight intention, this study would also help us understand more about human fight intention.

In addition, we also investigate techniques for predicting user intention in daily human-computer interaction tasks. There are few studies attempting to address this, but they are quite preliminary, either not studying on real applications nor using all reasonable modalities. On one hand, the approaches applied in prescribed tasks such as one mouse movement towards a predefined target are not extendable to real tasks, which may contain multiple interaction steps [5, 85, 124]. To our best knowledge, user intention prediction is not well investigated in the scenarios of daily computer interaction tasks. We therefore see the challenge of building an effective model to predict user intention in daily computer interaction tasks by using multiple interaction and body signals.

On the other hand, in order to utilize all reasonable modalities, we should be able to capture users' interaction and body signals during interaction tasks. Meanwhile, intrusive capturing methods would affect users' feeling and even intention. With the help of some webcam and computer vision techniques, we can capture the signals of mouse, gaze, head and body movements in non-intrusive

manners. However, most of the contemporary approaches for measuring human physiological signals are still relying on intrusive and expensive devices. We therefore see the challenge of investigating the feasibility of measuring human physiological signals in a non-intrusive manner. As physiological signals are among those useful indicators towards human emotions and intention, studying on the non-intrusive physiological signals measurement would also benefit the future studies in human-computer interaction.

To address the challenges described above, this thesis thus investigates on three studies. The details of the background and study overview of these studies are described in the following.

1.1 Background and Motivation

1.1.1 Understanding Human Fight Action and Intention

To understand human intention in human-human interaction, we focus on the detection of human fight in our work. Human fight is an important class of human-human interactions and social signals that has been drawing increasingly attention in recent years. Due to the growing need of fast response to social conflicts and security issues, there has been a mounting demand to automatically detect human fights in video surveillance scenarios. Therefore, in this thesis, we start our study by detecting human fights and understanding human real fight intention.

However, there are some drawbacks in recent fight detection or aggression detection studies. One of the major drawbacks is that the approaches proposed in most of these studies rely on extracting features from human gestures, actions or visual words [81, 114]. On one hand, extracting these features induces high computational cost in processing the surveillance videos. That may affect the

response time of detecting key events for a surveillance system. On the other hand, these approaches rely on the availability of relatively high quality videos, which cannot be readily provided by most of the surveillance systems in real life. This is partially due to the elevated surveillance angle and sometimes the large area to be covered. To address this challenge, in this thesis, we therefore propose to detect human fights by using motion analysis-based features, which can be obtained by a natural and low computation approach.

Another drawback of the prior fight detection studies is that most of them evaluated their approaches on simulated fight datasets [10, 29, 66]. On the contrary, little work had been performed on real fight surveillance scenarios. Although there are some similarities between real fight actions and simulated fight actions such as the gesture of raising a fist, they are triggered by different human intentions. The difference can be captured by body motion signals, such as the motion magnitudes. Since in the real fight scenarios, human really wants to knock down his/her opponent, the fight actions triggered by real fight intention would contain larger motion magnitudes. However, in simulated fights, since actors just want to express fight gestures, the simulated fight actions might not contain sufficient motion magnitudes to bring down the opponent. Simulated fights thus may not represent the real fight intention. Therefore, in this thesis, we would like to study fight detection in real application context. Since there is no standard surveillance dataset involving fights available, we proceed to collect and annotate human real fight dataset by ourselves. We would like to investigate whether our algorithms can detect real fight events in real surveillance effectively. Moreover, in order to further reveal the difference between real and simulated fights and gain a deeper understanding of human real fight intention, we would also like to evaluate our algorithms on detecting real fights from simulated fights in our study. Indirectly,

our results demonstrate that real fight events do differ from simulated fight events, especially in motion signals.

On the other hand, data-driven approaches, in particular, deep learning approaches have been explored to learn effective prediction model in many aspects [107]. Recent fight detection studies also attempted to apply deep learning approaches to learn fight detection model directly from data [107, 122]. However, well-performing data-driven methods require high model capacity and thus a substantial amount of well-annotated data, which is generally difficult to acquire, or even impractical in certain domains, such as rare disease diagnostic, extreme social action monitoring, which includes human fight detection.

Transfer learning such as adapting knowledge from real fight events in other scenarios to detect human real fights can be an effective solution. We would like to adapt knowledge from fights in real scenarios. An alternative source is using real fight videos, but not by human. Therefore, we turn our attention to fights involving animals. A natural question would then be "would human fight like animals"? Interestingly, there are a good amount of animal fight videos on the web. That would enable us in alleviating the data availability issue as well as attempting to answer this interesting question. The fighting actions exhibited by human and animals share intrinsic commonality, such as physical acceleration of moving body parts. Moreover, in contrast to the use of human videos, fewer privacy issues are involved in using animal fight videos. Inspired by the work of [105], we propose an ensemble-based method for cross-species fight detection to address this challenge. To the best of our knowledge, we are the first in investigating animal fights and cross-species learning in fight detection. In this thesis, we investigated the effect of different source data and feature sets on the adaptation of learning from fights in other scenarios to real human fights.

1.1.2 Understanding User Intention in Daily Computer Tasks

In addition to understanding human intention in human-human interaction and social events, user experience has also gained attention in recent researches in human-computer interaction. To improve user experience, some of these works attempted to estimate user's visual saliency to understand user's attention [112]. Based on that, they can also adjust the visual saliency of user interface [63] to enhance user's navigation experience. The other work tried to study on detecting user's activity state, skill learning progress [59] and task performance [11] as well as understanding user's action [90], etc. In addition, since complex and multi-step computer interaction tasks require a user to perform a series of actions, predicting the user's next interaction activity in advance could be a potential research direction of enhancing user experience. Recent studies thus start to explore user intention prediction. The major objective of these works is to predict potential human action in the near future and understand the underlying intention beneath [20]. If a computer has the ability to perceive user intention, it can offer help to improve user experience, such as enlarging or highlighting the potential target. It can also correctly understand user's action, even for the action that might not reflect users' real intention, such as understanding whether an interaction action is triggered by user's mistake [76]. To our best knowledge, there are only few prior studies tried to approach user intention prediction, but they are quite preliminary.

One of the major drawbacks of the prior related studies is that most of these works only studied on specific tasks, which are under well-controlled conditions, such as using predefined fixed interfaces. In their experiments, instead of interacting with a real interaction task, a subject is only required to perform some simple mouse movement trials, such as pointing the mouse to a particular target [5, 85, 124] in a fixed interface. However, in real applications, a task may contain

multiple steps of interaction activities, such as multiple mouse movements and clicks, etc. Compared with the prescribed trials, the interaction activities occurring in real multi-step tasks are more complex, of which the user interaction intention is difficult to predict. The user intention in the real interaction tasks might depend on the user, the task, and often the current as well as the previous state of the task, etc., which make predicting user intention in daily computer usage very challenging. The previous intention prediction approaches applied in the prescribed trials are restricted to predefined situations, and may not be extensible to daily multi-step computer interaction tasks. On the other hand, there are some works studying user browsing intention [25, 70]. Their approaches rely on browsing context, which is not accessible in other tasks. They thus are also not extensible to other computer interaction tasks.

Another limitation of the prior studies is that they did not investigate all reasonable modalities for user interaction intention prediction. Most of the related studies focus on modeling user intention from historical records [2]. According to these studies, user's activities in multi-step interaction tasks are sequential and each of them is dependent on historical activities. However, in addition to historical information, interaction from other modalities including mouse, gaze and head movements, etc. may contain additional information about user's next activity in the multi-step interaction tasks. In other words, they are not independent but correlated. Some work attempted to build user intention prediction model by these kinds of interactions. However, their model only involved one particular interaction modality such as modeling mouse movement profile only [85]. Although these approaches may succeed in prescribed trials, they may not be able to predict user intention in daily computer tasks. With the assumption that different interaction modalities may contain different information of user's intention,

modeling multiple interaction modalities combined with the historical information could be a potential solution to build a more effective user interaction intention prediction model. However, to our best knowledge, very few works have attempted to investigate multimodal user intention prediction approach in daily computer interaction tasks.

We therefore see that the major challenge of this study is how to effectively predict user's intention or what is the upcoming activity during a daily human-computer interaction task, with multiple modalities, which is also one of the core research questions in this thesis. To address this challenge, we therefore propose a multimodal user intention detection approach for multi-step human-computer interaction task based on mouse, eye, head and body motions. Specifically, we conduct experiments and study user interaction intention in a common daily computer interaction task, namely, web search, to predict the type of the next activity. Focusing on this task, we investigate the appropriate feature representation and the proper way to fuse multi-modalities for user interaction intention prediction. We believe that these studies would benefit future research in this area.

In addition to building an effective user intention prediction model, we are also interested in applying the model to enhance user experience during an interaction task. Detecting user selection slips could be one of those applications. Slips are a type of human error, which is describing the wrong action triggered by users even they formulated the right intention [109]. Given that slips frequently occur during human-computer interactions, especially in complex and multi-step tasks which require users to perform a series of actions, user experience would be affected if systems fail to handle them. For instance, a user may need to handle an unexpectedly opened window, if he or she clicks a non-intentional button or link.

On the other hand, in real applications, additional time cost is required to recover from a slip, for example, a user may need some time for additional navigation to find the real intended target after a selection slip [8]. The ability of user slips detection is likely to help enhancing user experience in multi-step computer interaction tasks. Therefore, we then conduct a pilot study to investigate the new research question: can the multimodal intention prediction model be applied to user slips detection? To our best knowledge, we are the first to investigate the feasibility of using multimodal intention prediction model to detect user slips.

1.1.3 Understanding Interaction and Body Signals

Prior studies from psychology suggest that humans are able to interpret other's intention during their communications. Their further investigations also suggest that humans tend to express their intention in human-human interactions via head and body orientation as well as vocal signals [9, 76]. According to these studies, the head, body motion, and vocal signals can be regarded as the important cues for understanding human intention. However, speech and vocal signals are not available in some scenarios such as CCTV surveillance, which is a major source of recording human-human interactions and social events. Meanwhile, speaking is not common when users are interacting with daily computer interaction tasks, such as web search. We therefore do not consider speech and vocal signals in this thesis.

Instead, we observe that mouse and gaze interactions are important in daily computer tasks. Users can use the mouse to achieve different tasks, such as clicking buttons, selecting text, etc. Prior studies are focused on analyzing users' mouse behavior to perform authentication [19, 121], and gaze position alignment [41], etc. In daily computer usage, the mouse can be regarded as an extension of the human body in daily computer interaction. And users' interaction intention can

also be expressed via mouse movements. On the other hand, gaze interaction can indicate attention content and intended clicking targets in computer interaction tasks, therefore, it may also reveal users' intention. Both of them can be captured in non-intrusive manners. Mouse coordinates can be obtained via system logs, while gaze position estimation can be achieved by eye track devices such as Tobii, and webcams through some gaze estimation algorithms [7, 42, 120]. Therefore, our studies on understanding human intention also involve the analysis of mouse and gaze interaction.

In human face-to-face interaction, human can interpret partners' intention simultaneously using the interaction and body signals from multiple modalities. By doing this, human can understand other's intention in a more flexible and robust way. Even interpreting intention from one particular modality fails, they can still interpret other's intention from other modalities [84]. We therefore propose to address the challenge of modeling multiple interaction and body signals to predict human intention in this thesis. In CCTV surveillance, the only signal that can be captured is the body movement. We therefore focus on body motion analysis-based approaches to study the human fight intention. In daily computer interaction tasks, we are able to capture multiple interaction and body signals from system logs and webcam, etc. We then study the multi-modal intention prediction approaches in daily computer interaction tasks.

In addition to these interaction and body motion signals, physiological signals can also be applied to detect human emotion and even intention. Moreover, when compared with other body signals, physiological signals can even represent the unaware intention that is hard to control. However, the major drawback of traditional physiological signal measurement is that the measurement needs to rely on some intrusive devices. In order to make use of physiological signals to detect

human emotion and even intention in daily computer usage, we also investigate a non-intrusive way of capturing human physiological signals by enhancing standard input device in this thesis.

1.2 Study Overview

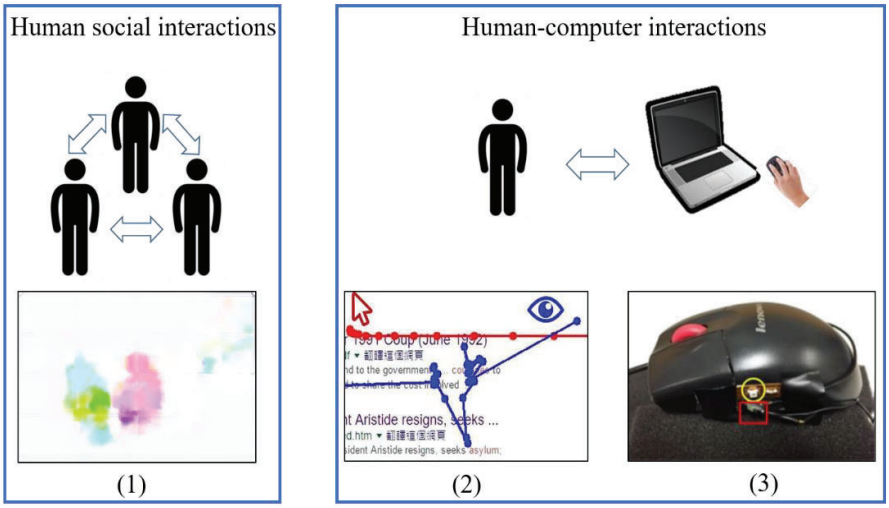


Figure 1-1 The flow of this thesis. This thesis studies human intention understanding in two aspects. In human social interactions, this thesis investigates (1) automatic human real fight detection. In human-computer interactions, this thesis explores (2) multi-modalities user intention prediction and (3) non-intrusive measurement of physiological signals for emotion detection.

In this thesis, we studied human intention understanding in two aspects. Figure 1-1 presents the flow of this thesis. To have a better understanding of detecting human action and intention via body motion analysis, we first conduct our study in the scenario of human-human interaction. We focus on detecting human fight and understanding human real fight intention in our work. Another essential issue of this thesis is to build a well-performing user intention prediction model via modelling multiple modalities for common daily computer interaction tasks. In our study, we investigate the user interaction intention in the web search task. The modalities applied in our studies include history activity records, mouse interaction, eye gaze, head as well as body movements, which can be captured by

non-intrusive devices. Moreover, since physiological signal could be a potential modality to predict user intention, we also conduct a pilot study to make use of user's physiological signals as a new form of modality in determining human emotion, in a non-intrusive manner.

1.2.1 Detecting Human Real Fight Action via Motion Analysis

To investigate human-human aggressive interaction and understand human real fight intention, we first propose motion analysis-based approaches to detect human fights. Compared with deep learning and prior approaches that rely on recognizing human gestures or visual words, our proposed approach can be an ideal method to detect human fight in a low-computational cost manner.

In our approach, we extract motion information from video clips based on optical flow. We then propose two approaches to further extract features from the motion information. In the first approach, we propose to extract features from the motion signals occurring with different types of motion, which can be detected by using the nature of motion regions present such as size of the motion region, without relying on the complicated mechanisms of gesture or action recognition. The motion signals involved in this approach include motion magnitude, acceleration and a new concept called motion attraction. We refer to the features extracted via this approach as the “motion signal features”. In our second approach, we propose to extract features from the motion signals occurring in different locations of a frame. More specifically, inspired by the local binary features proposed in [37], we try to extract the local motion features that describe motion amplitudes and accelerations within different regions of video frames. We refer to the features extracted via this approach as the “local motion features”.

The proposed approaches are evaluated on a human real fight dataset, which is collected by our own, and another publicly available dataset which collected

fight events simulated by subjects. The evaluation results show that both of the proposed approaches can achieve promising performance on human real fight detection, and the local motion features can obtain better accuracy than the motion signal features. Moreover, this study also reveals that simulated fights are fundamentally different from real fights and our motion analysis-based features can detect the difference.

Another challenge of building a well-performing human fight detection model is the lack of a large amount of training dataset, especially for the data-driven approaches such as deep learning. To address this challenge, we propose an ensemble-based method for cross-species fight detection with the proposed features. Interestingly, the motion attributes such as physical acceleration of fighting actions from human and animals share some intrinsic similarities. We then apply an ensemble technique to adapt useful knowledge from similar animal fights to learning human fights.

To evaluate the cross-species fight detection approach, we collect a dataset of animal fight. In the evaluation experiment, we apply the ensemble techniques to train on animal fights combined with a modest amount of data from the collected human real fight dataset and test on real human fights. Our experimental results show that our approach achieves a competitive accuracy with adaptation from animal fights to that of real human fights. A close scrutiny reveals that our proposed local motion feature representation describes the intrinsic motion attributes and thus is well generalizable across species and it is very beneficial to the learning of cross-species fight detection. We believe that our study can shed light on the appropriateness of feature representation for cross-species learning for human fights as well as the effectiveness of adaptation from different sources.

1.2.2 Modeling User Intention in Daily Computer Tasks

Apart from understanding human fight intention (human aggressive intention in human-human interaction), it is also important to study user intention in daily computer interaction tasks. We therefore conduct a study to investigate user intention prediction in the scenario of multi-step human-computer interaction tasks. We focus on studying in natural web search task, as it is a common and multi-step interaction task. In this study, we aim to utilize the interaction and body signals captured by non-intrusive and low cost devices to predict users' next web search activity. Fortunately, prior studies in computer science show that mouse interaction, eye gaze interaction, head movements, and body motion signals can be captured by non-intrusive and low cost devices. We then focus on modeling these interaction and body signals for user intention prediction.

A well-performing model relies on effective features. In our study, we propose two feature representations to model mouse, eye gaze, head and body movements. Inspired by prior related studies, we first apply statistics-based features to summarize the interaction signals. We also hypothesize that considering the movement magnitude, orientation information, as well as temporal information together, could help in modeling the interaction signals. We then propose a histogram-based feature representation which can encode all the information at the same time. In order to investigate the proper feature representation for user intention prediction, we conduct an experiment to evaluate the intention prediction performance of using different feature representations. Meanwhile, to bridge the gap between the limited approach of modeling individual modality and modeling multiple interaction modalities, we conduct an experiment to explore the performance of the prediction models with different ways of fusing multiple modalities. The experimental results show that our proposed histogram-based

feature representation is more useful for user interaction intention prediction. Besides, the study also indicates that the performance of the prediction model can be improved by modeling multiple modalities together.

The ability of predicting user's interaction intention could be applied to enhance user experience in multiple ways. One potential application of the intention prediction model is to detect user slips. If the actual coming activity is different from the predicted activity, then the actual activity may be triggered by a slip. The system may need to take some action to help the user to correct the mistake. In this study, we also conduct a pilot study to investigate the performance of user slips detection by applying our multimodal intention prediction model.

1.2.3 Monitoring Human Physiological Signals Non-intrusively

In addition to interaction and body motion signals, physiological signals may also represent user intention, even for the implicit intention that the user is unaware of, and even cannot control or hide. Therefore, exploiting human physiological signals could be useful for user intention prediction. However, there are some drawbacks in the traditional manners of measuring physiological signals. First, these physiological signals measurements rely on expensive devices such as Mindset [79] and EPOC [26]. Though these devices might provide a higher measuring accuracy, they are not commonly-found equipment and not affordable and accessible to common users. Second, users need to wear these devices or attach them to the body during the measurement. Therefore, they are intrusive, imposing a burden on the user and even potentially affecting user behavior or emotion.

It would be better if we can measure users' physiological signals without the user feeling of the existence of the measuring devices. To address this challenge, one of the potential solutions is to utilize a standard device of personal computers.

Therefore, in this study we enhance the daily-used mouse by attaching a low-cost LED and a light sensor, to build the physiological mouse prototype and measure the photoplethysmographic (*PPG*) signal [3] in a non-intrusive manner. We then measure the infrared light which is emitted from the LED and reflected off users' skin, and then process those raw signals into appropriate physiological signals, specifically, the heart beat rate and respiratory rate. The mouse is virtually available in all computers, and the attached LED, as well as light sensor, are low-cost and common devices, they are all affordable and accessible to common users. Meanwhile, physiological signals can be captured while the mouse is held, without the user being consciously being aware of the measuring devices.

Our evaluation shows that the physiological mouse can accurately measure users' heart beat rate and respiratory rate. Going one more step, we further conduct a pilot study to investigate the relationship between measured physiological signals and human emotions. Experiment results show that physiological signals captured by the physiological mouse could be another potential modality to detect human emotion and it might also be applicable to user intention prediction and enhance user experience in the future.

1.3 Thesis Aims and Outline

The aims of this thesis, as outlined in the study overview, are as follows:

- To propose low computation approaches based on motion analysis to automatically detect fights and discriminate fights with real fight intention against simulated fights.
- To investigate cross-species learning in human real fight detection with a set of low-cost and effective local motion features.

- To collect a human real fight dataset and an animal fight dataset for fight detection models evaluation and cross-species learning.
- To collect a user interaction intention dataset in web search task, with non-intrusive devices.
- To propose the user intention prediction model with multiple interaction and body signals as well as history activity sequence. The proposed model can be further applied to detect user slips in mouse clicks.
- To design and build a prototype for the novel physiological mouse, which can measure users' physiological signals during daily computer interaction tasks in a non-intrusive way.

The remaining chapters of this thesis will cover the following:

Chapter 2 provides the literature reviews on the research works about human interaction and body signals analysis. More specifically, it covers related prior research studies related to human fight detection, user intention prediction, mouse behavior analysis, gaze behavior analysis, body motion analysis, as well as physiological signals analysis.

Chapter 3 presents two low computation motion analysis-based approaches as well as a cross-species learning technique for human fight detection. This study shows that the proposed approaches could accurately detect human real fights. Our model is even able to distinguish fights with real fight intention from simulated fights, that indirectly indicates that body motion is a potential modality to reveal underlying intention behind human action and it can be used to predict user intention.

Chapter 4 describes the tasks and experiment details of the collected user interaction intention datasets as well as presents the multimodal approach to predict user interaction intention in daily common computer interaction tasks. This

study explores the performance of the intention prediction models with different ways to fuse the different modalities. The evaluation results indicate that the proposed approach is useful to predict user interaction intention. It also demonstrates a potential application with our intention prediction model to detect user slips click.

Chapter 5 explores the performance of measuring user's physiological signals, specifically, heart beat rate and respiratory rate, by the physiological mouse. A pilot study on the relationship between measured physiological signals and human emotions are presented as well. This study shows another potential modality to predict user interaction intention with non-intrusive devices.

Chapter 6 summarizes the contributions and limitations of this thesis and the potential future work. This chapter also introduces other contributions I have made that are related to the scope of this thesis.

Chapter 2 Literature Review

This chapter begins with a review of the literature on the approaches of fights and aggressive interaction detection. In order to gain a better understanding of cross-species learning, which is an important technique that we adopt to address the problem of limitation of dataset availability, we therefore review the studies about animal action recognition and transfer learning as the basis of cross-species learning. This chapter also presents the studies about animal action recognition and transfer learning. These are followed by the review of user interaction intention studies. Finally, this chapter presents the studies that contribute to interaction and body signals analysis. The purpose of this chapter is to provide an understanding of the prior research in fight detection, human intention detection as well as interaction and body signals analysis. Based on that, this chapter outlines the rationales for the proposed studies.

2.1 Human Fights and Aggressive Behavior Detection

Researchers have made much effort to investigate fight, violence or aggression detection. In the work of [18] and [33], a violence detector was built by using audio features. In practice, this approach is not quite viable, since real life fight or violence is often captured by video surveillance systems without audio recording.

Nievas et al. [81] attempted to detect violence from videos with visual features, and they introduced two video datasets that contain fight events for evaluation. In each of the two datasets, the videos were equally divided into two groups, and were annotated as “fights” or “non-fights”. They computed two video spatio-temporal motion descriptors, namely, Space-Time Interest Points (STIP) [17] and Motion SIFT (MoSIFT) [60] to generate a set of visual words, which can

incorporate local motion information into appearance features. After obtaining the visual words, they then applied the Bag-of-Words (BoW) [21] approach to represent each video as a histogram over the visual words for further classification. Their approach can yield a relatively good performance of detecting violence, and they proved that human fights and aggressive behavior can be automatically detected from some video surveillance scenarios. However, their approach is quite complicated, with high computational demand.

Yang et al. [114] attempted to adopt a rule-based detection approach to implement automatic detection aggression inside a train. In their approach, they generated some low level features from the raw video data including energy signatures and motion paths, which were combined into high level concepts in the rule-based model for aggression detection. This approach was limited to an environment inside a train and yet required much computation. Hassner et al. [37] proposed the Violence Flows (ViF) descriptor to investigate Violence in crowded scenes. In their work, they focus on measuring how the motion magnitudes change over time. By comparing magnitudes, they can measure the significance of observed motion magnitudes in each frame compared to its predecessor. Changes in motion magnitude higher than a threshold are accumulated for each pixel in building a histogram for the local regions in the video frames, to create the feature vector for fight detection. Gao et al. [31] then proposed the OViF features to detect violence events by further considering motion orientations in the ViF. In addition, Improved Fisher Vectors (IFV) [86] that delineates local features by their deviation from the generative Gaussian mixture model was also suggested to be useful for violence detection. There are also some other works trying to detect human aggression behaviors and activity patterns [39, 117].

Besides the hand-crafted motion feature representations, recently, deep

learning approaches have been very successful in action recognition, and it can also be applied in fight detection. Wang et al. [107] proposed the Temporal Segment Networks (TSN) structure to perform deep action recognition. Their network achieved promising performance in many human action datasets such as UCF101 [95]. The TSN network [107] can be used to build a "FightNet" [122] to perform fight detection in deep learning. However, this network is quite complicated and demands a lot of training time and training data. Furthermore, it has only been evaluated in a subset of UCF101 with little fight scenarios, such as boxing game, but not on real fight events. Most of the publicly available human fights or aggressive actions datasets are not collected from real life surveillance. Some of these datasets are collected from specific sport games such as ice hockey [81]. Some are collected in simulated scenarios, where the fight events were acted by some subjects [10, 29]. These datasets may not reflect the fighting actions and intentions in real fight scenarios.

In addition to detecting human fight actions, there are some other related studies attempting to study another aggressive social interaction: fierce argument. Predicting the conflict level of argument in conversations and debates could be useful in real life. To study conflict level detection in debates, Kim et al. [55] constructed a conflict detection dataset, based on the videos from some televised political debates [101], which contain also the audio channel. For each video in this dataset, they annotated a numeric conflict level. They then tried to predict continuous conflict level in political debates. In their study, they applied prosodic and conversational based features to detect the conflict level of those debate videos. Their experiment results suggested that the prosodic and conversational based features can be very useful for conflict predicting in the debate scenarios. However, their approach is not applicable to the surveillance scenarios, in which the audio

channel is not accessible.

2.2 Animal Action Recognition

This thesis involves the study of cross-species learning in fight detection. In order to gain a better understanding of cross-species learning, we review the studies about animal action recognition and transfer learning. There are some interesting studies on animal action recognition. Lu et al. [67] tried to estimate sheep pain level from the facial units of sheep. Burgos et al.[13] proposed a novel method to automatically segment and recognize social behaviors of mice, such as approaching, attacking, cleaning, and walking away. Mazur et al. [73] investigated automatic analysis of the aggressive behaviors of laboratory mice through thermal video processing. They recorded images of thermal sequences, which allowed them to track the mice. Corner detection technique was used for temperature analysis of mice to identify their aggressive behaviors such as biting. Ladha et al. [57] utilized a wearable sensor (accelerometer) to recognize the actions posed by dogs. Their study aimed to monitor and track the health and wellbeing of dogs. In the same spirit, Iwashita et al. [46] mounted a camera on the back of dogs and recognized dog actions from the first-person animal videos. Most interestingly to us, Wang et al. [105] studied cross agents action recognition, based on transfer learning across different agents including different groups of people and species. However, their method does not generalize well for cross-species fight detection.

2.3 Transfer Learning

Transfer learning can be broadly categorized into instance-, feature-, and model-based methods [83]. Instance-based transfer learns from samples that minimize the difference between target and source distributions, where adaptation can be conducted from either a single source [22] or multiple sources [115].

Feature-based transfer aligns the source and target distributions by finding a new feature space through, for instance, domain-guided regularization [111], transfer component analysis [82], and maximum independence domain adaptation [113]. Model-based transfer exploits the pre-trained classifiers to build an adaptive classifier in the target domain, using such as domain-dependent regularization [24], different domain metrics [44], or parameter fine-tuning [116]. Despite the success of transfer learning in addressing the data limitation issue, only little attention has been paid to the computational cost of these methods and their practicability in real use. Sangineto et al. [89] mitigated the computational cost of a personalized model, using a regression function to determine the target parameters. Similarly, Zen et al. [118] used support vectors to achieve parameter transfer without model retraining. The closest work to this study is from Huang et al. [44]. In their approach, they proposed to reorganize the pre-trained weak source classifiers based on a domain metric. However, their work leveraged the person identity information which is agnostic in our case. This makes our problem more challenging.

2.4 User Interaction Intention Prediction

Our study about user interaction intention is related to the work about user intention detection as well as the application of user errors detection in daily computer interaction tasks.

2.4.1 User Intention Detection

In order to improve user experience, researchers are recently interested in detecting user intention and predicting what a user wants to do, when the user is interacting with a robot [20, 50, 99], mobile device [36, 78, 96], and vision-based interface [90, 103], etc. For instance, Kato et al. [50] attempted to predict the next

step of human behaviors by modeling human body movements and gestures for human-robot communication. Negulescu et al. [78] tried to predict user intention on mobile phones based on detecting the grip type of the phone. In the work of [90], Schwarz et al. attempted to detect whether a user gesture or movement is intended to control or not in vision-based interaction by modeling the user's gesture type.

When it comes to interacting with personal computers in daily interaction tasks, some works attempted to understand user behavior and the underlying intent. For instance, Toker et al. [97] focused on real time user skill level detection by modeling eye gaze interaction. Hu, et al. [40] tried to understand the underlying intention behind user's query, while Mandayam, et al. [69] demonstrates that understanding users' intent could be helpful for estimating relevance documents. Besides, some of these works tried to predict users' visual attention by modeling information available to the interface such as mouse interaction [112], as well as tracking and modeling eye gaze interaction [72]. However, to our best knowledge, few works attempted to predict users' next interaction activity in daily multi-step tasks. Most of these works attempted to investigate user intention in the scenario of prescribed trials. For instance, some of these works are interested in predicting the target point of a mouse movement [5, 85, 124]. They conducted their studies in a simple task of pointing mouse to some specified targets. Among these works, Pasqual et al. [85] achieved the state-of-the-art. In their study, they summarized a mouse movement velocity profile and model its velocity time series as a 2D stroke gesture. By doing this, they can record the velocity profile from prior mouse movements as templates and can find the most likely template for the testing mouse movement in predicting the mouse endpoint, by applying the template matching approach. User interaction intention in real multi-step interaction tasks

was not investigated in these works.

Besides, some of these works focus on exploring user intention prediction in browsing activities. For instance, in the study of [70], Maniu et al. investigated user's search behavior on photo sharing platforms. They attempted to use historical search sessions and query types to predict the next URL class that the user is going to visit. In the work of [35], Guo et al. attempted to model user's history query, page content and mouse interaction in the context of e-commerce websites to predict whether a user is ready to buy or just browsing. However, these methods relied on extracting features from browsing context, such as the historical search query. Thus, it is difficult to generalize these approaches to other daily computer interaction tasks.

The works of Alexander et al. [2] and Fitchett et al. [30] are kind of studying on predicting user intention in real daily interaction tasks. In their studies, they noticed that users will frequently return to previously visited regions within their document, during reading tasks. They then tried to investigate this revisit intention further to predict the possible revisit document regions by modeling historical reading records. Based on the prediction model, they then designed a scroll bar which can help users to easily select and jump to the possible revisit document regions. Their user study found that user experience can be improved by understanding user's revisit intent. However, this study only focused on the reading task which is relatively more restricted and has relatively fewer types of intention compared to our web search task. Fitchett et al. [30] also studied on predicting re-visitations. They extended the revising contents to file accesses, website visits, window switches, etc. However, they still utilized log records only. Evans et al. [27] also focus their study on daily computer usage. In their study, they attempted to build a tool to automatically segment text entry and mouse pointing input

streams occurring in daily computer usage into “trials”. However, they did not try to model these interaction signals for further detecting user’s next interaction activity. To our best knowledge, user intention prediction via modeling multiple interaction patterns is not well investigated, especially in the scenarios of daily computer usage.

2.4.2 User Errors Detection

Human-machine interaction experience could be hampered by different types of errors. Errors detection thus has gained attention in recent studies in various scopes. Vandewynckel et al. [100] tried to perform real time activity error detection for Alzheimer's patients with accelerometer. In their study, they defined the deviation from the regular movement as an error. Jiang et al. [49] studied on whether a user is skipping a relevant search result without clicking on it in the web search tasks. They defined this action as skip error. Besides, Lin et al. [65] tried to predict user's errors in the task of numerical typing, based on the EEG signal. However, in our work, we are interested in user slips in daily computer usage, which defined as selecting the wrong target or triggering a wrong action with the right intention.

To understand user errors in target pointing and selecting tasks, some studies attempted to predict error rates in target selecting [62, 110]. Some of these works investigated the trade-off between the mouse moving speed and target selecting accuracy [123]. These studies suggested users may adjust their mouse moving behavior to reduce clicking errors. Banovic et al. [8] further investigated the time cost associated with user selection errors and attempted to predict task completion time when the cost of error is involved. Their study also indicates that the user selection error would induce additional time cost, such as the time for recovering from the error and new target navigation. These kinds of additional time cost

would affect task completion time as well as user experience.

These studies about user errors all focused on mouse pointing and selecting tasks. However, they only studied in prescribed mouse pointing trials, which only required subject to select some particular targets. Moreover, these works only utilize mouse movement profiles to study target selection errors. To the best of our knowledge, we are the first in investigating multimodal intention-based user selection slips detection in the scenario of daily computer usage. In our study, we therefore focused on studying user interaction intention in common daily computer interaction tasks, by using multimodal interaction and body signals.

2.5 Interaction and Body Signals Analysis

The studies in this thesis involve the analysis of multiple interaction signals such as mouse and gaze interactions, and body motion signals, such as head movements and body motions, as well as physiological signals. We therefore review the related studies about these interaction and body signals.

2.5.1 Interaction Signals

Traditional human-computer interaction involves three common KVM devices: the keyboard and mouse for input, and the screen for output. Users can use the mouse to achieve a lot of tasks, such as selecting text, clicking buttons to trigger events, etc. These actions generate a good volume of information highlighting the interaction from human to computer and previous studies suggested that mouse activities dominate interaction in daily computer use [15, 74]. Due to the importance of mouse interaction, many researchers have worked on analyzing and modeling information about user behavior and intention from recent mouse interaction data in recent. For instance, some prior studies utilized mouse interaction to detect user's search attention [58], and predict users' eye gaze

position [42], etc. Some recent work tried to model mouse interaction information to detect the quality of crowdsourcing workers [75], and users' engagement [4]. In the study of [43], Huang et al. attempted to detect users' stress level by modeling mouse movements. Meanwhile, some other studies suggested that different users may behave differently in terms of mouse movements, when they are interacting with some daily computer tasks. These studies then attempted to perform user authentication by modeling mouse behaviors [19, 121].

In addition to mouse interaction, users' gaze interaction is also a fundamental interaction signal in the scenario of daily computer interaction tasks. Since the gaze positions and movements may indicate the attention contents and intended targets of a user. There are some related studies about user gaze interaction analysis in computer interaction tasks. For instance, in the work of [64], Li et al. attempted to correlate users' reading attention with gaze interactions. In the work of [93], the authors tried to detect users click intention in web pages, by modeling eye gaze behaviors combined with EEG signals. While in [23], Debnath et al. attempted to detect drivers' visual focus of attention by utilizing drivers' gaze behaviors. Besides, some recent works attempted to consider both mouse and gaze information in their models. For example, Wang et al. [106] tried to detect users' stress level by modeling mouse and gaze interactions as well as the interface contents together.

Gaze positions and interactions can be obtained by using some eye tracker devices, which are not affordable and accessible for common users in daily computer usage. However, there are some gaze estimation techniques that can estimate gaze positions through standard webcams [7, 42, 120]. With the help of these techniques, we therefore can extract gaze information in daily computer usage in a non-intrusive and low cost manner.

2.5.2 Body Motion Signals

Social interaction pattern study has become a very popular topic recently in human-computer interaction and affective computing. Motion analysis is an important approach for analyzing the interaction patterns in the context of social signal processing. For instance, in the work of [88], Ramseyer et al. proposed an approach based on frame differencing and motion energy analysis for synchronized movement in social interaction. While in [80], Nguyen et al. attempted to use a mixture of body communicative features to predict the personality and job interview rating of employees in the employment interview scenarios. Some of their features were based on motion analysis. They computed the dense optical flow map for each video frame, and then extracted the hand likelihood map by assuming that the hands are the fastest moving part in the optical flow map. They extracted the hand velocity and acceleration as features for predicting, after obtaining the hand likelihood map.

The approach of motion analysis can also be applied in the area of analyzing audience behavior. For instance, in [77], Navarathna et al. utilized the face and body motions features generated by motion history images to learn and represent the individual and group behaviors of audience in watch movie scenarios. Then they used these representations to predict movie ratings. Motion analysis can be also be applied in group behavior analysis, such as predicting dominance in group conversation [48], extracting hand position and communicative cues in conversations [71]. These recent works demonstrated that motion analysis is an effective technique for extracting social interaction information from video. Since then, motion analysis techniques have been widely used in social signal processing. However, most of these works applied motion analysis for processing videos with good resolution, clear background, and simple actions. It is unclear whether the

motion analysis techniques can work well on the video surveillance scenarios where the video quality is not good enough for human or even face detection, with a relatively noisy background, for instance, in fight detection applications.

2.5.3 Physiological Signals

Prior studies suggest that facial expression, vocal information, hand gesture, body posture, and language, etc. are the most common input factors for human affect inference [14, 119]. In addition to these signals, physiological signals also appear to carry important information related to human emotions and intentions. For instance, some studies attempted to detect happy by analysis EEG signals [47], some tried to study stress by using human physiological signals including galvanic skin response, and heart rate variability, etc. [68], some attempted to detect users' typing errors based on the analysis of EEG signals [65], while some other studies suggested to use physiological signals such as skin conductance level to predict users' actions [61], when they are playing computer games. However, different from the extraction of facial expressions, gestures, and body postures, etc. which are easily captured by webcam or other non-intrusive devices, most current approaches of physiological signals measurements are often intrusive. These approaches require users to have additional sensors attached onto their body for electrocardiogram (heart-related), electromyogram (muscle-related), electroencephalogram (brain-related) signals and so on. For instance, BP@Home system [56] relies on an A&D blood pressure sensor that must be worn by the user. The MobiSense system [104] is capable of returning heart-beat and activity information to a server based on accelerometer and ECG sensor information, but also requires the user to wear sensors on the body.

In this thesis, we therefore study on a non-intrusive approach to measure physiological signals in daily computer tasks. Specifically, we explore the usage

of the photoplethysmographic (*PPG*) signals to extract human heart beat rate and respiratory rate. There are some prior studies working on using non-intrusive video-based methods to measure human physiological signals for health monitoring. For instance, Scully et al. [91] used the video taken by smartphone cameras to determine human heart beat rate in their study. In the work of [1], Ahsan et al. attempted to detect hemoglobin level via fingertip video images which were collected by smartphone cameras. In [6], Balakrishnan et al. attempted to detect a periodic pulse from captured head movement videos via principal component analysis (PCA). Based on the detected pulse, heart beat rate and respiratory rate could be estimated in their study. In [53], independent component analysis (ICA) was adopted to reduce the impact of motion in the captured signal. These techniques, however, are very easily affected by movement of the body and changes in the head orientation. After physiological signals are extracted, one could proceed to associate them with human affects. It is observed that heart rate variability will decrease when human feel fear, sad or happy, while peak heart rate will increase with pleasure [87]. Slow respiration can be regarded as manifestation of relaxed emotion, while irregular rhythm and quick variations correspond to anger or fear [54, 87, 94]. Nevertheless, establishing a good mapping from the collection of physiological signals to human emotions remains an interesting and perhaps open research problem.

2.6 Summary of Related Works

Inspecting from the prior related research, we see some gaps between the existing studies and approaches about understanding human intention in both human-human and human-computer interaction tasks. The constraining issues of the prior studies include studies being conducted only on the predefined tasks or

on simulated applications without making use of all reasonable modalities.

Besides, the literature review on the interaction and body signals suggests the feasibility to capture human interaction and body signals including mouse, gaze, head, body movements as well as physiological signals, in a non-intrusive manner. These studies also suggest that these interaction and body signals can be further applied to interpret human emotions and even intention. The studies described above together provide some practical approaches to further understand human intention in real applications.

Chapter 3 Automatic Fight Detection via Motion Analysis

Selected notations and abbreviations used in this chapter

$A(t)$	motion acceleration in optical flow image t
$A_{lr}(t)$	motion acceleration of local region lr in optical flow image t
C	a set of classifiers for ensemble learning
c_l	a classifier in set C
D_{ij}	Euclidean distance between motion regions i and j
G_{ij}	motion attraction between motion regions i and j
$G(t)$	overall motion attraction among motion regions in optical flow image t
L	number of ensemble classifiers for cross-species learning
lr	a local region in a frame
$M(t)$	motion magnitude of optical flow image t
$M_r(t)$	motion magnitude of motion region r in optical flow image t
$M_{lr}(t)$	motion magnitude of local region lr in optical flow image t
n_f	number of frames in a video
n_c	number of columns for segmenting local motion regions in a frame, default value is 4
n_r	number of rows for segmenting local motion regions in a frame, default value is 4
n_τ	number of optical flow images for type τ in a video
p	a pixel in an optical flow image
r	a motion region in an optical flow image
$R(t)$	set of motion regions in optical flow image t
t	index of current frame
$v_{x,t}$	horizontal component of a motion vector in optical flow image t

$v_{y,t}$	vertical component of a motion vector in optical flow image t
τ	a motion type
δ^S	source training set for cross-species learning
δ^T	target training set for cross-species learning
δ_l	a sub training set to learn classifier c_l
LMF	abbreviation of local motion features
MSF	abbreviation of motion signal features

We start our study of understanding human intention by investigating human-human interaction in social signal processing. We focus on detecting human fight, which is a special social event and interaction. This chapter presents our proposed motion analysis-based automatic fight detection approaches. Compared with data-driven approaches or approaches relying on gesture recognition, etc. the proposed approaches require less computational cost.

Different from the prior fight detection studies, we aim to detect real fight events happening around people in daily life. In order to evaluate the proposed fight detection models in real fight scenarios, we collect and annotate the human real fight dataset. We then present the details of the human real fight dataset, evaluation experiments and results. From the results, we find the appropriate way to model the body motion signals for human real fight detection.

At the same time, we are also interested in revealing the fundamental difference between real fights and simulated fights, as well as understanding real human fight intention. To this end, we conduct experiments to discriminate real fights from simulated fights. This chapter presents these experiments and results.

In addition, in real applications, only a small amount of human fight videos are available for training. That might affect the performance of fight detection models, especially when we build the model by data-driven approaches, such as deep learning. To address this challenge, we propose a cross-species learning approach based on the motion analysis features, to adapt knowledge from animal fights to human real fights. This chapter presents the collected animal fight dataset, cross-species learning approaches, as well as the evaluations.

The rest of this chapter is organized as follows. Section 3.1 describes the motion analysis models that we build and adopt for the purpose of fight detection, as well as the underlying motion features contributing to the models. Section 3.2

introduces the human real fight dataset collected by ourselves, which is involved in our experiments for evaluating the proposed fight detection models and cross-species learning approaches. Section 3.3 describes our evaluation experiments. Section 3.4 presents the study of evaluating proposed fight detection models in simulated fights as well as discriminating real fights from simulated fights. Section 3.5 describes the work of cross-species learning in fight detection. Finally, this chapter is concluded in Section 3.6.

3.1 Fight Detection by Motion Analysis

In this study, we investigate in extracting a series of features based on motion analysis to build human fight detection models, which are resilient to relatively low resolution videos with noise induced by camera movement etc. Under the context of low resolution video with noise, it is often hard to perform action recognition or even human detection. As a result, we refrain from having to recognize the human and the associated body parts, e.g., the arm, the hand or the leg, which would normally be involved in a fight. Instead, we reckon that a fight will be associated with fast moving objects or motion regions. Stationary objects residing in the background would not contribute to a fight event and should not be considered for fight detection at all. We thus aim at detecting the presence of rapidly moving motion regions and their representative trajectories.

Our approaches adopt a two-level statistical aggregation technique to generate the feature sets. We first implement an algorithm based on optical flow images to extract motion information from video clips. We extract the motion pixels and then the motion regions from consecutive frames by computing the optical flow vectors after eliminating the noise. We then extract features from the

motion information in two approaches. In the first approach, we consider the motion signals occurring with different types of motions present, which include motion magnitude, acceleration and a new concept called *motion attraction*. We do not want to rely on the complicated mechanisms of gesture or action recognition to detect different motion types. We therefore simplify the motion type detection by using the nature of motion regions present such as size of the motion region, etc. We then compute motion statistics according to the classified types in the video to generate the feature set for fight detection. We refer to this as the "motion signal features". On the other hand, the locations of fight scenes in video frames may vary across the whole video. Therefore, in our second approach, we consider the motion signals occurring in different locations. In this approach, we propose to extract motion signals including motion magnitude and acceleration, within different regions of video frames. We then compute motion statistics of these motion signals according to the local motion regions as our features, which can effectively encode the spatial motion attributes for fight detection. We refer to this as the "local motion features".

3.1.1 Optical Flow Images

Recall that our goal is to develop an effective algorithm that can deal with low resolution videos and resilient to noise. This would imply that common motion analysis techniques based on action recognition or even human detection could not be adopted. Instead, we need fast and simple but robust motion extraction and analysis approaches. Candidate methods include optical flow, frame difference, etc. In this study, we prefer to adopt optical flow for motion analysis, since in the application of human fight detection, the magnitude of optical flow vector is a very strong cue for measuring the amount of motion and the direction of optical flow vector is able to provide us with more motion information. Our approach is

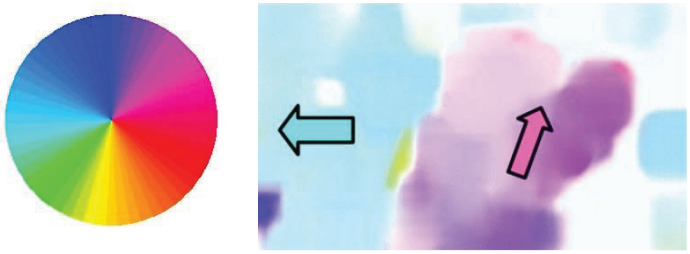
actually based on extracting motion information from optical flow images, and in our approaches, we applied the algorithm described by Farneback, et al. [28] to compute optical flow images between two consecutive frames.

Our motion analysis-based approaches consist of several steps: optical flow computation, noise removal, motion signal extraction and feature extraction. After extracting the motion analysis features, we then adopt machine learning approaches to recognize and classify the video. As illustrated in Figure 3-1, we generate an optical flow image in Figure 3-1 (b) from two consecutive raw frames from a video, such as Figure 3-1 (a). We then try to remove noises from the optical flow image, as shown in Figure 3-1 (c).

(a) (b) (c)

Figure 3-1 Computing optical flow image. (a) original image, (b) optical flow image, (c) noise removal

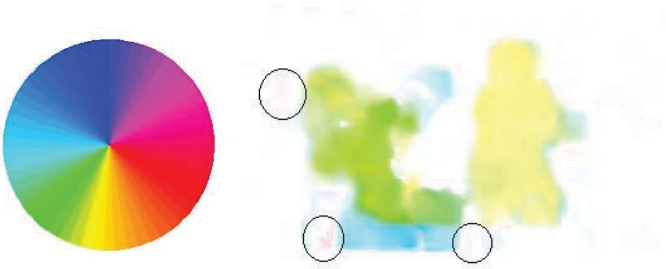
Figure 3-2 demonstrates a sample of optical flow image with camera motion noises. Figure 3-2 (a) indicates the color code scheme for visualizing the optical flow image. The hue of the color code represents the motion direction in Figure 3-2 (b), while the intensity represents the motion magnitude.



(a) (b)

Figure 3-2 Color code scheme (a) and optical flow image (b).

The videos involved in our human real fight dataset are collected from video surveillance. These videos are in bad resolution and contain a lot of background movements or environmental changes, such as environmental light change, etc. These background noises and environmental changes would introduce some noises when we generate the optical flow images. Figure 3-3 (b) shows some raw optical flow image noises, under the color coding scheme in Figure 3-3 (a). Therefore, we need to remove some noise before further processing. We observe that the main noises are introduced by environmental light changes. In our experiments, since most of the videos are collected from the videos produced by relatively stable camera, there is little camera movement happening during the video recording. The main causes of the noise can be attributed to environmental light changes, while these noises usually appear in fragmentary pixels. For this type of noise, we can remove them by filtering the motion regions with very small size. Therefore, we identify and extract connected motion regions from optical flow images, and then compute the size of each connected motion regions in our approach. The regions with very small size which contain few fragmentary pixels are usually the noise caused by light changes. We then remove those regions to remove light changing noises.



(a) (b)

Figure 3-3 Color code scheme (a) and light change in optical flow images (b).

3.1.2 Extracting Motion Signal Features

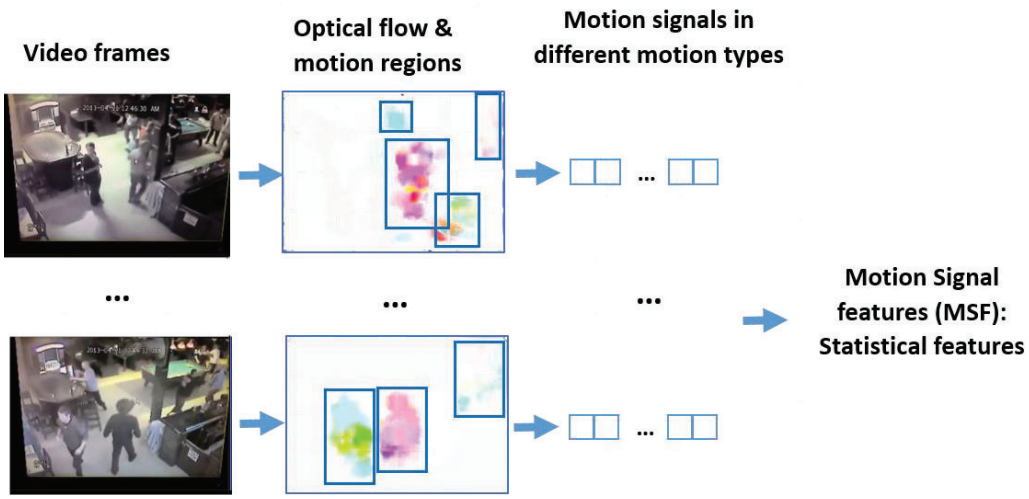


Figure 3-4 Example of motion signal features extraction.

After performing noise removal to produce denoised optical flow images, we can extract and analyze motion signals from these optical flow images. Figure 3-4 illustrates the process of extracting motion signal features. As mentioned before, we would like to implement fight detection without performing behavior, gesture or action recognition. However, there are inherently multiple types of motions present in the raw optical flow images. Thus, in our first approach, we perform some preprocessing to simplify the task of detecting motion type.

Detecting Motion Type

It is intuitive that different types of motions may tell different stories about the underlying motion activities. For instance, an optical flow image that contains one big motion region with big motion magnitude may mean a fast whole body movement, whereas two small motion regions both with big motion magnitudes may well represent two fighting fists. Though the motion magnitudes in both of the two cases are big, they may be manifesting different scenarios, due to the difference in number and size of motion regions. In general, most fighting scenes should involve at least two persons, often characterized by the presence of two or more fast moving regions. To improve the discriminative power of the motion detection algorithm without incurring excessive overhead, we would like to classify all the different possible types of motions into a few types based on the number of motion regions, the average size of regions, their moving directions, etc. Then for each optical flow image, the representative motion type reflecting the whole image is detected.

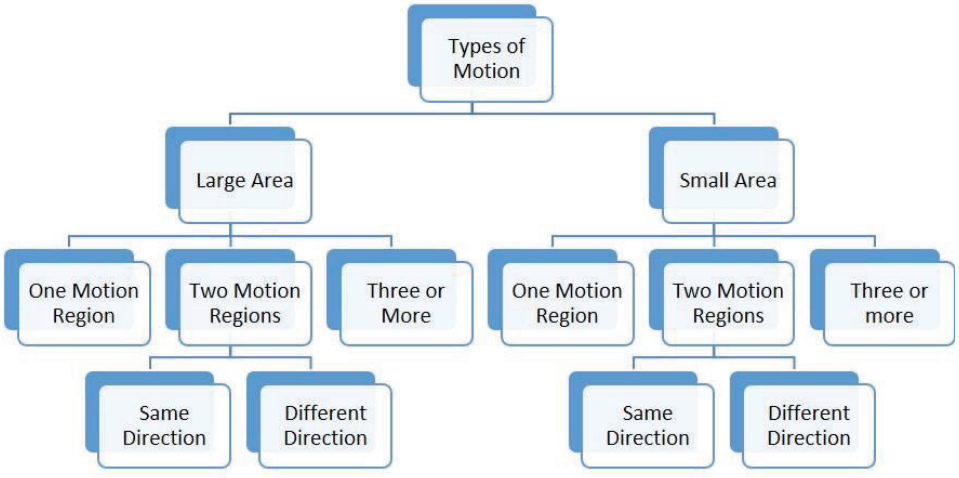


Figure 3-5 Decision tree for classifying motion types.

As illustrated in Figure 3-5, we specifically classify the motion regions into

8 motion types. All these motion types can be determined through the relative size and number of motion regions, as well as the motion directions. To make the detection task more efficient, we adopt a simple decision tree approach, as depicted in Figure 3-5. For instance, for an optical flow image containing two motion regions, we follow the appropriate branch of the decision tree and compare the directions of the two motion regions on whether they are moving in the same or different direction. This is achieved by computing the angle between the directions of the motion regions. If the difference is smaller than 90-degree orientation, the motion is regarded as a "Same Direction" motion; otherwise the motion is regarded as a "Different Direction" motion.

Computing Motion Signals

After computing for the optical flow images, removing noise thereof and classifying the resultant representative motion patterns, we extract the key features for machine learning. There are three representative attributes arising from our motion analysis approach: motion magnitude, motion acceleration, and strength of motion region relationship, collectively known as motion signals in this approach. Motion magnitude reflects the amount of movement observed in an optical flow image. It is a first order quantity. Acceleration measures the rate of change in motion magnitude, which is a second order quantity. We propose a third type of useful motion signal to cater for measuring the strength of the relationship between several motion regions, called *motion region attraction*, which is also a second order quantity. In statistics, distribution is often approximated via some first order and some second order parameters, the most common ones being the mean and the standard deviation. Additional parameters would be needed if the distribution deviates from standard ones, for example, mode or median as a variation to the mean, and the use of range (minimum and maximum) or inter-quartile range.

Occasionally, higher order quantities like third order skewness or fourth order kurtosis may also be considered. However, we demonstrate that our first and second order statistics already suffice in contributing to a good performance.

In this approach, we first compute the motion magnitude from each optical flow image. Each pixel p in optical flow image index with t actually represents an optical flow vector $(v_{x,t}, v_{y,t})$, and the magnitude implied of each pixel can be computed as the length of this vector:$\sqrt{{v_{x,t}}^2 + {v_{y,t}}^2}$. The magnitude $M_r(t)$ of a motion region r is defined as the average magnitude of all the pixels within the motion region, as in Equation 3.1. Let the optical flow image consist of a set of motion regions: $R(t)$. The motion magnitude of that optical flow image is then computed by the sum of all the motion region magnitudes, as in Equation 3.2.

$$M_r(t) = \frac{1}{|r|}\sum_{p \in r}\sqrt{{v_{x,t}}^2 + {v_{y,t}}^2} \tag{3.1}$$

$$M(t) = \sum_{r \in R(t)} M_r(t) \tag{3.2}$$

We next compute the acceleration, as the change in motion magnitude. We would take two consecutive optical flow images and compute the absolute difference of the motion magnitudes between two consecutive optical flow images as the acceleration, as in Equation 3.3, where t and $t-1$ are the indices of the current and previous frame respectively.

$$A(t) = |M(t) - M(t-1)| \tag{3.3}$$

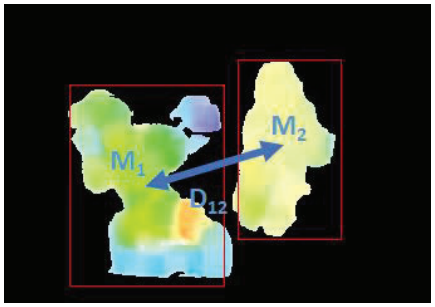


Figure 3-6 Motion attraction computation. M_1 and M_2 are the motion magnitudes of motion region 1 and 2 respectively. While, D_{12} is the Euclidean distance between the centroids of the two motion regions.

We finally propose the notion of motion attraction $G(t)$ among motion regions in an optical flow image to summarize the relationship between different motion regions. If there are more than one motion regions in an optical flow image, we then measure the strength of the relationship between each pair of the motion regions in that image. The attraction between two regions is stronger if the regions are larger, and also stronger if they are closer to each other. We then compute the motion attraction G_{ij} by Equation 3.4, which computes the product of the magnitudes for the two regions, normalized by the distance between the centroids of those regions. i and j are the two motion regions belonging to $R(t)$ and D_{ij} is the Euclidean distance between the centroids of the two motion regions. Figure 3-6 illustrates the example of computing motion attraction. The overall attraction for an optical flow image is the average of all the pairwise attraction values, which can be computed by Equation 3.5, where the number of the total pairs of motion regions is $|R(t)|(|R(t)| - 1)/2$.

$$G_{ij} = \frac{M_i M_j}{D_{ij}} \tag{3.4}$$

$$G(t) = \frac{2}{|R(t)|(|R(t)| - 1)} \sum_{i,j \in R(t) \wedge i<j} G_{ij}(t) \tag{3.5}$$

Extracting Motion Signal Features

After obtaining the motion signals, namely, motion magnitude, motion acceleration, and motion region attraction, we would like to extract motion signal features based on their statistics across all the optical flow images throughout the video, as well as the individual statistics for optical flow images belonging to each of the 8 motion types described before. Starting from a video consisting of n_f frames, we first compute a total of $n_f - 1$ optical flow images, removed of noise. Then we classify the optical flow images into one of the 8 motion types following the decision tree in Figure 3-8. Now there are n_τ images for a motion type τ, and $\sum_\tau n_\tau = n_f - 1$. For the set of optical flow images belonging to each type, we compute common statistics for each of the three motion signals, namely, mean, maximum, minimum, median, and standard deviation. Furthermore, we count as the final feature the percentage of optical images belonging to each motion type, hereby reflecting the distribution of the different motion types. All these features are used in the machine learning algorithm to detect the presence of a fight.

To summarize, Table 3-1 shows the motion signals and statistics that we use to generate useful features. There are 3 motion signals and 5 statistics for each signal. As a result, there are a total of 15 features generated for frames belonging to each motion region type. This number is reduced to only 10 for those with just one motion region (without motion region attraction). Table 3-2 indicates the complete set of features that we generate for each video for machine learning. There are a total of 110 statistical features for the 8 motion types, plus the 8 frame count percentages for each type, and a final set of 15 global statistical features covering the whole video, regardless of the motion type. This results in a potential set of 133 features.

Motion Signals	Statistics
Motion magnitudes	Mean, maximum, minimum, median, standard deviation
Motion accelerations	
Motion region attractions	

Table 3-1 Motion signals and their statistical features.

Region	Statistics
Single	Magnitude / acceleration statistics, count
Multiple	Magnitude / acceleration / attraction statistics, count
Global	Magnitude / acceleration / attraction statistics

Table 3-2 Full set of motion signal features generated for a video.

3.1.3 Extracting Local Motion Features

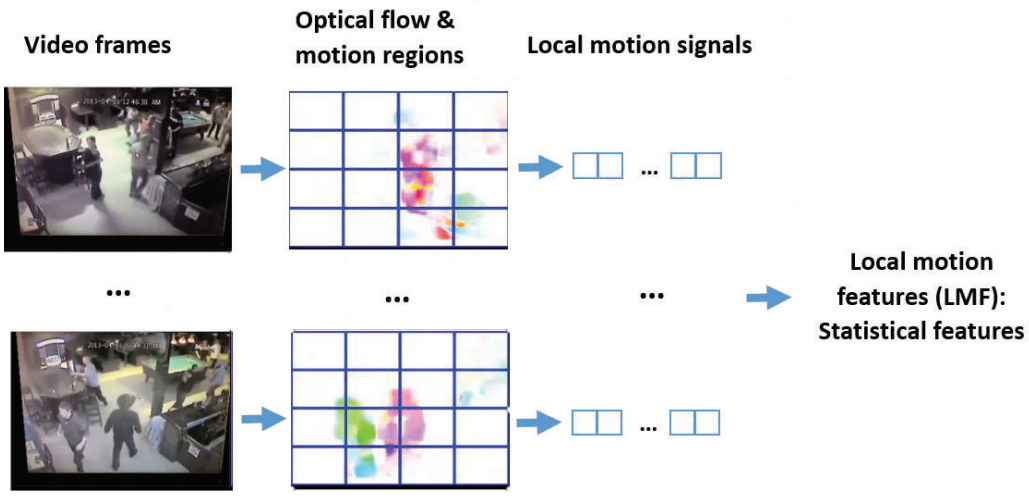


Figure 3-7 Examples of the optical flow image and local motion regions for extracting local motion features.

Extracting Local Motion Sequences

Fight scenes captured in the wild vary very much in nature and manifest with diverse camera locations and orientations. It is not uncommon to see that some

fight motions occupied a large central area of a video frame, while others only took place in a small corner region. Inspired by the spatial robustness of local binary features [37], we propose to extract local motion signals to account for spatial variations of the fight motions across the full image.

To extract the generic motion features that depict signals, we focus on the motion magnitude, i.e. the velocity of the moving object in a pixel-based manner, rather than the moving direction. This is because the direction of an actual fight motion can vary from instance-to-instance, and even more according to the viewing angle of the camera and the location of the people, as many surveillance cameras are installed overhead. To further distill the essential motion attributes for fight detection, we also extract the motion acceleration. The acceleration reflects the suddenness of actions, which generally attains a high value in the real fight actions.

More specifically, given a frame in a video, we first equally segment the video frame into $n_c \times n_r$ regions, where n_c and n_r are the number of columns and rows for segmenting a frame, respectively. In our evaluation, we adopt $n_c = n_r = 4$. After obtaining the optical flow image described above, we then measure and compute the amount of motion magnitude $M_{lr}(t)$ and motion acceleration $A_{lr}(t)$ by:

$$M_{lr}(t) = \frac{1}{|lr|} \sum_{p \in lr} \sqrt{{v_{x,t}}^2 + {v_{y,t}}^2} \tag{3.6}$$

$$A_{lr}(t) = |M_{lr}(t) - M_{lr}(t-1)| \tag{3.7}$$

where t and $t-1$ are the indices of the current and previous frame respectively, each pixel p belongs to local region lr, $(v_{x,t}, v_{y,t})$ represents the motion vector of p in the optical flow image with the index t. $M_{lr}(t)$ represents the average motion magnitude within local region lr. Processing the frames of a video

segment gives us $n_c \times n_r$ sequences of the motion magnitude and motion acceleration. We then consolidate these information to compute for statistical features as the temporal feature representation. We then feed these features into a conventional machine learning module, which is Support Vector Machine (SVM) [52] in our study, to learn a fight detection model. Examples of the optical flow image and local regions for motion features extraction can be found in Figure 3-7.

Extracting Features from Local Motion Sequences

After obtaining the local motion sequences, we analyze their motion patterns based on low-cost hand-crafted features. As shown in Figure 3-7, we measure the motion dynamics using the descriptive statistical features, which is similar to our motion signal features described above.

Our approach extracts temporal features based on human heuristics and adopts traditional machine learning algorithm for fight detection. There are two representative attributes arising from local motion sequences: motion magnitude and acceleration, which we term motion signals. We represent the temporal information of motion signals based on their statistics throughout a video. Similar to the motion signal features, we compute five statistics for each motion signal as shown in Table 3-3. Extracting the five statistics from both sequences of motion magnitude and acceleration gives us $5 \times 2 \times n_c \times n_r = 10 n_c n_r$ statistical features for each video.

Motion Signals	**Statistics**
Local motion magnitudes	Mean, maximum, minimum, median, standard deviation
Local motion accelerations	

Table 3-3 Local motion sequences and statistical features.

3.2 Constructing Human Real Fight Dataset

Figure 3-8 Example of fight scenes from human real fights (first row) and simulated fights (second row).

In order to evaluate the proposed approaches, we need to apply them to a video dataset containing human fight events. However, to our best knowledge, most of the prior related studies investigated fight, violence detection or human actions recognition on the datasets that are collected from sports, movies or simulated scenarios. The fights or actions involved in these datasets either only occur in some specific scenarios, such as hockey games, or are acted by subjects. These kinds of fight events may not truly represent the real situations of fights occurring in real daily life and may not reflect the real human fight intention. In order to study human real fight, we would need to collect the real fight dataset that contains the real fight events in daily life. However, it is difficult to collect real human fights videos by researchers in the laboratory environment. It is impossible or unethical to ask someone to fight for real in front of a camera. Fortunately, there are many videos uploaded to some video websites such as YouTube, including those containing real fight events. These events actually happened around the people involved and therefore represent the real situation of fighting.

We then identify and download real surveillance fighting videos from

YouTube. The videos involve the events of real bar fights, prison fights, street fights, etc. They are all taken from a top angle view by some stable CCTV cameras installed in a specific location, such as a bar, or a post around the street corner. The first row in Figure 3-8 demonstrates some examples of real fight scenarios. These recorded videos contain events happening in this specific location. Thus they contain both the fight scenes and non-fight scenes.

Since each video contains a sequence of fight and non-fight scenes, it would not be appropriate to label the whole video as fighting. Therefore, we partition these videos into several sub-clips and annotate each of the sub-clips to produce our dataset for evaluation. In most public datasets, especially those acted upon by actors, the annotation is frame-based. Each video frame was given a label of an event type such as fight, walk, etc. However, an event in the real situation should contain a series of actions and should last for a period of time. We believe that the frame-based annotation may not be suitable for representing a fight event, due to the overly fine granularity and the high cost of annotation. Instead of annotating based on each frame, we segment the videos into several semantically related sub-clips based on the scenes of event and then annotate for each clip. For our experiments, each of the segmented video clips lasts for about 10 seconds, which we think is long enough to represent an event but is short enough to separate different events.

Based on the scenes of event, our video clips can be categorized into two classes: Fight and Non-Fight. To establish the ground truth for performance evaluation, we annotate these video sub-clips manually. We ask three independent judges to perform the annotation task. Each of them should watch the video sub-clips, and then give each sub-clip a label of either Fight or Non-Fight independently. After they finished with the labeling, we summarize the annotations

and retain only those video clips that receive the same annotation from all the three persons and remove the others, in order to ensure that our annotations represent the real ground truth. Finally, from 19 videos, we semantically partition them into 299 clips. Among them, 266 clips receive the same annotation from all the three independent annotators and are included in our dataset, resulting in a yield of 89%. We then construct our dataset by using these 266 clips, the resolution of which is 320×240. Among these clips, there are 147 clips that contain fight scenes while the remaining 119 clips contain non-fight scenes.

3.3 Evaluating Fight Detection Model in Real Fights

3.3.1 Experiments and Results

We would like to evaluate the performance of our proposed approaches under real surveillance scenarios. As mentioned before, little work had been done to detect fights from real scenarios. In this experiment, we evaluate our approaches based on our collected human real fight dataset described in Section 3.1. For the videos in our dataset, we extracted both the motion signal features and the local motion features as shown in Table 3-2 and Table 3-3 respectively. We then conduct experiments to evaluate the performance of these features.

For each of the approaches, we use the feature vectors generated from videos and apply a classifier to build our fight detection model for classification. In our experiments, we choose Support Vector Machine [52] as our classifier to build fight detection model. For the evaluation, we divide our dataset into training set and test set and train the classifier with the training set and evaluate on the test set. We adopt a standard 10-fold cross-validation in classification performance evaluation. We proceed by dividing the dataset into 10 partitions. We then train the

classifier with 9 partitions and evaluate on the remaining one. This is repeated 10 times for the 10 partitions for testing. We then summarized the overall performance for the evaluated approach.

Classified as / **Ground truth**	Fight	Non-Fight	**CCR**
Fight	119	28	82.7%
Non-Fight	18	101	
Total	137	129	

Table 3-4 Performance of motion signal features on real fights.

We first evaluate the proposed motion signal features (*MSF*). Table 3-4 summarizes the performance of our motion signal features in the form of confusion matrix. We can observe that out of a total of 266 video clips, we are able to correctly classify 119+101 of them. This translates to a correctly classified rate (CCR) of 82.7%. The performance of our motion signal features is quite encouraging. The baseline of this study is 55.3%, by taking the size of the majority class. There is a performance improvement of over 27%. We are interested in those fight events, and report the precision and recall of the events in Table 3-5. It can be seen that for our interested events: fight events, the precision is 86.9%, while the recall is 81.0%. These results show that our fight detection algorithm can be applied in real surveillance scenarios with good accuracy without generating too many false positives.

Class	**Precision**	**Recall**	**F-score**
Fight	86.9%	81.0%	83.8%
Non-Fight	78.3%	84.9%	81.5%

Table 3-5 Precision and recall of motion signal features on real fights.

We also evaluated the local motion features on the dataset. The performance of the local motion features is presented in Table 3-6. According to the table, we can observe that the local motion features can also achieve a promising performance of 87.6% accuracy, which is even better than that of motion signal features. Again, we are interested in those fight events, and would like to report the precision and recall of the events, which are presented in Table 3-7. It can be seen that for the fight events, the precision is 91.3%, while the recall is 85.7%. These results suggested our local motion features can accurately detect human fight events in real applications.

Classified as / **Ground truth**	Fight	Non-Fight	**CCR**
Fight	126	21	87.6%
Non-Fight	12	107	
Total	138	128	

Table 3-6 Performance of local motion features on real fights.

Class	**Precision**	**Recall**	**F-score**
Fight	91.3%	85.7%	88.4%
Non-Fight	83.6%	89.9%	86.6%

Table 3-7 Precision and recall of local motion features on real fights.

3.3.2 Comparison with the State-of-the-Art Approaches

It is encouraging to see that both motion signal features and local motion features give promising results of detecting human real fights. Next, we would like to compare the performance of our proposed approaches with the state-of-the-art approaches. We thus applied the state-of-the-art methods to our human real fight dataset, which include MoSIFT [81], ViF [37], and OViF [31]. Table 3-8 shows

the evaluation CCR of these approaches. We can observe that our local motion features achieve the best performance against its counterparts. Though not as good as the local motion features, the performance of motion signal features can still beat those of the other approaches.

We see that our approaches are promising. To further evaluate their effectiveness against state-of-the-art fight detection, we also compare our methods with FightNet [107, 122]. FightNet has a more complicated network structure than ours, thus requires much longer training time. In this experiment, we train the FightNet on the human fight dataset. We use 90% of the dataset for training and the remaining 10% for testing. We adopt the same data augmentation approach described in [107, 122] to the training set. For a fair comparison, we utilize the same settings in our approach with local motion features. The performance of the FightNet is around 88.5%, while the performance of our approach can also reach around 88.5%. Our model achieves promising performance even compared with the state-of-the-art FightNet. The FightNet model is a complex neural network. If we want to train an effective fight detection model using a normal personal computer, for instance, with a 2.90 GHz CPU, it would take more than one month. However, we only need around 100 milliseconds to train an SVM fight detection model by utilizing our proposed local motion features on the same machine. Our approaches are capable of performing the tasks at a very low cost in terms of computational power demand.

Approach	**CCR**
MoSIFT [81]	74.4%
ViF [37]	80.1%
OViF [31]	81.6%

Motion Signal Features	82.7%
Local Motion Features	**87.6%**

Table 3-8 Evaluations on human real fight dataset.

3.4 Discriminating Real Fights from Simulated Fights

The experiments above demonstrate that our proposed approaches perform well on the human real fight dataset collected by us, which contains the fight events in real surveillance context. After that, we would like to cross-validate it against other commonly adopted datasets for research studies in fight detection. There are several common video datasets which contain fight events. Among them, the BEHAVE dataset [10] and CAVIAR dataset [29] are the most famous ones adopted by prior research works in detecting fight and other human actions. These datasets involve many common human actions such as walking, meeting, running, fighting, etc. However, the actions in these datasets were acted by some subjects. Even though the actions of these datasets are simulated, there are still a lot of research works conducting their studies on these datasets for the detection of different actions. Since they are so well-received, we would like to evaluate our approaches also on these datasets.

3.4.1 Evaluating in Simulated Fights

In our second experiment, we employ the publicly available BEHAVE and CAVIAR datasets to evaluate our algorithm in simulated fight scenarios. These datasets involve many common human actions, including fighting. Similar to the videos that we collect from YouTube, the videos from these two datasets are also taken from a top angle view by a stable camera. The biggest difference between them is that the former's fight actions are real, while the latter's are simulated.

Several subjects were required to act some human actions in generating these two datasets. In addition, as mentioned earlier, our datasets have been annotated based on each sub-clip, while both BEHAVE and CAVIAR datasets were annotated based on each frame. Since we regard a fight event as a series of fight actions, in order to compare the two kinds of datasets fairly, namely, YouTube versus BEHAVE and CAVIAR datasets, we transform the annotations of both BEHAVE and CAVIAR from the frame level to the clip level. We segment the videos in these two datasets into several sub-clips based on their events. Each of the sub-clips lasts for about 10 seconds, and is associated with only one label of either Fight or Non-Fight. Finally, we complete with the simulated fight dataset for our experiment with some post-processing upon the BEHAVE and CAVIAR datasets. The dataset now contains 46 fight video clips and 123 non-fight clips. The second row in Figure 3-8 illustrates some examples of the simulated fight dataset.

We adopt a similar evaluation procedure as in the experiments presented above. We generate a feature vector for each video clip with our proposed features and apply machine-learning algorithm on the generated feature vectors to build the fight detection models from the training set, and then test on the testing set. Same as those experiments, we adopt SVM as the classifier and 10-fold cross-validation for evaluation.

We first evaluate the *MSF*. Table 3-9 and Table 3-10 show the results of the experiment. It seems that the performance of our features is not good as that in real fight detection. The CCR is around 79.9%, only beating the relatively high baseline of 72.8% by around 7%. For the more interested fight events, the precision of this model is around 68.8% and the recall is only about 48%. It can be seen that our algorithm pessimistically classifies most fight instances as non-fight ones.

Classified as / Ground truth	Fight	Non-Fight	CCR
Fight	22	24	79.9%
Non-Fight	10	113	
Total	32	137	

Table 3-9 Performance of motion signal features on simulated fights.

Class	Precision	Recall	F-score
Fight	68.8%	47.8%	56.4%
Non-Fight	82.5%	91.9%	86.9%

Table 3-10 Precision and recall of motion signal features on simulated fights.

We then evaluated the *LMF*. Table 3-11 and Table 3-12 illustrate the results. According to the results, we can also see that the performance of our proposed motion analysis-based features in simulated fight scenarios is not as good as that in real fight detection, though the performance of *LMF* is better than that of *MSF*. For the fight events, the precision and recall are 67.4% and 63.0% respectively. Moreover, the model still tends to classify fight instances as non-fight ones.

Classified as / Ground truth	Fight	Non-Fight	CCR
Fight	29	17	81.7%
Non-Fight	14	109	
Total	43	126	

Table 3-11 Performance of local motion features on simulated fights.

Class	Precision	Recall	F-score
Fight	67.4%	63.0%	65.2%
Non-Fight	86.5%	88.6%	87.5%

Table 3-12 Precision and recall of local motion features on simulated fights.

We then further compare the performance with the other state-of-the-art approaches. According to Table 3-13, we can observe that the two proposed features cannot outperform the other features. The performances of MoSIFT with BoW approach is much better than that in real fight detection.

Approach	CCR
MoSIFT [81]	86.4%
ViF [37]	82.2%
OViF [31]	82.8%
Motion Signal Features	79.9%
Local Motion Features	81.7%

Table 3-13 Evaluations on human simulated fight dataset.

It is interesting to observe that this experiment seems to suggest that our algorithm is not able to precisely classify between simulated fight events and non-fight events. However, in the real fight scenarios in the first experiment, the performance of our algorithm is much better. The two experiments have been run following the same procedure with the same algorithm. Furthermore, the videos of these two experiments are all taken from top view angle by stable cameras. Yet the performance is very different. This can be partially attributed to the much higher baseline in the second experiment, but that is just a minor factor. Since the biggest difference between the two types of datasets is that the fight actions in our dataset are real, while those in the two public datasets are simulated, a plausible

explanation would be due to the nature of the videos, namely, real versus simulated fights.

In this second experiment, our algorithm is not doing as well, and the key ingredient of our algorithm is based on motion analysis. It is not difficult to see that the acceleration feature for real and simulated fights would be quite different. In real fights, the "force" exerted by the fighting parties would be large. The laws of physics imply a high acceleration, and then when the target is hit, a high deceleration is experienced, to bring in a stronger impact on the body being hit. In simulated fights, there is no intention for anyone to hurt anybody, so that the real "force" exerted will be small, leading to a small acceleration, and in general, a lower velocity and slower motion. In many Kung-Fu or martial art movies, the scenes are often shot at a frame rate of 16 per second, and played back at the normal rate of 24 per second, implying an automatic increase in perceived velocity to make them look more real. These differences in velocity and acceleration bring in much impact to our algorithm, whereas there would be less impact on research works not based on motion analysis.

A related question with our motion analysis-based approaches is: does it matter if we cannot detect fights effectively in the simulated scenarios? Recall that our goal is to detect real fights and understand human real fight intention in a possible surveillance application. A simulated fight may just reflect some playing acts by children in real life, which may not reflect a real fight intention of human. Thus, it is not really necessary for us to identify those simulated fight events as real fight events. Rather, they are semantically non-fight events in real world setting, and may not represent the real fight intention. We thus transform the problem a little bit: can we detect real fight events against simulated fight events when they both occur in the same dataset? Next, we will conduct another set of

experiments to study this issue.

3.4.2 Manual Detection

Simulated fight events are different from real fight events, and they may not represent the human real fight intention. In order to understand more about the difference between real and simulated fight events as well as human real fight intention, we design two experiments for discriminating them. In the first experiment, we invited some human judges to watch and label some selected real fight and simulated fight video clips. They need to label each video clip as fight or non-fight. From this experiment, we want to see whether human can tell apart real fights from simulated fights, and whether human will regard simulated fight events as fights or not. In addition, we also tried to evaluate our approaches to see whether machine can tell apart real fights from simulated fights. Both experiments would help us understand more about real and simulated fight events as well as human real fight intention.

First, we design a manual detection experiment. To avoid bias, we recruit 10 judges to participate in this manual detection experiment. Besides, the three persons who helped to annotate videos downloaded from YouTube for our real fight dataset are excluded in this experiment to avoid any bias. Since we have hundreds of real and simulated fight video clips, requiring all the judges to perform a comprehensive manual detection task for each clip is very time consuming. Therefore, in this experiment, we only pick some real fight video clips from our real fight dataset and some simulated fight video clips from the BEHAVE and CAVIAR datasets to perform the manual detection. More speicificaly, we group all the real and simulated fight clips into 18 and 13 groups respectively, based on the variety of fight scenes and environment (e.g., bar fight, street fight, with different number of persons involved). We then randomly pick one clip from each

group to cover the varieties. Finally, we get 18 real fight clips and 13 simulated fight clips for this manual detection experiment.

Before the experiment, each subject was told that there two types of videos: with fight events and without fight events. They were required to watch the video clips and determine the type for each clip. They had to make the judgment by themselves and they can only choose one type of either fight or non-fight for each clip. In order to give judges enough time and sufficient information to make their judgment, there was no time limit when they participated in this experiment. Furthermore, they were allowed to watch the clips over and over again if desired, but for each clip they only had one chance to label.

Since there are 18 real fight clips and 13 simulated fight clips, and 10 judges participating in the experiment, we finally receive 310 labels, with an expected number of 180 labels for real fight clips coming from YouTube and 130 labels for the simulated fight clips coming from the BEHAVE and CAVIAR datasets. The results are shown in Table 3-14.

Detected as / **Fight scenarios**	Fight	Non-Fight	**Total**
Real fight	152	28	180
Simulated fight	18	112	130
Total	170	140	310

Table 3-14 Manual detection on discriminating real fights from simulated fights.

According to the table, there are 152, about 84.4%, of real fight clips correctly labeled as Fight, whereas 112, around 86.2% of simulated fight clips are correctly labeled as Non-Fight. This translates to an overall CCR of 85.2% by human. In general, human can tell apart real fight events from simulated fight events

reasonably well. We would like to know the comparative performance by our algorithm on its ability to distinguish the events against human. It is interesting to notice that there are still a good number of cases (14.8%) that even human can mistakenly classify. When the mistakes are considered, human seems to tend to consider more real fight events (28) as non-fight events than to consider simulated fight events as fight events (18).

3.4.3 Machine Detection

In order to determine whether our algorithm can distinguish real fight events from simulated fight events, we select all the real fight videos and simulated fight videos from our dataset and all the simulated fight videos from BEHAVE and CAVIAR datasets to form a new dataset. We apply the same evaluation procedure on this new dataset as in our first two experiments. We report the average correct classi_cation rate as well as the precision and recall of fight events in Table 3-15 and Table 3-16, with a baseline of 76.2%.

Classified as / **Ground truth**	Real fight	Simulated fight	**Total**
Real fight	145	2	147
Simulated fight	16	30	46
Total	161	32	193

Table 3-15 Performance of motion signal features on discriminating real fights from simulated fights.

Classified as / **Ground truth**	Real fight	Simulated fight	**Total**
Real fight	142	5	147
Simulated fight	18	28	46

Total	162	31	193

Table 3-16 Performance of local motion features on discriminating real fights from simulated fights.

From the tables, we can observe that our proposed features can accurately tell apart real fight events from simulated fight events, attaining an average accuracy around 90%, and 88% for the *MSF* and *LMF* respectively. It is interesting to note that the performance of machine detection for discriminating real fights from simulated fights is even better than that of human classification. Since our algorithm is based on motion analysis, the results give us the insights that the motion signals between real fight events and simulated fight events are largely different and machine can distinguish these two types of events very well based on their motion information. On the contrary, most existing works capable of identifying simulated fight events as fight events would experience difficulty to differentiate between real and simulated (perhaps child play) fights. That could have imposed some limitation on their applicability in real world surveillance scenarios.

3.5 Cross-species Learning in Fight Detection

It is encouraging to see that our proposed approaches can perform well in detecting human fights in real scenarios. Moreover, they can tell the difference between real human fights with simulated fights, which in turn can help us understand more about human real fight intention. At the same time, our model can achieve promising performance even compared with the state-of-the-art FightNet, which however may be limited by the small amount of training data. A large amount of well-annotated data is required for building a well-performing FightNet. However, it is difficult or even impractical to collect a large amount of

human real fight videos. Adapting knowledge from similar data sources can be an effective solution. However, our experiments demonstrate that simulated fights do exhibit fundamental different behaviors from real fighting scenarios, indicating that simulated fights generalized to real fights poorly. An alternative source is using real fight videos, but not by human. Interestingly, there are a good amount of animal fight videos on the web. And there are some similarities between human and animal fighting actions. We therefore propose to explore our study further on cross-species fight detection, being also curious about whether human are fighting like animals.

3.5.1 Source Datasets

This study aims to adapt useful knowledge from similar source subsets to learning human fights. We noticed that human real fights and animal fights may share some intrinsic similarity. We therefore are interested in adapting knowledge from animal fights to learning human fights. In addition, there are some other public datasets containing human fights in other scenarios, which are more similar to real human fights than the simulated fight dataset. Therefore, our experiments attempt to evaluate the performance with adaptation from animal fights and human fights in other scenarios to that of real human fights.

Our experiments were conducted based on 4 datasets. The target dataset is the human real fight dataset described in Section 3.1, which is collected by ourselves. The source datasets of this study involve a public dataset consisting of human fights in hockey games [81], a public dataset capturing fight scenes from action movies [81] and finally an animal fight dataset collected and annotated by ourselves. Figure 3-9 depicts some examples of these datasets.

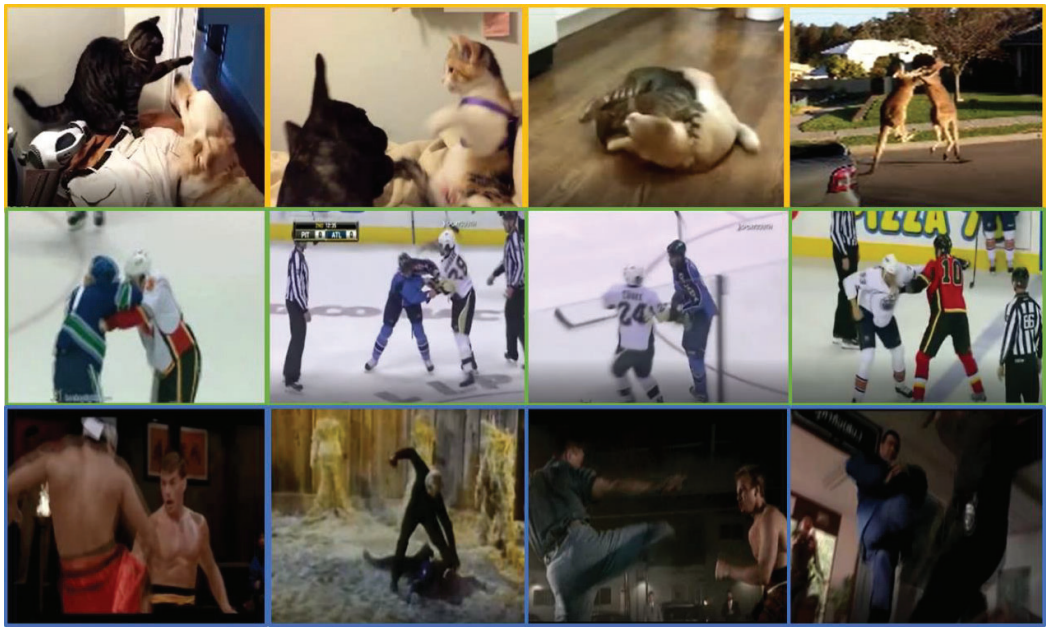


Figure 3-9 Example fight scenes from animal fights (first row), hockey fights (second row) and action movies (third row).

Animal Fight Dataset

In order to investigate fight recognition for animals and the cross-species learning in fight recognition, we need a video dataset of real animal fights. To our best knowledge, there was no available annotated video set containing animal fight events. As a result, we proceeded to collect our own animal fight dataset. We first found and downloaded videos containing animal fights from YouTube. The videos we found involve the events of real animal fights, such as dogs fighting with cats. The first row in Figure 3-9 depicts examples of animal fight scenarios.

After getting these videos, we segmented each collected video into several sub-clips and annotated them for our evaluation. Similar to real human fights, a fight event in the situation of real animal fights should be composed of a series of actions and last for a period of time. Considering that, instead of annotating based on each frame, we segmented the videos into short sub-clips based on the scenes

of event and then annotated for each of them.

Again, we categorized these video slips into Fight and Non-Fight, based on the event occurred within. We then recruited some judges to annotated these clips in order to establish the ground truth for evaluation. In the annotation task, judges were required to watch these video clips independently, and label each of them as either Fight or Non-Fight. There were three judges participated in this annotation task. When the three of them finished their labeling, we summarized all their annotations and only preserved those video clips that received the same annotation from all the three persons, but removed others receiving inconsistent annotations, to ensure that our annotations are close to ground truth. Finally, we got 206 clips, which received the same annotation from all the annotators. Among them, there are 111 clips that contain fight scenes while 95 clips contain non-fight scenes.

Public Available Datasets

In addition to animal fights, there are also some fight datasets capturing fights from other sources like sport events and action movies. In order to investigate the effect of different source data on transfer learning in real human fight detection, we need to also evaluate the transfer learning performance from other source domain. Nievas et al. [81] published two datasets capturing fights from hockey events and action movies separately, which are closer to real human fights compared with the simulated fight dataset. We therefore also adopted their datasets in our study. The second row and third row of Figure 3-9 show samples of the two datasets, respectively.

3.5.2 Ensemble-based Adaptation

Collecting a large number of human fights or aggressive videos is challenging. We observed some intrinsic commonalities of human and animal fighting actions,

such as moving amplitude and acceleration, which fits naturally into our motion analysis-based recognition approach. We believe that investigation of cross-species fight detection would be highly beneficial to this application domain, namely, fight detection.

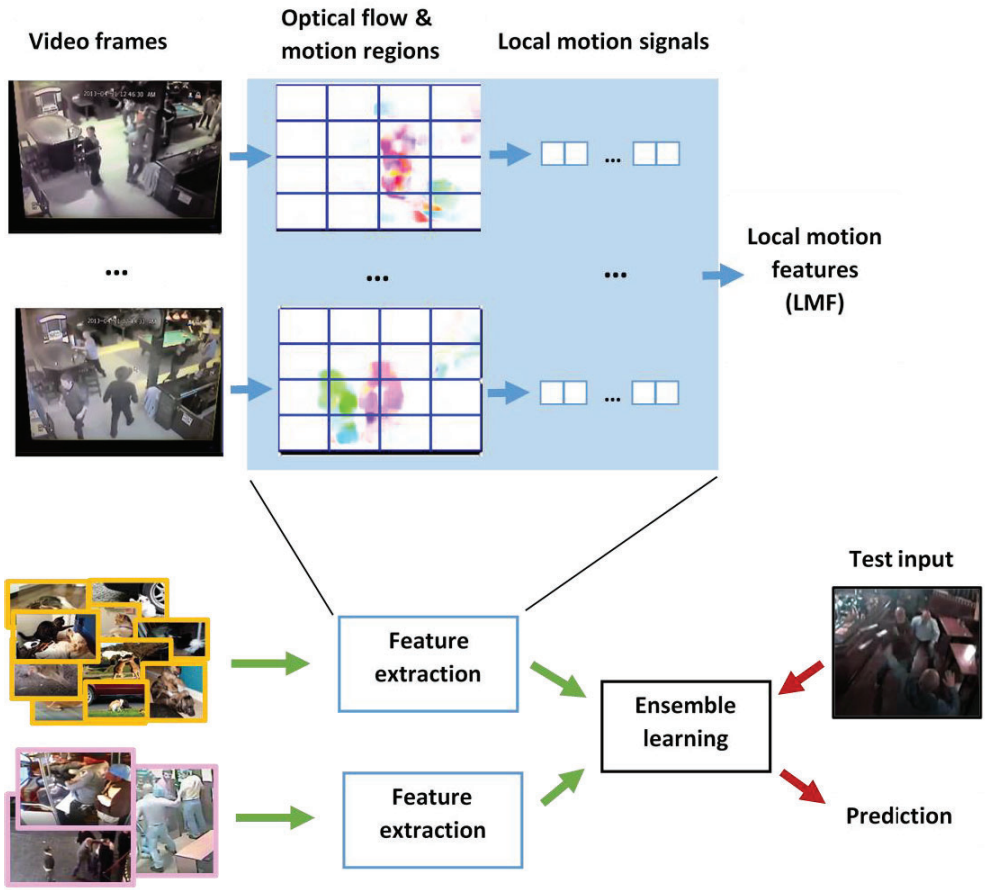


Figure 3-10 System architecture. Our cross-species learning is achieved through ensemble learning. Taking local motion features as an example, we extracted features from animal fight videos and a small amount of available human fight videos for ensemble learning.

For cross-species fight detection, the training data set is normally from one species, and the test data set from another. To address the original research problem, we mainly investigate training on animal fights and testing on real human fights. In practice, if merely a modest amount of target data is available compared with that of source data, directly training on the combined set of source and target data

generally results in unsatisfactory accuracy, mainly due to the domination of the source data. This is the exact issue we need to solve, as the data amount of the target species is limited in our problem setting.

Input: source training set $\boldsymbol{\delta^S}$ and the training set of the target specie $\boldsymbol{\delta^T}$
Output: A set of classifiers $\boldsymbol{C}$ for the target specie
Randomly dividing source set into $\boldsymbol{L}$ subsets
$\boldsymbol{\delta^S = \delta_1^S, \delta_2^S, \dots, \delta_L^S}$
Learning ensemble classifiers
for $\boldsymbol{l}$ = 1 to $\boldsymbol{L}$
Form a training data set by $\boldsymbol{\delta_l \leftarrow \delta_l^S \cup \delta^T}$
Learn a classifier on $\boldsymbol{c_l}$ on $\boldsymbol{\delta_l}$
Update the classifier set $\boldsymbol{C \leftarrow C \cup \{c_l\}}$
end for
return $\boldsymbol{C}$

Table 3-17 pseudocode of learning ensemble classifiers.

To address this problem and achieve an appropriate adaptation, inspired by [44], we propose an approach of using ensemble classifiers to perform cross-species fight detection, where we adapt animal fight data to build a human fight detection model. Figure 3-10 depicted the architecture of our approach. When source data dominates the training set, the resulting hyperplane tends to discriminate the difference in a specific domain, e.g. between animal fight and animal non-fight. However, when we regroup the subset of source data and the target data in a reasonable manner, there is a higher chance for each ensemble classifier to be able to identify the difference between fight and non-fight across

domains. However, some of the animal fight data may help in detecting human fight but some may not. In order to reduce the effect of bad animal fight data and maintain balanced training sets, we randomly divided the source data into L subsets, and then we trained L classifiers on each subset combined with partial human data. Table 3-17 presents the pseudocode of learning our ensemble classifiers. Finally, we bag these L classifiers and apply a voting strategy to derive the final predictions.

Figure 3-11 illustrates an example of our cross-species learning approach. The dashed lines denote the ideal hyperplanes for the target (blue) and source (green) domain. They may not align well due to differences in species. As in practice there is only a small amount of available target samples, the resulting target hyperplane (blue solid line) tends to deviate much from the ideal one. However, training with different source subsets enhances the chance to produce more proper hyperplanes (pink and purple solid lines) for the target domain, thus increasing the likelihood of learning the correct hyperplane.

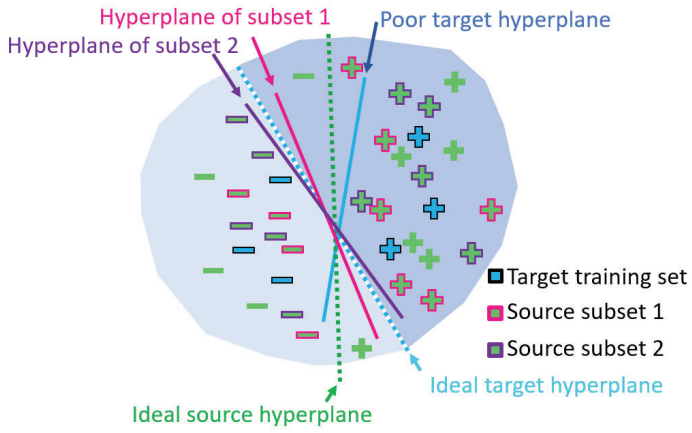


Figure 3-11 Example to illustrate the cross-species learning.

3.5.3 Evaluating Cross-species Learning

Evaluating on Various Datasets

It is encouraging to see that our approaches with *MSF* and *LMF* yield promising results for human fight recognition tasks. In order to compare with the state-of-the-art methods in fight or violence detection, we then evaluated our approach with various datasets containing fights from different scenarios and species. Table 3-18 summarizes the evaluation performance of our approaches.

Approach	Human Fights	Animal Fights	Hockey Fights	Action Movie
MoSIFT [81]	74.4%	75.2%	**90.9%**	89.5%
ViF [37]	80.1%	80.1%	81.6%	93.0%
OViF [31]	81.6%	79.6%	84.2%	89.0%
Motion Signal Features	82.7%	79.6%	84.5%	98.5%
Local Motion Features	**87.6%**	**84.0%**	87.5%	**99.0%**

Table 3-18 Evaluations on various fight datasets.

According to the table, we can observe that for real human fight, animal fight and action movie dataset, our approach with *LMF* outperforms the state-of-the-art methods using MoSIFT [81], ViF [37], OVif [31]. *LMF* consistently outperforms other features. As for hockey fights, the accuracy of *LMF* is slightly lower than that of MoSIFT features, but it still outperforms those of other features. This indicates that *LMF* is a good representation of intrinsic actions inherent in fight events. More importantly, it can be well generalized across different scenarios and species.

Cross-species Fight Detection

We are also interested in cross-species fight detection. In this situation, the training data set is from one source (e.g. animal fight) and the test data set is from

another (e.g. human fight). We therefore evaluate our method on this task. In order to make a comparison, we also evaluate our method on adapting from other sources including hockey fights and action movies, to real human fight detection.

Besides, the effect on the choice of features in cross-species learning is another interesting research question to answer. In our experiments, we also investigated the effect of different features on cross-species learning in human fight detection. We are interested in whether our proposed features are really suitable for cross-species learning.

For studying cross-species fight detection, we first investigated the adaptation through ensemble learning using our proposed approaches. Using ensemble classifiers to do the cross-species fight detection is a solution to reduce the effect of some useless fight data from the source dataset. In practice, the amount of target data is oftentimes limited, especially in the case of fight detection. To evaluate the learning curve with incremental target data, we present the performances of human fight recognition while training on an incremental amount of target data. Following the previous practice [16], we evaluate on the training when 10% to 90% samples (increments by 10%) are available. For each experiment, we first selected 10% samples from target data in random as the test samples. Then from the remaining 90% target samples, we incrementally added 10% to the training set to perform the adaptation learning with source data. In order to reduce the effect of randomization, we repeated the procedure 20 times for all the adaptation learning experiments to even out performance fluctuation.

The cross-species adaptation by ensemble learning yields promising results. For the comparison purpose, in our experiments, we evaluated the ensemble learning models with 3 different choices on the number of ensemble classifiers, L: 5, 10 and 15 for all the sources. Figure 3-12 depicts the performance of the

experiments. The x-axis shows the number of samples (10% to 90%) from the human fight dataset. Different curves present the results of 1) the single classifier trained on only human data as well as 2) the L-ensemble classifiers trained on human data and different subsets of source data.

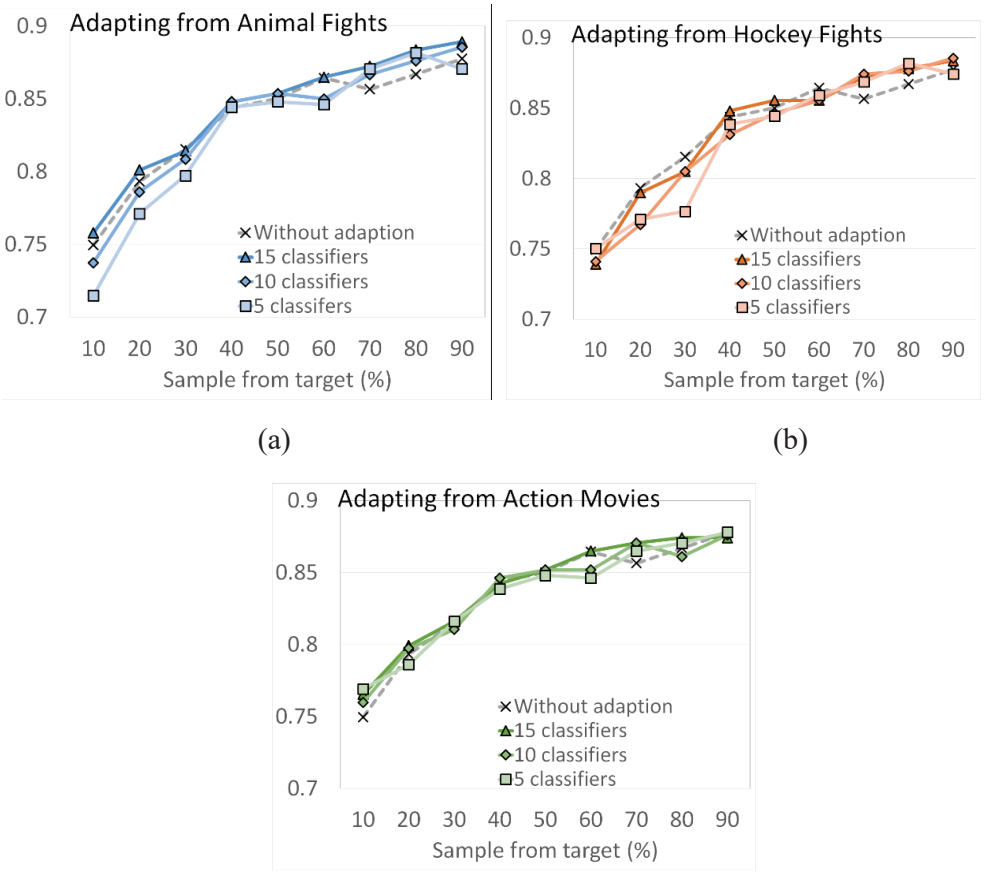


Figure 3-12 Performance of adaptation by learning from ensemble classifiers. (a) adaptation from animal fights, (b) adaptation from hockey fights, (c) adaptation from action movies.

For adapting from animal fights (Figure 3-12 (a)), in general, given a sufficient amount of human fight data ($> 60\%$ samples), adaptation by ensemble can consistently outperform that of without adaptation. Even without sufficient target data, the ensemble model can also slightly outperform the model without adaptation. Additionally, as expected, the performance of adaptation improves as

the amount of human data increases as well as the number of classifiers L for the ensemble goes up. In particular, training with 90% samples of real human fight with 15 classifiers achieves an accuracy of almost 89%.

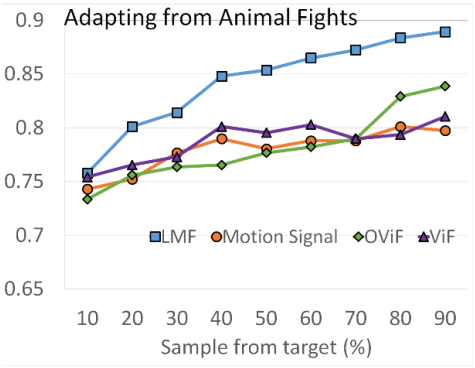
In addition to animal fights, we also evaluated our method on adapting from hockey fights and action movies to real human fight detection. Figure 3-12 (b) and (c) depict the results. For adapting from hockey fights and action movies, we can also find that the performance of adaptation improves as the amount of human data increases as well as the number of classifiers L for the ensemble goes up. However, the performances are not as good as adapting from animal fight data. This further suggests the usefulness of cross-species learning, since both animals and human appear to fight in a similar natural manner. Fighting in hockey games is limited by the venue, their wearing and the device (hockey stick), and fighting in action movies is acted upon with possibly fast-motion for playback.

Effect of Features on Ensemble Learning

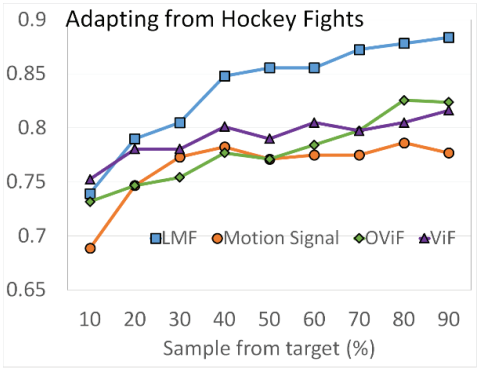
Since an appropriate feature representation that captures the intrinsic characteristic of fight motion is essential to facilitate cross-species learning, we start with experiments to answer the effectiveness of the proposed feature representation. To this end, we compare with the commonly used representations for human fight or action recognition, including motion signal, OViF [31], and ViF [37]. Since when compared with other approaches, the MoSIFT [81] approach does not perform very well in human real fight and animal fight detection according to Table 3-18, we thus did not adopt the MoSIFT approach in this experiment. As the very first step to investigate the practicability of cross-species learning in fight detection, we only focused on detecting fight events from hand-crafted features. Once we understand more about it, cross-species learning in fight detection can be conducted in a better way by applying well-performing data-

driven approaches like deep learning.

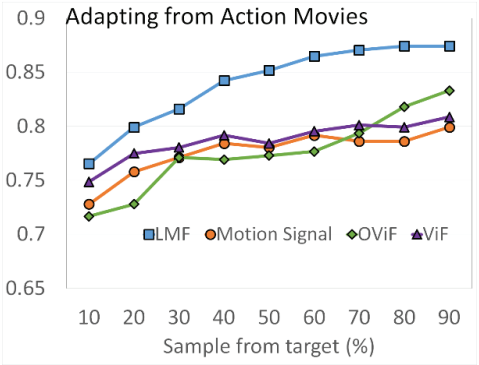
According to the experiments above, we know that an ensemble with 15 classifiers can contribute to a better performance than an ensemble with 5 or 10 classifiers. Therefore, we adopted $L = 15$ as the number of ensemble classifiers consistently in this experiment for all the feature sets. Figure 3-13 shows the performance of real human fight detection with adaptation from animal fight dataset, hockey fight dataset, and action movie dataset. Different curves denote the performance of using different feature representations. Most encouragingly, we see that *LMF* (blue squares) can considerably outperform its counterparts. This is consistent across adaptations from either other human motion datasets or animal fight dataset. It thus indicates that the proposed *LMF* feature set is most suitable for cross-species learning.



(a)



(b)



(c)

Figure 3-13 The effect of different feature sets on ensemble learning in human fight detection when adapting from (a) animal fights, (b) hockey fights, and (c) action movies. The proposed *LMF* features generally outperform the state-of-the-art motion features across adaptation from different datasets.

Another interesting observation is that adaptation from animal fights can outperform those from hockey fights and action movies with the *LMF* and motion signals features. In contrast, this is not consistently true with the two ViF based features. This may be because *LMF* and *MSF* features encode the motion information based on the value of motion amplitude or acceleration, while the ViF based features are more focused on counting the changes of significant motions. A potential interpretation is that the animal fights and human fights share some similarities in motion amplitude and acceleration. This also implies that among different motion features, the proposed motion analysis-based features can identify the intrinsic motion representation in the most appropriate manner, and thus facilitates the cross-species learning task in fight detection.

3.6 Summary

This chapter presents the efficient approaches of using motion analysis-based features to detect human real fights, without relying on recognizing complex behavior, gesture or actions, etc.

To verify the performance on spontaneous fights in real scenarios, we proceed to collect real surveillance videos from YouTube, which contain real fight events. Then we annotated them to form a new dataset. The experiments demonstrate that the performances induced by the use of our features are quite encouraging. The use of *LMF* features can outperform the other state-of-the-art features in real fight

detection.

However, the performances of our proposed features are not good when we evaluate them on simulated fights. We suspect that this is due to the fact that real and simulated fights are different in motion behaviors, though they might look similar in gestures. We then conduct experiments to investigate the difference between real and simulated fights. The insights we learn from the experiments include: (1) human can easily tell real fights from simulated fight events; (2) human tend to regard simulated fight events as non-fight; (3) our algorithm can differentiate between real and simulated fights; (4) our algorithm seems to be capable of doing a better job than human. We believe that our approaches benefit from the motion analysis nature, being able to distinguish between real fights and simulated fights, which exhibit quite different motion behaviors. The study can help us understand more about the fundamental difference between real and simulated fights, as well as the real fight intention.

The chapter also presents the study of cross-species learning to address human fight detection, where real spontaneous data is rare and insufficient for data-demanding learning algorithm. Our approach adopts ensemble learning to adapt useful knowledge from similar subsets of source data and achieves adaptation with our proposed *LMF* and *MSF* features. For evaluation purpose, we prepared the first animal fight dataset. In our experiments, the proposed *LMF* features are demonstrated to be generalizable across scenarios as well as species. Our results indicate that learning with animal data can improve the performance of human fight detection. In addition, adapting from animal fights to human fights produces comparable and even favorable results with those of adaptation from other human fight from other scenarios such as sports and action movies. This seems to suggest that human perhaps fight like animals in some way. Further, the proposed *LMF*

feature representation is also empirically shown to be effective in cross-species learning in fight detection. We believe that our studies would shed lights to the studies of some other human and social interactions.

Chapter 4 Exploring Multi-modalities User Intention Prediction

Selected notations and abbreviations used in this chapter

a_i	an activity, belonging to one of the five search activities introduced in Section 4.1
c_i	correct classified rate of modality i
D_k	duration sequence containing the durations of k most recent activities
$\widehat{D}_k$	duration level sequence, obtained by quantizing the continuous values in D_k
d_i	actual value of δ_i, a continuous value
fv	feature vectors extracted for user intention prediction
I	the set of modalities
k	number of most recent activities involved in our models
l_i	actual value of $\hat{\delta}_i$, belonging to one of the discrete duration levels
$MI(\vec{v})$	movement magnitude of movement vector $\vec{v}$
n_s	number of subjects
p_t	pre-selected threshold for user slips detection
S_i	a search activity in T
s_i	activity type value of S_i, belonging to one of the five activities
T	activity sequence of a web search task
T_k	sequence that containing k most recent activities before the current moment
$\vec{v}$	movement vector of interaction and body movements
v_x	horizontal component of $\vec{v}$
v_y	vertical component of $\vec{v}$
W	size of the time window for extracting interaction and body signals
w_i	weight of modality i for fusing multi-modalities in decision level

X	time (seconds) preceding the next activity
$\alpha(\vec{v})$	orientation of movement vector $\vec{v}$
δ_i	time duration between activities S_i and S_{i+1}
$\hat{\delta}_i$	discrete level of the duration between activities S_i and S_{i+1}
HF	abbreviation of histogram-based feature
SF	abbreviation of statistics-based feature

Chapter 3 presented the body motion analysis-based approaches to build a well-performing model for fight detection and demonstrated that the proposed approaches can be applied to reveal the underlying real fight intention behind human action. In this chapter, we investigate user intention in a different direction, that is, predicting user interaction intention in daily computer interaction tasks. User intention may differ a lot in different tasks. We are more interested in multi-step interaction tasks. We thus focus on natural web search task in our study presented in this chapter.

In the fight detection task, we only utilize the body motion signals captured by camera. However, in this study, we aim to build a user intention prediction model by using multiple modalities, including mouse, gaze, head movements and body motions, as well as the historical activity sequence. To study and evaluate user intention prediction models, we collect a user interaction intention dataset by using non-intrusive and low cost devices. This chapter contains the description on the details of the dataset as well as the captured interaction and body signals.

The major challenge of this study is to find the appropriate features and multi-modalities fusion approaches to build effective user intention prediction models. To understand the performance of each individual modality and find the appropriate feature representation to model the interaction and body signals, this chapter investigates the user intention prediction models with different modalities and with different feature representations. Moreover, to bridge the gap between the limited approach of modeling individual modality and modeling multiple interaction modalities, this chapter explores the performance of the prediction models through different ways of fusing multiple modalities. Besides, a pilot study of user slips detection is also presented in this chapter.

The rest of this chapter is organized as follows. Section 4.1 introduces the

user intention task that we adopt in our study, namely, the web search task. Section 4.2 presents our prediction models and the extraction of the underlying multiple modalities. Section 4.3 describes the nature of our collected dataset. Section 4.4 presents the evaluation experiments for user intention prediction. Section 4.5 describes the pilot study of applying the proposed user intention prediction model to user slips detection. Finally, this chapter is concluded in Section 4.6.

4.1 User Intention Task

User behavior varies from task to task. We choose our target experimental task to be one which involves complex and multi-step interactions, in which a user needs to perform a series of actions to trigger a series of intended activities. Web search is a very common computer interaction task in daily life. Compared with the prescribed task and the task with fixed interface often adopted in other contemporary research works, the web search task is by far more open-ended and thus the interactions are more complicated. Therefore, we choose web search task to study user intention prediction in this thesis for better generality.

In our web search experiment, we asked subjects to play a game similar to the search game “A Google a Day” [34]. In the experiment, users are asked to find the answer to a given question by searching on the web. The question is formulated in such a way that the answer cannot be found via a simple web search. We restricted the browser and search engine to Chrome and Google respectively.

One of the questions associated with user intention prediction is how to categorize a user's intention. User's interaction intention may vary across different tasks. In our web search task, we define the user's intention based on the web search activities, such as reviewing search results. Since the web search may return any page, the interface of web search task is not fixed and we cannot classify a

user's activity based on the interface. We define an activity as a change in the currently-viewed page in the browser and do a rough classification of the content on the page, which serves as an indication of the users' intent. For instance, a “New Tab” probably means the user wants to start a new search, and users visiting a page linked to from the search results are probably in the state of parsing information. We thus classify five types of activities: 1) forming a searching goal, 2) starting a new query, 3) reviewing result page, 4) parsing information, and 5) submitting answer. Figure 4-1 demonstrates an example of our web search task and the five types of activities.

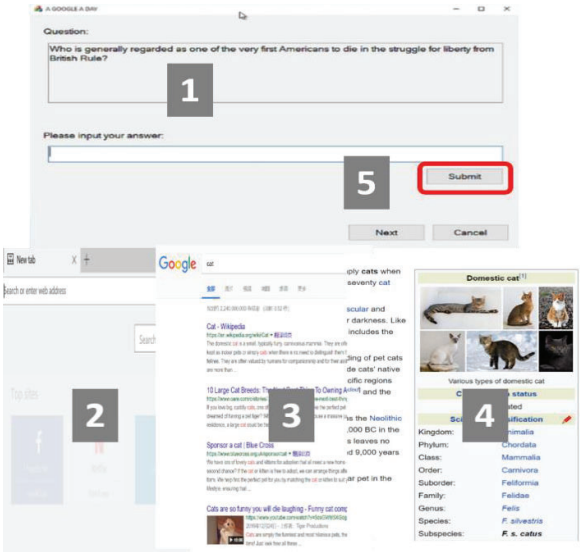


Figure 4-1 Web search task and the five types of activities

We formally describe the users' activity sequence through the visited page sequence $T =< S_1, S_2, \ldots, S_n >$, where T is an instance of the search task and S_i is a search activity. For instance, if a user extracts search keywords before clicking “New Tab” to start a new query, and finally obtains the answer from search result page as well as submits it, then the activity sequence is $T =< 1, 2, 3, 5 >$. Predicting a user's search intention involves predicting the value of S_{n+1}. Recall the two different scenarios that we would like to study in user intention prediction,

i.e. predicting the next activity when it is about to occur, and predicting it when it would soon occur. In the first scenario, in which it is known that an activity is about to occur, we detect the value of S_{n+1} when S_{n+1} is about to occur. While in the second scenario, we predict the value of S_{n+1} a few seconds ahead before it occurs.

Once we can efficiently predict user intention, we could be able to perform slips detection based on the intention prediction model. Sometimes users may trigger some wrong actions even after they formulated the right intention. These wrong actions are referred to as slips [109]. A user slips detection model can automatically detect whether a user is selecting the wrong target or triggering the wrong activity and can help to save the time from recovering from slips. Therefore, in this study, we are more interested in whether a triggered activity meets user's intended activity. To achieve this, instead of detecting the value of S_{n+1}, we try to detect whether the given value of S_{n+1} is correct or not, when it occurs.

4.2 Extracting Features from Multi-modalities

We are interested in whether we can predict the next activity, i.e. the type of next mouse click event, given some interaction signals and history activity. Prior related studies suggest that historical records contain information about user activities and are highly useful. We also utilize historical information in our model. At the same time, our approach involves multiple interaction modalities, including mouse interaction, eye gaze, head movements and body motions, etc. We believe that these interaction and body signals could also convey useful clues to indicate user's intention. This chapter presents the details of our proposed features.

4.2.1 Features from Historical Activities

In a multi-step computer interaction task, user's activities are sequential and

each activity is probably dependent on past ones. Prior studies have investigated in modeling user's intention from historical records. And these studies demonstrate that hidden information about users' intentions can be deduced from the activity history. In this study, we also make use of the historical activity sequence as one of our feature sets. Specifically, we consider historical activity sequence in two ways: probability model and classification model.

Probability Model

We first try to consider adopting a statistical approach to utilize historical activity sequence. We refer to this as the "probability model". This approach relies on a classical *n-gram* model on the historical activity sequence. In this approach, we describe the activity sequence as a tuple: $T =< S_1, S_2, \dots, S_n >$. The value of a random variable S_i indicates one of the defined activity types introduced in Section 4.1. For practical purpose, we consider the most recent k activities. We then represent those k activities as: $T_k =< S_{n-k+1} = s_{n-k+1}, S_{n-k+2} = s_{n-k+2}, \dots, S_n = s_n >$, where S_n is user's current activity, s_n is the value of the activity type. We refer to this as the k^{th} order probability model. Predicting the next activity S_{n+1} is a matter of computing the value of s_{n+1} that can maximize the conditional probability given the previous k activities by:

$$\mathrm{P}(s_{n+1} \mid T_k) = \Pr(S_{n+1} = s_{n+1} \mid T_k) \qquad 4.1$$

In addition to historical activities, we consider the time spent on each activity, which we refer to as the duration of the activity. By using the duration information, we then build a second prediction model. Similarly, we represent the time spent on the activities by a duration sequence: $D_k =< \delta_{n-k+1} = d_{n-k+1}, \delta_{n-k+2} = d_{n-k+2}, \dots, \delta_n = d_n >$, where δ_i is a random variable that models the time duration between activities S_i and S_{i+1}, d_i is the actual value of that duration,

which is a continuous value. In order to apply it to the probability model, we quantize the continuous value into 3 discrete levels: {long, medium, short}. That makes it possible to generate a duration level sequence $\widehat{D}_k =< \hat{\delta}_{n-k+1} = l_{n-k+1}, \dots, \hat{\delta}_n = l_n >$, where l_i belongs to one of the levels described above. This time, we consider the conditional probability, given the most recent k activities and the corresponding time durations. Predicting the next activity is to compute the value of S_{n+1} that can maximize the conditional probability of:

$$\mathrm{P}(s_{n+1} \mid T_k\widehat{D}_k) = \Pr(S_{n+1} = s_{n+1} \mid T_k\widehat{D}_k) \qquad 4.2$$

Classification Model

Probability model can be one solution to model historical information. However, it is difficult to extend the probability model by adding additional features to it. With the expansion of the number of conditions, the number of potential states grows exponentially, leading to a sparse data problem and large error with the probability distributions. If we want to make use of more features for prediction, a better choice is to apply machine learning algorithms. We therefore consider historical activity information in a second way, that is, to extract features from it and build classifiers to detect user intention. We refer to this approach as the "classification model", which can be further augmented to a multimodal approach by fusing features from other interaction modalities together.

In the classification model, our historical activity features include the k most recent activities and their durations. In total, the history sequence information provides us $2 \times k$ features. Table 4-1 lists the features extracted from historical activity sequence.

Attributes	Features
k most recent activities	$s_n, s_{n-1}, \dots, s_{n-k+1}$
k most recent activity durations	$l_n, l_{n-1}, \dots, l_{n-k+1}$

Table 4-1 Features from historical activities.

4.2.2 Features from Interaction and Body Signals

We hypothesize that users' intention may be concealed behind interaction and body signals including mouse, gaze, head, and body movement, especially in complex tasks. In the complex computer interaction task, the mouse serves as a key input method. The mouse movements may vary depending on whether the user is reading text, watching videos, clicking links or making decisions. Gaze behavior may also indicate user's interaction activity and attention saliency which may indicate some intended targets. In addition, the head movement and body motion may also reflect human intention as we discovered in our fight detection study. Moreover, these interaction and body signals can be captured by computer log, standard camera, and other non-intrusive devices. We therefore focus on modeling these commonly available signals in our study. For these interaction and body signals, we consider two feature representations: statistics-based features (*SF*) and histogram-based features (*HF*). In our study, we extract the interaction features by using both the feature representations, and conduct experiments to investigate which one is more appropriate for user intention prediction.

Capturing interaction and body signals

Before extracting features from the interaction and body signals, we need to be able to capture them. In daily computer interaction tasks, the mouse serves as a key input device. Therefore, in this study, we extract features from mouse movement (sequence of the x and y coordinates). The mouse movement

information is logged by the operating system. It thus can be easily extracted from the system log.

Eye gaze and head movement are also important interaction and body signals that contain the information about user intention. Fortunately, with the help of the computer vision techniques, both gaze and head movement can be captured from a standard webcam directly. By using computer vision techniques, human face can be detected from the webcam, and then the head movement can be directly linked to the movement of the detected human face. Meanwhile, the gaze direction and gaze points can be estimated by appearance-based estimation method from the detected face [120]. In this study, we apply OpenFace toolkit [7] to detect human face and estimate user's gaze direction. Figure 4-2 demonstrates the example of capturing head and gaze movement from a webcam.

In addition to capturing from webcam, there are other possible methods to extract user's gaze interaction information. One common way is to rely on additional eye tracking device such as Tobii to estimate a relatively precise eye gaze location. However, these eye tracking devices are expensive. They are not affordable and accessible for common users. It is interesting to understand the performance gap between the features collected from these two methods, and explore how good we can achieve if we do not rely on an eye tracking device. Therefore, in our study, we collected and extracted user's gaze information by both methods for comparison.

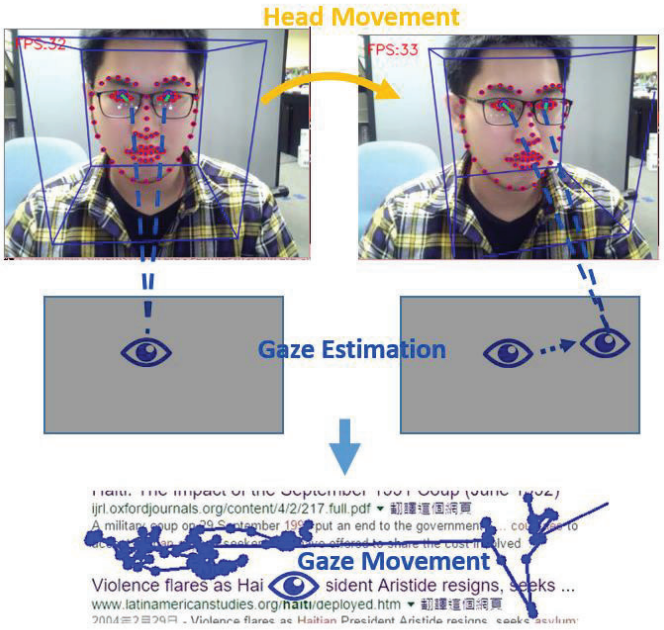


Figure 4-2 Example of capturing gaze and head movement from a webcam with the help of OpenFace toolkit [7].

By using the approaches described above we can capture the mouse, gaze and head movement. However, detecting and tracking hand or other body part is not as easy as detecting human face, as the hand moves much faster and may manifest with various gestures. Tracking the movement of the other body parts may not be stable and may introduce additional noise. On the other hand, our fight detection study shows that it is possible to detect human intention based on body motion analysis. We therefore propose to extract the body motion signals from the other body parts as the additional information for user intention detection by using optical flow. Inspired by the local motion signals we applied in our fight detection task, we propose to extract local motion signals to account for spatial variations of the body motions in this task. Similar to the local motion features introduced in our fight detection study, given a frame in a video, we first segment the video frame into several regions. However, this time, we segment the frame based on the heuristic human body parts instead. As demonstrated in Figure 4-3, we segment

the frame into 7 regions. Body motion occupied in different regions may be related to the movement of different body parts, and it may indicate different underlying intention. For instance, a user holds his face in his left hand while he is thinking, then the left body region and mouth region may capture the corresponding hand motions. We then compute the optical flow [28] and measure the amount of motion magnitude of each motion region by Equation 3.6. We then obtain 7 sequences of the motion magnitude from the video.

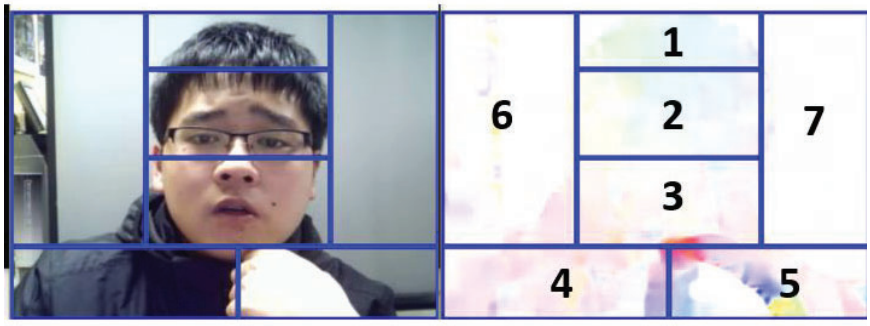


Figure 4-3 Example of the body based local motion regions. 1 head region, 2 eye region, 3 mouth region, 4 right body region, 5 left body region, 6 right shoulder region, and 7 left shoulder region

Statistical Features

After obtaining the interaction and body signals, we need to extract features from them and apply those features to build user intention prediction models. We first introduce our statistics-based features in this section.

For extracting features from the interaction and body signals, we first select a time window W. Only the interactions that occur within W will be considered in our model. These interaction signals contain rich information and occur in various forms. We focus on extracting statistics-based interaction features from the movement attributes for all the interaction modalities. Taking mouse interaction as an example, a user moves the mouse cursor for various reasons, e.g.

to aid reading, select text, click buttons, etc. Thus different mouse movements may reflect different intentions of the user. We therefore extract features from the movement attributes to model user intention. There are many ways to define an interaction movement period and describe the information contained within. In our study, we segment a movement instances according to the moments when the movement speed effectively drops to zero. The trajectory traced out by movements with non-zero speed is then processed to generate movement attributes. In our work, we consider the following movement attributes for mouse, gaze and head movement (please refer to the mouse movement example in Figure 4-4):

- Travel distance: total traversed distance of a movement, e.g. the black curve.
- Shortest-path distance: length of the straight line connecting the start and end points of a movement, e.g. the red line.
- Movement angle: angle between the shortest path defined above and the horizontal x-axis, e.g. θ.
- Angle of curvature: for consecutive recorded points A, B, C, the angle between line AB and line BC, e.g. ϕ.

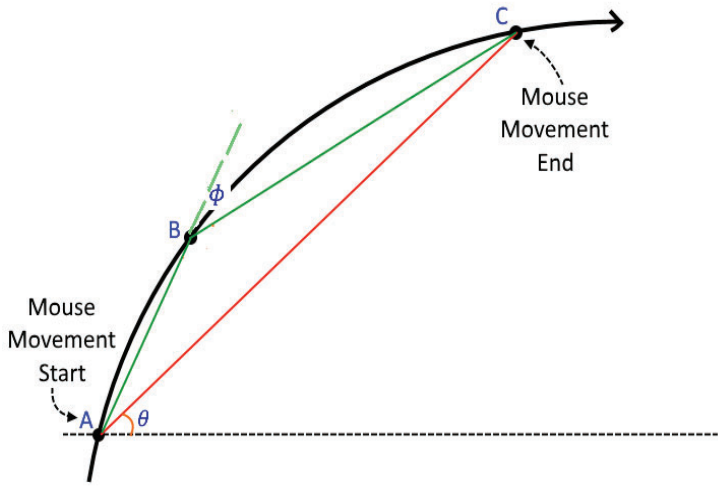


Figure 4-4 Example of an interaction movement and the attributes

Given a temporal window W, we might expect to see multiple movements. We thus would obtain a set of values for each of the above attributes. For each set of attribute values, we calculate the descriptive statistics, i.e. the mean, maximum, minimum, median and standard deviation, for these attribute values, as shown in Table 4-2.

Attributes	**Statistics**
Travel distance	Mean, maximum, minimum, median, standard deviation
Shortest-path distance	
Movement speed	
Movement acceleration	
Movement angle	
Sum of angle of curvature	

Table 4-2 Movement attributes and statistical features.

For mouse, gaze and head movement, we extract statistical features based on the movement attributes. For the body motion signals, we extract the local motion features that are applied in our fight detection task. After obtaining the local motion sequences based on the different body regions, we then measure the motion dynamics using the descriptive statistical features throughout the time window W. We then obtain the statistics-based features (SF) from all the interaction and body signals, which are summarized in Table 4-3.

Attribute	Features
Mouse interaction	Statistics of the movement attributes
Gaze interaction	
Head movement	
Body motion	Statistics of the local motions

Table 4-3 Statistical features from multi-modalities.

Histogram Features

Statistical features can summarize some information of the interaction and body signals. However, the drawback of the proposed statistical features is that they consider the movement attributes separately. Interaction signals may contain rich information, and can be summarized in different ways. Considering different movement attributes at the same time may reveal additional information about user's intention. Therefore, in our second approach, we aim to encode different movement attributes at the same time, specifically, we focus on extracting the mouse movement magnitude and orientation, which are the most import attributes of a movement.

We propose to use histogram as our second feature representation method which can encode the mouse movement magnitude and orientation information at the same time. Similar to extracting statistical features, we first select a time window W, and then only consider the interactions occur within that window. For a movement vector $\vec{v} = (v_x, v_y)$ of two consecutive recorded points, occurring in the defined time window (e.g. $\overrightarrow{AB}$ in Figure 4-4), we calculated the movement magnitude $MI(\vec{v})$ and orientation $\alpha(\vec{v})$:

$$MI(\vec{v}) = \sqrt{{v_x}^2 + {v_y}^2} \quad 4.3$$

$$\alpha(\vec{v}) = tan^{-1}(\frac{v_y}{v_x}) \quad 4.4$$

Where v_x and v_y are the movement speeds in horizontal and vertical direction respectively.

Our movement magnitude histogram features use both the magnitude and orientation information of one particular movement vector. We equally divided the orientation between 0 to 2π into 16 orientation bins. We then mapped each movement vector to the corresponding orientation bin by accumulating the movement magnitude to that bin. For instance, if the orientation of a mouse movement vector belongs to bin 8 and the magnitude is 2.0, then 2.0 will be accumulated to the bin 8 of the histogram. Figure 4-5 illustrates an example of the mapping procedure. After mapping all the mouse movement vectors occurring in the defined time window, we then finally obtained the movement magnitude histogram features with 16 dimensions.

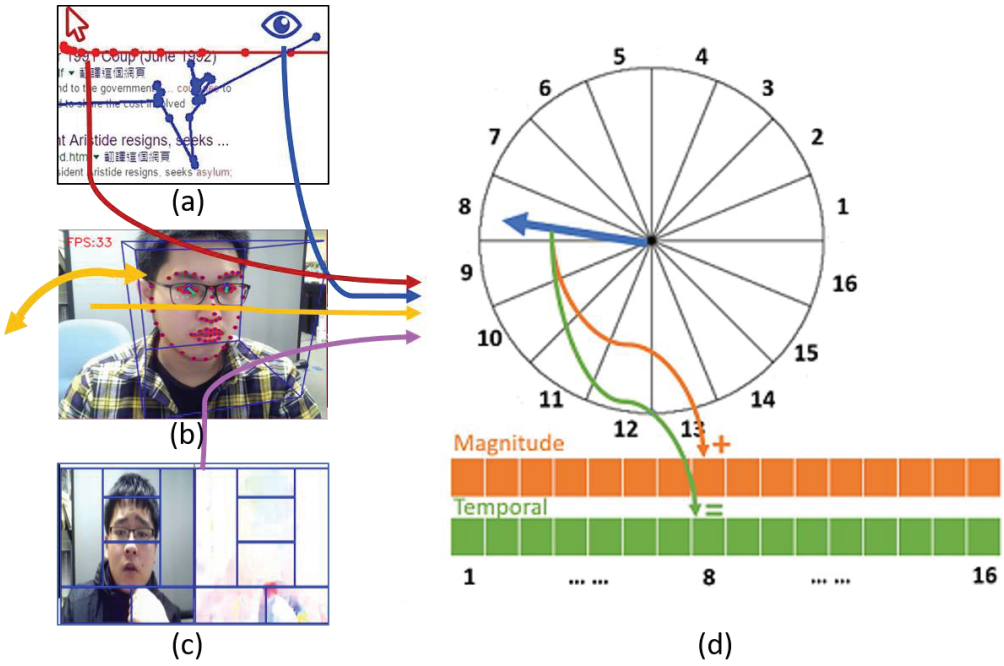


Figure 4-5 Mapping interaction movements of (a) mouse, gaze movements, (b) head movements, and (c) body motions, to corresponding histogram bin to generate histogram-based features (d).

However, the current features cannot model the temporal information of the

mouse movements. In order to utilize the temporal information, we induced an additional histogram. Same as the movement magnitude histogram, the temporal histogram is formed by dividing the full orientation 0 to 2π into 16 orientation bins. Similarly, we mapped each movement vector to the corresponding orientation bin. But this time, we assigned the relative time stamp of the mouse movement to that corresponding histogram bin. For instance, if a mouse movement vector occurring at the relative 0.9 seconds of the time window belongs to bin 8, then 0.9 will be assigned to bin 8 of the temporal histogram. Figure 4-5 illustrates the example of mapping temporal information of a movement. After mapping all the mouse movement vectors, the final temporal histogram recorded the latest updated time for each orientation bin. Combining the two histograms together brings us the final set of histogram features with 32 dimensions.

For mouse, gaze, and head movement, we can all extract the histogram features by the same approach described above. However, we cannot apply this approach to extract histogram features of body motion signals, since we cannot capture the movement vectors from body motions.

To capture the body motion signals, we compute optical flow and measure the amount of motion magnitude of each body motion region. As introduced in Section 3.2, each pixel in optical flow images actually represents an optical flow vector, and the magnitude implied of each pixel can be computed as the length of this vector. We can also get the orientation of the optical flow vector. Therefore, for extracting histogram features from body motion signals, we consider the magnitude and orientation of the optical flow vectors occur in each of the body motion regions within W. We then also obtain the magnitude histogram and temporal histogram for each body motion region.

We finally obtain the histogram-based features (HF) for the interaction and

body signals. Table 4-4 shows the histogram features for all the interaction and body signals.

Attributes	Features
Mouse interaction	Magnitude, temporal histograms of the movements
Gaze interaction	
Head movement	
Body motion	Magnitude and temporal histograms of the local motions

Table 4-4 Histogram features from multi-modalities.

4.3 Constructing User Intention Dataset

To study on the user interaction intention prediction and user slips detection, we need ground truth and interaction data. We designed and conducted experiments to collect user intention dataset in the web search task. The experiment involved 19 subjects (9 females), all comfortable with computer usage. A user interface, only used for displaying questions and submitting answers, was designed for the purposes of repeatability and controllability. In our experiment, each subject was asked to answer 6 questions. In total, 6598 mouse click events were collected. All the experiments were run on a standard desktop computer running Microsoft Windows.

Since we are interested in predicting user intention with multiple interaction and body signals, in addition to recording historical information and mouse interaction by system logs, we also recorded the videos that captured subjects' body movements by a standard webcam as well as eye gaze interaction by a Tobii EyeX Eye tracker.

4.4 Evaluating User Intention Prediction

An effective user intention prediction model should be able to automatically predict a user's next potential activity. As a user stays longer on one page or one search activity, the user intention and the intended activity could be different. In fact, the closer to the next click moment, the more our prediction performance matters in assisting the user. Therefore, we are more interested in the intention prediction performance when the next click is nearing (e.g. next click is coming in 1 second). Specifically, we first evaluated our model in the extreme case, that is predicting the next activity at the moment when it is about to occur. We then evaluated the model in predicting the next activity 0.5 to 2.5 seconds (in increment of 0.5 seconds) before the event occurring moment.

In our experiments, we first evaluate the model that only makes use of historical activity information. Since the historical activity information can be easily obtained from the log information and it does not involve any interaction patterns, modeling from historical activities can be regarded as one of the baseline models for user intention prediction. We then evaluate the performance of the model with each individual interaction modality (mouse, eye gaze, head movements, and body motions) as well as investigate the proper set of feature representation for user intention prediction. This is followed by our study on multimodal user intention prediction models using different ways of fusion.

In real application contexts, a model may well be used to predict the intention of a new user, never seen before. Therefore, we need to investigate the performance of our model on an unseen user. Hence, we adopt the *leave-one-subject-out* approach for evaluation. We train our model with data from $n_s - 1$ subjects (training set), and evaluate on the data from the left-out subject (test set),

and repeat for each subject. The overall accuracy is the performance averaged over all iterations. Since we have $n_s = 19$ experimental subjects, we iterate 19 times and report the average correct classification rate (CCR) for both models. For all the experiment presented in this study, we adopt the same evaluation approach. The experiment results of this study are summarized in the following.

4.4.1 Modeling Historical Activity Sequence

We first evaluate the models that only consider historical activity sequence. As mentioned in Section 4.2.1, in our approaches, we perform our study based on the conditional probability model and classification model. We start with evaluating our probability model. We experiment with two conditional probability models. One only considers the sequence of user activities, while the second considers both the activities and their durations. In this experiment, we focus on evaluating our model in the scenario of predicting the next activity when it is about to occur, which is the extreme case in our study. Based on that, we also investigated the impact of the value of k on the performance of conditional probability models in Equation 4.1 and 4.2.

Figure 4-6 presents the performance of the different models of using only historical activity sequence as a function of k. According to the figure, we can see that for our experiment, the baseline is around 31%, by taking the size of the majority class. While the best performance of our probability model is around 57%, achieved by $k = 3$, when the model involves both activity and duration sequences. That suggests that the past three activities are helpful for predicting the next activity. However, this is already the peak performance, which drops rapidly when we consider more past activities.

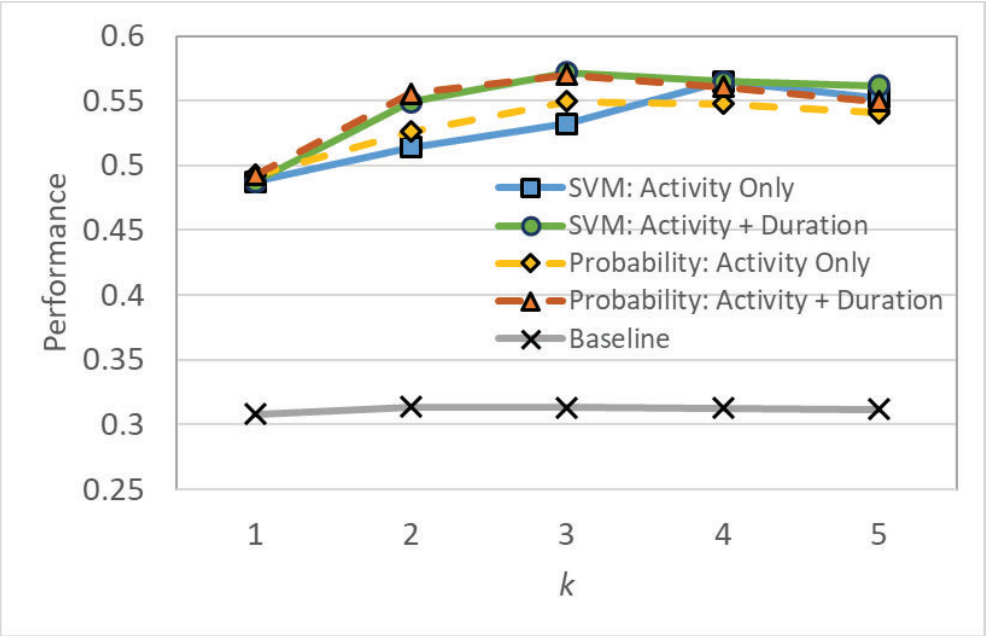


Figure 4-6 Performance of intention prediction by modeling historical activity sequence only.

From the results, we observe the effect when more information becomes available, comparing the curves for the model considering activity only versus considering both activity and duration. The performance of the latter is slightly better than the former. On the other hand, we also observe that when k is larger, the performance drops with additional information. This is probably due to the fact that the data becomes sparse rapidly as the length of the sequence increases, and the number of conditional states with duration is twice that without duration. It may also be due to the fact that there is no obvious lengthy interaction pattern that exhibits a high frequency. The improvement with more information is quickly offset by the increase in history length. Generally, these results show that it is possible to detect the user's next activity type from the historical activity and duration of the interaction sequence. It also suggests that it is not informative to go too far back: user activities from more than 3 events back in time are unlikely to be useful.

We next experiment with using machine learning approach to build a more sophisticated model for comparison. In all experiments in this study, we adopted support vector machine (SVM) [51] as our classification model. As with the classical probability model, the SVM is provided with k features and $2 \times k$ features depending on whether duration information is adopted, as depicted in Table 4-1. At a glance, the performance of the classification model is similar to the probability model for both tasks. The best performance is also around 57% for the model that considers both activity and duration sequences. This time, we also achieve the peak performance at $k = 3$ for the model considers both activity and duration sequences. Although for the SVM model with activity sequence only, the performance peaks at $k = 4$, we can still observe a trend of performance drop when we involve too much historical information. These results suggest again that the historical activities from too far back in time are not useful for predicting the next event.

When we look into further details with the classification model, we do observe some interesting difference. First, the performance drop as k increases is not as serious as in the probability model. This is due to the exponential increase in potential (ordered) sequences with k for the probability model, as compared with a modest increase with (unordered) feature sets for SVM. Second, when more data is available, SVM is able to produce better performance. This can be witnessed by the curves between activity only versus activity plus duration. This exhibits a different trend with the probability model. This is because compared to probability models, classification models is more powerful to ignore features that cannot contribute to classification and to alleviate the data sparseness problem due to more features.

In summary, we see that classification is more versatile to change in data

volume and number of attributes, producing more stable performance. Furthermore, there is no need to consider and record a long path of historical sequence for our prediction model since the best performance is achieved with small value of k. That provides useful insights for our further studies.

4.4.2 Modeling Individual Interaction and Body Signals

After evaluating the models of using historical activity information, we then conduct experiments to evaluate the models of using different interaction modalities. As we observe from the results above, historical information from too far back in time is not useful for predicting the next event. We wonder whether this phenomenon also exists in interaction and body signal features. Thus, we first conduct an experiment to find the appropriate window size for extracting interaction and body signal features. Moreover, we proposed two feature representations to encode the interaction and body signal features. We therefore also investigate on finding the appropriate feature representation in our experiment.

Finding the appropriate window size

We investigate the performance with the interaction and body signal features by varying the temporal window size W instead of the length of the historical sequence k. We anticipate that the interaction and body signal features will exhibit a similar phenomenon as before, i.e. that interaction data that is too old would not be helpful in prediction. We experiment with a window size W of 1 to 5 seconds, and extract features from the interaction and body signals that occur within the given time window prior to the prediction moment. Since we proposed two feature sets for the interaction and body signals, we run the experiments for the two feature sets respectively. In this experiment, we also evaluated in the extreme case: predicting the next activity at the end of the current activity. After finding out the appropriate window size, we then conduct further experiments to investigate the

performance of different feature representations.

We first conduct experiments with the SF feature. The performance is encouraging especially for mouse movements. The best performance is around 65%. In this experiment, we focus on finding the proper W for extracting features. According to the results demonstrated in Figure 4-7, the best performance for mouse and gaze interactions is achieved when $W = 1$ and it slightly but continuously decreases as W increases. While for the head movement and body motion features, there is a slightly declining trend when the window size is enlarged, and the highest performance is achieved when the window size is set to be 2 seconds. According to these results, it seems that the window size cannot be too large, as the information from interactions too far back in time generates noise that affects the prediction performance. For the SF, the optimal window is about 1 second for mouse and gaze movement and 2 seconds for head movement and body motion.

Based on the experiments above, we then extract SF for mouse and gaze movement from the most recent 1 second, and extract SF for head movement and body motion from the most recent 2 seconds in our further experiments presented in this chapter.

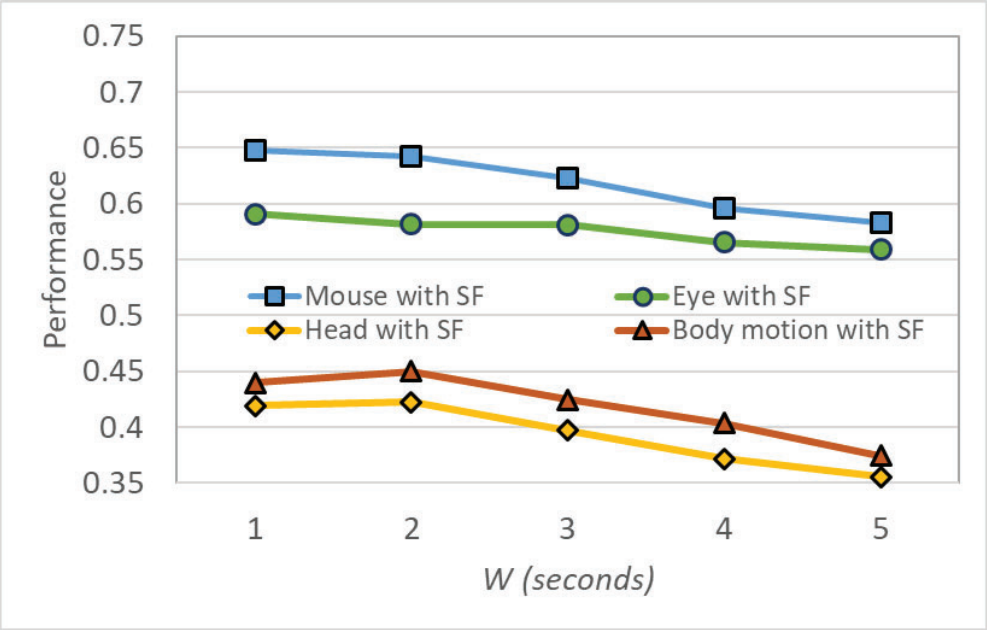


Figure 4-7 Performance of intention prediction. Features: *SF* within the W most recent seconds.

We next conduct experiments with the *HF* feature. Overall, the performance is even better than that with the *SF* features. The best performance of mouse movement can be around 70%. The experiment results are illustrated in Figure 4-8. For the *HF*, the best performances for all the interaction and body signals are achieved when $W = 2$ and there is a slightly declining trend when the window size W increases. According to these results, we again observe that the window size cannot be too large for extracting interaction and body signals, as the information from interactions too far back in time generates noise that affects the prediction performance. For our *HF*, the optimal window is about 2 seconds for all of the interaction and body signals.

Based on the experiments above, we then extract *HF* for interaction and body signals features from the most recent 2 seconds in our further experiments presented in this chapter.

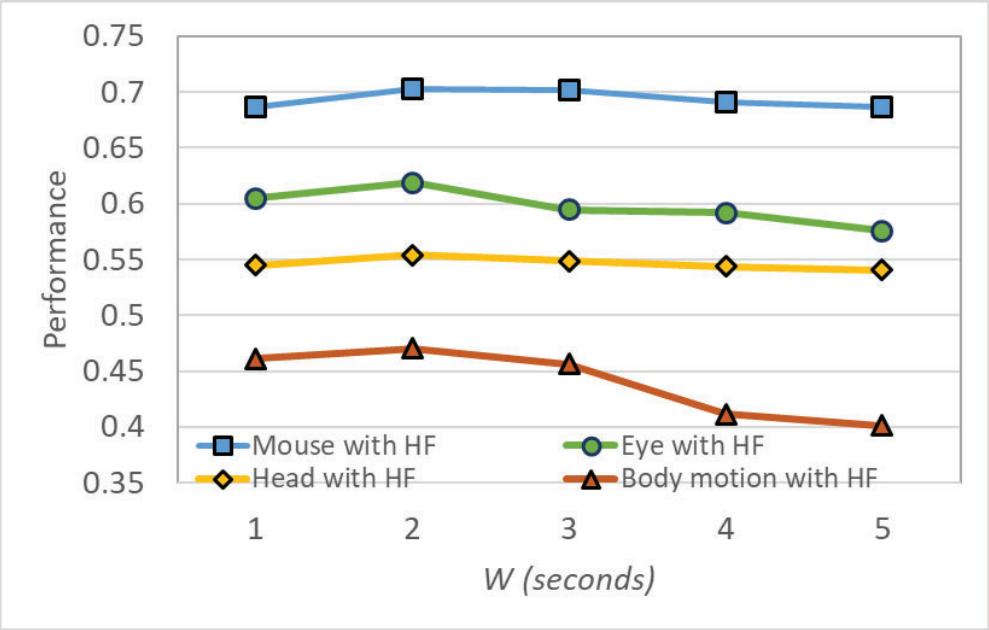


Figure 4-8 Performance of intention prediction. Features: *HF* within the *W* most recent seconds.

Finding the appropriate feature representation

According to the experiments above, we find the appropriate *W* for extracting features. We also observe that it seems that there are some performance gap between *HF* and *SF*. In order to further understand the appropriate features for user intention prediction, we then evaluate our model of using individual interaction modality with different feature representations. Particularly, we utilize the interaction and body signals including mouse interaction, eye gaze, and head movement as well as body motion in our study. For each modality, we extract the proposed *SF* and *HF* from the signals to build the user intention prediction models.

In the real application, in order to provide the chance of assisting users beforehand, a system needs to predict the next potential activity few seconds before it actually occurs. We then conduct experiments to investigate how well our prediction models could do in predicting a user's next activity *X* seconds ahead before it actually occurs. Since the closer to the moment of the next activity, the

more our prediction accuracy matters. The model does not need to be very precise when the next activity is too far away. We believe that 2.5 seconds should be adequate for such purposes. In this experiment, we evaluated our prediction model with X varying from 0 to 2.5 seconds, with the same features described above. The experiments involve the situation of $X = 0$s, which can be considered as predicting at the occurrence moment of the next activity. This case is corresponding to the scenarios, in which we need to understand user's action correctly when it is completing, such as detecting whether an action is an error.

Figure 4-9 shows the results. In this experiment, we set $k = 3$ for the historical activity sequence features, set the appropriate W for the corresponding interaction and body signal features to achieve the best performance of each individual modalities. Overall, as we move further away from the interaction event, obviously, the performance drops, as it is difficult to predict too far into the future. However, we can see that it is still possible to predict slightly ahead into the future with reasonable accuracy. Considering taking the size of the majority class as the baseline, the best performance (achieved by mouse movements with *HF*) can beat the baseline by 39% when $X = 0$ s and about 32% when $X = 0$ .5s. While considering the historical activity model as the baseline, the best performance can beat the baseline by 14% and 13% when $X = 0$s and $X = 0.5$s respectively. The results show that our models can predict user intention a few seconds ahead.

Meanwhile, among the interaction modalities, modeling with mouse interaction can achieve the best performance for both feature representations across the experiments. Specifically, the models of using mouse features can achieve 70.2% and 64.7% for our histogram-based feature and statistics-based feature respectively, when $X = 0$ s. As mouse interaction is dominating in computer interaction tasks, it should contain abundant cues to indicate user's

interaction intention. Though not as good as mouse interaction, eye gaze and head movement could also indicate user's intention to a certain degree. Features from eye gaze movements can beat the baseline by more than 31% and 22%, when $X = 0$s and $X = 0.5$s respectively.

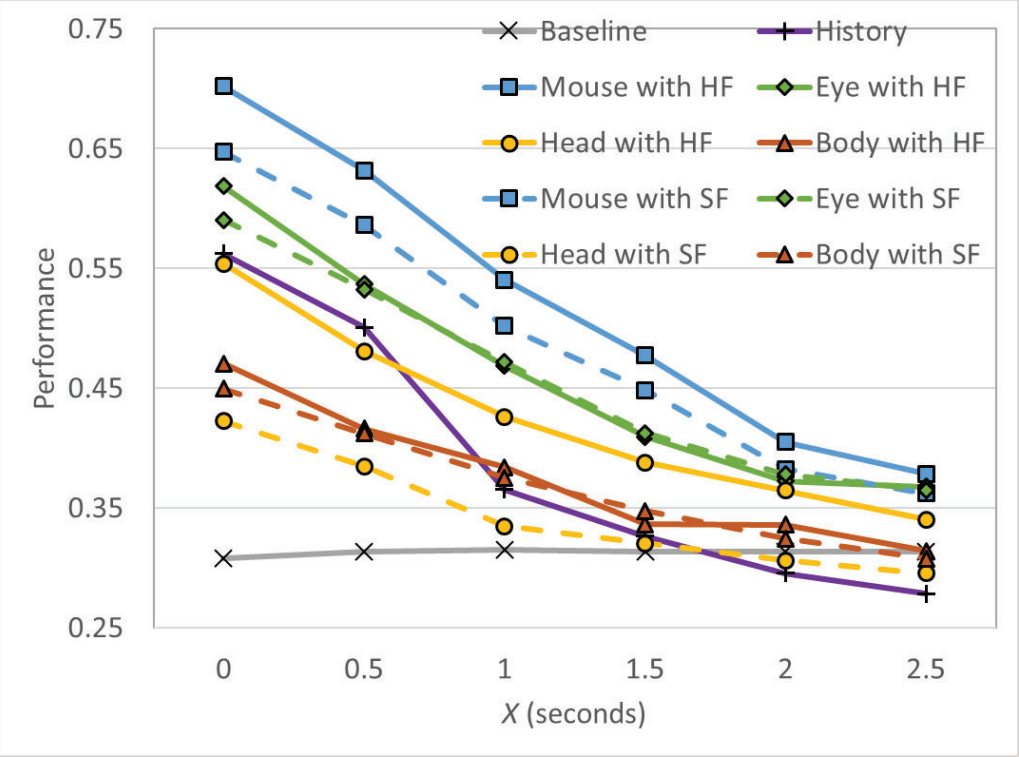


Figure 4-9 Results of predicting users' intention X seconds ahead by using different modalities individually.

It is also interesting to know that our *HF* representation is more beneficial to indicate user interaction intention than *SF* representation across different interaction modalities. The *HF* representation encodes the movement magnitudes, orientation as well as temporal information together, and it thus may well describe the interaction patterns. For instance, consider a mouse movement moving left and then right, and the other one moving right and then left with the same speed and

distance, their *SF* should be very similar. However, they may be triggered by different intentions and the *HF* representation can distinguish them, as it captures the orientation and temporal information. Figure 4-10 demonstrates a particular case that the *HF* representation can describe the interaction pattern more precisely. The *SF* of the two mouse movements depicted in the figure may be similar to each other. However, the *HF* can extract the difference between them.

(a) (b)

Figure 4-10 Examples of mouse movements (a), and their corresponding histogram representations (b). Longer radius means more movement magnitudes occur in that direction, deeper color means movement in that direction occur in more recent. The histogram representations can distinguish the two movements.

We also notice that the *HF* is much more appropriate for encoding mouse and head movements. The performance gain between the two features is around 5.5% and 13.2% for mouse and head movements respectively. Since our *HF* representation encodes the orientation and temporal information along with movement magnitudes together. From these phenomena, we can learn that the orientation and temporal information of mouse and head movement are important indicators for interpreting user intention.

4.4.3 Going towards Multi-modalities

Our previous experiments have shown that mouse, eye gaze and head movements, as well as body motions are all helpful for indicating user intention, especially for the situations when the prediction moment is close to the occurrence moment of the next activity. However, we believe that the performance of the prediction model can be improved by modeling all the signals extracted during human-computer interactions together. We then further investigate into combing all the interaction and body signals as well as historical information together. In our study, we first tried to model the multi-modalities data by directly fusing all the features together in the feature level. With this strategy, we trained our prediction model once on all the modalities described above. Besides, we also tried to fuse the different modalities in the decision level. In this way, we trained multiple prediction models on each of the feature sets. Then we adopt a weighted average approach to fuse the final prediction. In our study, we decided the weight for each modality based on the performance obtained from itself only. Since the modality that can predict a better performance may contain more useful information, it should contribute more to the model. Given the set of modalities I, we then compute the weight for each modality by the equation:

$$w_i = \frac{c_i}{\sum_{j \in I} c_j} \qquad 4.5$$

where w_i is the weight of modality i, while c_i is its correct classified rate (CCR) , when it validated on the validation set. For comparison, we also try another decision fusion approach that is using equal weight for all modalities. We refer to the two approaches as decision level with “dynamic weight” and decision level with “equal weight” in our experiments.

As demonstrated in our previous experiments, the *HF* features are more useful

for encoding interaction and body signals. We then adopt the *HF* as our feature representations in this experiment. According to Figure 4-11, we observe that applying multi-modalities helps in improving performance. The performances of different fusion methods are quite similar; however, applying the approach of fusing in decision level with dynamic weight can achieve a better performance.

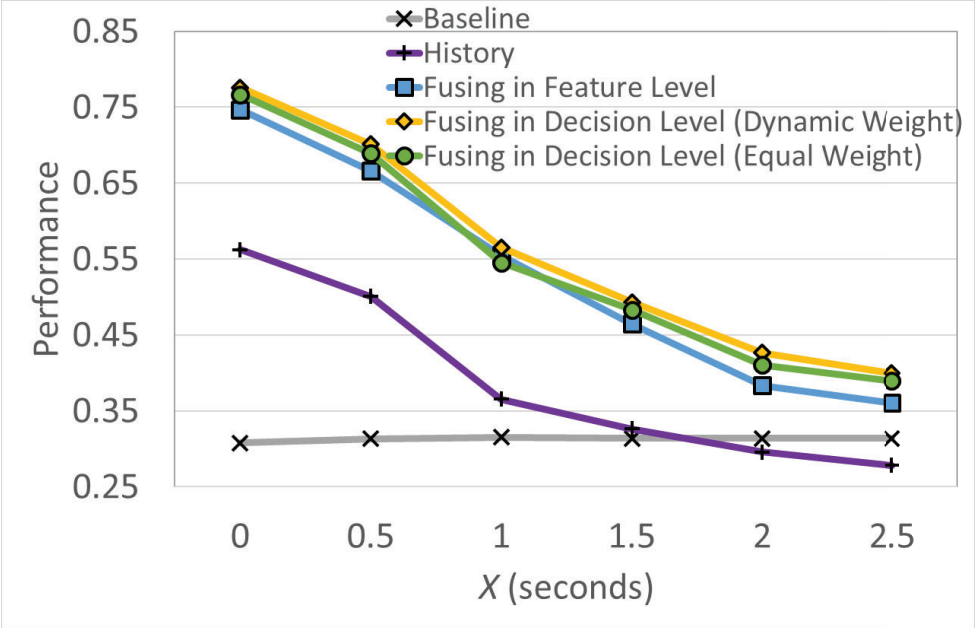


Figure 4-11 Results for predicting users' intention X seconds ahead with multi-modalities.

For $X = 0\text{s}$ the best performance of multimodal approaches can beat the model using only the historical information by 21.3%, and can beat the model using only mouse interaction information by 7.4%. While for $X = 0.5\text{s}$ the multimodal approach can beat the model using only the historical information by 20.0%, and can beat the model using only mouse interaction information by 6.9%. These results further illustrate the feasibility of modeling multiple interaction and body signals to predict user intention in computer interaction tasks.

So far, the prediction performance is encouraging. However, in our current approach, we extract gaze information through the use of a dedicated Tobii eye tracker device. We refer to this method as the Tobii method. Although the device

is non-intrusive, we still do not want to rely on it as it is not affordable for common users. We therefore try another method to estimate gaze information from webcam with the help of OpenFace toolkit [7]. We refer to this method as the webcam method.

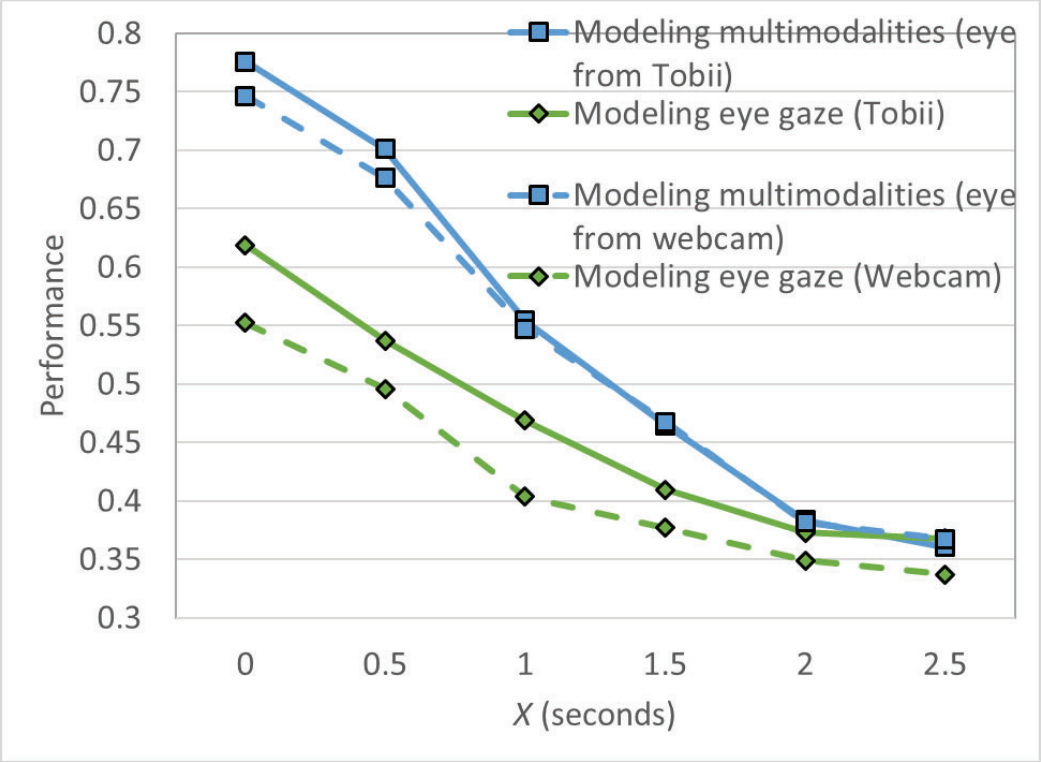


Figure 4-12 Results for predicting users' intention X seconds ahead with different gaze estimation methods.

Based on our previous experiments, we apply appropriate time window, $W = 2$ and *HF* features here. Figure 4-12 shows the results of using different gaze estimation methods. It is encouraging to see that using the gaze movements estimated from webcam can still achieve reasonable performance, even though not as good as that of using Tobii eye tracker. The performance of modeling individual gaze movements as well as multi-modalities for the webcam method are 55.2% and 74.6% respectively for $X = 0$s, which are close to those of the Tobii method.

If we consider the model of using all the reasonable modalities, the performance gap between with and without using Tobii is around 3% and 2.4% for $X = 0\text{s}$ and $X = 0.5\text{s}$. And there is almost no difference between the model with and without using Tobii when $X \geq 1\text{s}$. The experiment results suggest that it is feasible to build effective user intention prediction model by using non-intrusive and low cost devices.

4.5 Towards User Slips Detection

Our model can successfully detect a user's next intended activity. The ability to predict users' intention could be used to enhance an intelligent system in multiple applications. Detecting user slips could be one of the potential applications. To detect a user's slips, a system should be able to distinguish the intended and non-intended activity. When a non-intended activity is triggered, the user is probably making a slip, such as clicking an unexpected link by accident. Such a system should be able to understand user's intention which has been studied in our previous experiments. We now are interested in studying whether our intention prediction model can be applied to detecting user's slips. This section presents the pilot study that we conducted to investigate the feasibility of intention-based user slips detection by using our intention prediction model.

As the very first step of intention-based user slips detection, we design and conduct a toy experiment to study slips detection in a simulated scenario with the same web search task. In this experiment, we simplify the slips detection to estimate the likelihood of an interaction event. For instance, there is a mouse click event that triggers an activity a_i, which belongs to one of the five web search activities described in Section 4.1. To detect whether the mouse click event is a user slip, we first extract the feature vectors fv from the interaction and body

signals preceding the click event within the time window W. We then feed fv and a_i to our model and obtain the value of $\mathrm{p}(a_i)$, which is the likelihood of activity a_i estimated by our intention prediction model. We then compare $\mathrm{p}(a_i)$ with a pre-selected threshold p_t. If the value of $\mathrm{p}(a_i)$ is smaller than or equal to p_t, then the activity a_i is classified as a non-intentional activity and the mouse click event is detected as a user slip, otherwise it is considered as a correct action.

We then use this approach to perform our intention-based slips detection. For the evaluation purpose, we then produce slip click instances in our toy experiment. In the experiment, we randomly select 50% of the test instances in the test set and randomly assign a different activity class to these instances. This gives us a slips detection baseline of 50%.

Figure 4-13 illustrates the results of our slips detection with different thresholds p_t. According to the results, We observe that when $p_t = 0.3$, the model achieves the best balance between precision and recall. Our slips detection model can reach a performance of 89.4%, which is encouraging. However, it is just a very first step towards user slips detection. In the future, we will try to study the user slips detection in real applications.

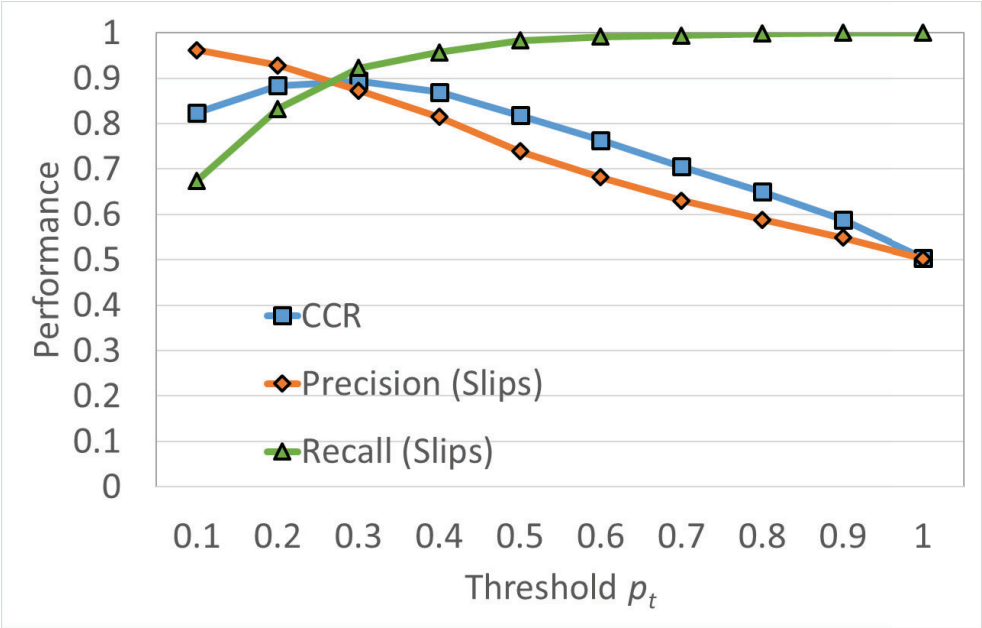


Figure 4-13 Results for user slips detection.

4.6 Summary

This chapter presents several multimodal approaches to predict user interaction intention in a natural web search task. In the study, we proposed two feature representations: a statistics-based feature set and a histogram-based feature set to encode users' interaction and body signals including mouse, eye gaze and head movements as well as body motion signals. We then utilize this information combined with historical activity sequence to build multimodal user intention prediction models.

To evaluate, we collected our user intention dataset in the scenario of web search. In our study, we are more interested in the intention prediction performance when the next interaction activity is closing. Specifically, we evaluated our model in the situation of a few seconds (0s - 2.5s) preceding the occurrence of the next activity. The experiment results show that our proposed approaches can achieve encouraging performance in predicting user intention.

In this study, we also investigate user intention prediction by modeling

individual modality with different feature representations. The experiment results show that all of the proposed modalities contain some information about user interaction intention. Compared with other modalities, modeling mouse interaction can achieve the best performance. Modeling mouse interaction with the histogram-based feature can attain an accuracy of 70.2% and 63.2%, for predicting 0 and 0.5 seconds ahead. The results also indicate that our proposed histogram-based feature representation is more proper for describing interaction patterns and predicting user intention, compared with the statistics-based features. We further evaluate the prediction model of using multi-modalities. The experiment results suggest that performance can be improved by fusing multi-modalities, and fusing the modalities in the decision level with our proposed approach can achieve the best prediction performance. The best performance can be around 77.6%. The experiment results also suggest that it is possible to detect user intention by using non-intrusive and low cost devices.

Finally, we also applied the intention prediction model to detect user selection slips. The experiment suggests that users' intention prediction model has a good potential as a means of providing information to the intelligent system. Besides, corrective action in the event of a user slip could also be taken to improve user experience.

Chapter 5 Physiological Mouse - Non-intrusive Measurement of Physiological Signals

Selected notations and abbreviations used in this chapter

B_H	frequency band that we adopt for heart beat rate calculation
B_R	frequency band that we adopt for respiratory rate calculation
C	candidate frequency for computing respiratory rate
D_i	heart beat rate series measured by iHealth device for subject i
$d_{i,t}$	t^{th} heart beat rate in D_i
E	average heart beat rate error series for all subjects
E_i	heart beat rate error series for subject i
$e_{i,t}$	t^{th} error value in E_i
f	sampling frequency of raw signal, default value is 200 Hz
F_H^*	peak power frequency of signal L', after applying band-pass filter B_H
F_R^*	peak power frequency of signal H', after applying band-pass filter B_R
F_C	candidate peak power frequency for computing respiratory rate
F_j	power frequency in local region of F_{C_i}
F_{C_i}	i^{th} candidate of F_C
H	heart beat rate variability signal
h_i	i^{th} data point in H
H'	equally-spaced interpolated series of H
h_i'	i^{th} data point in H'
hw	half window size of the moving window for computing the smooth series, default to 10 data points
L	input series of raw signal

l_i	i^{th} data point in L
L'	series of smoothened signal of L over the moving window
l'_i	i^{th} data point in L'
M_i	heart beat rate series measured by physiological mouse for subject i
$m_{i,t}$	t^{th} heart beat rate in M_i
n_s	number of subjects
N_T	size of neighborhood for extracting local maxima power
P_T	threshold for extracting candidate frequencies, default to 5% of total power
R_C	respiratory rate candidates
R_{C_i}	i^{th} candidate in R_C
w	moving window size for computing the smooth series, default to 21 data points
W_H	moving window size for computing heart beat rates, default to 5 seconds
W_R	moving window size for computing respiratory rates, default to 60 seconds

IBI	abbreviation of inter-beat interval
PPG	abbreviation of photoplethysmographic

In previous chapters, we attempted to build user intention prediction models by using multiple interaction and body signals, including mouse, gaze, head and body movements. Apart from these signals, physiological signals may also reveal human emotion or even intention and are useful in user intention understanding. However, traditional methods of measuring physiological signals rely on intrusive and expensive devices, which make measuring physiological signals inconvenient in daily computer usage for common users. Moreover, intrusive measuring methods may affect users' behavior and even emotions during the measurements. A better way is to measure the physiological signals without users feeling of the existence of the devices. This chapter presents the design and prototype construction of a physiological mouse to measure human physiological signals by using non-intrusive and low cost devices.

We enhance a daily-used mouse by some low cost components to capture photoplethysmographic (*PPG*) signal. Based on the *PPG* signal we then develop algorithms to measure heart beat rate and respiratory rate. A user's physiological signals can be easily detected while he/she is holding on and using the physiological mouse. This chapter describes the detailed design of the physiological mouse, measuring algorithms as well as the evaluation experiments in this study.

Moreover, to investigate on the feasibility of determining human emotions by using the physiological mouse, we also invited subjects to use the physiological mouse while they are watching movies and playing games. The mouse measures the *PPG* signals of the subjects and determines their heart beat rate and respiratory rate throughout the tasks, in the form of data sequence. We then conduct a pilot study to correlate the physiological signals to the subjects' emotions.

The rest of this chapter is organized as follows. Section 5.1 describes the

design and construction of the physiological mouse prototype and the devices involved. Section 5.2 introduces the algorithms of measuring heart beat rate and respiratory rate by using the physiological mouse through the processing of the acquired photoplethysmographic signals. Section 5.3 presents the experiments for evaluating our proposed approaches and the corresponding results. Then Section 5.4 presents the pilot study of correlating the physiological signals with human emotions by using the physiological mouse. Finally, this chapter is concluded in Section 5.5.

5.1 Physiological Mouse Prototype

One potential solution of non-intrusive physiological signal measurement is to measure physiological signals via a standard device of personal computers. Mouse is one of the standard input devices when people are using personal computers. It would be better if users' physiological signals can be measured when they are using a mouse. We therefore enhance a daily-used mouse to capture human physiological signals.

In this study, we built a prototype of the physiological mouse for both proof-of-concept and validation of our physiological signals capturing and processing algorithms. Figure 5-1 (a) demonstrates the prototype. We attach a small light sensor (a photodiode) to the left side of the mouse (red box in the figure), where the thumb of the user is placed. An infrared light emitting LED (yellow circle in the figure) is attached next to the sensor. The device picks up and relays the intensity of the reflected infrared by the thumb when the user is holding the mouse, as demonstrated in Figure 5-1 (b), via a connected Arduino board. Essentially, this is equivalent to attaching a second light sensor and light source to the side of the mouse (the first pair is at the bottom of a conventional optical mouse) and this can

be easily integrated by product engineers.

(a) (b)

Figure 5-1 The prototype of the physiological mouse (a) and its usage (b).

After attaching these optical components to our prototypical physiological mouse, we are able to capture the *PPG* signals by using the mouse. The infrared LED sends off infrared light, which is blocked and reflected by the user's finger when the mouse is in use. We use the photodiode light sensor to measure the intensity of the reflected light, which will vary over time. This light intensity signal is then analyzed and physiological signals can be extracted. Currently, the response time for the sensor is 5ms, which means that we are able to capture 200 readings per second, i.e., 200 Hz. The range of the light intensity reading returned by the sensor is $l_i \in [0, 1023]$.

We record the intensity of the reflected infrared light and the 200 Hz time series signal is passed to a connected computer for processing. We use modeling clay to hold the gadget together. Though there are wires coming out of the mouse and the use of modeling clay does not look nice, users do not have a bad feeling when using it in general. This is a good proof-of-concept feedback. We believe that the market will react to the availability of useful or interesting technology. Had the idea of physiological mouse become well-accepted, product engineers will design a more user-friendly physiological mouse, as well as creating small

add-ons to transform a normal mouse into a physiological mouse. They would also be able to integrate more sensors within the device or add-on in order to capture additional inputs for processing into other physiological signals, for instance, temperature sensor and skin conductivity sensor.

5.2 Measuring Physiological Signals via Physiological Mouse

After building the physiological mouse, we then investigate the approaches of measuring human physiological signals through the mouse. In this study, we focus our signal processing work on two key physiological measures, namely, heart beat rate and respiratory rate (per minute). Since the infrared signal is monochromatic, it is not necessary to consider the more complex RGB signals.

5.2.1 Measuring Heart Beat Rate

In order to compute the heart beat rate, the first step is to clean the input signal by applying a smoothing function. The frequency of the input signal L is $f = 200$ Hz. We employ a moving window approach, with half window size $hw = 10$ (window size $w = 21$). Given an input series of raw signal $L = < l_1, l_2, \dots, l_n >$, we compute the smoothened series over the moving window, $L' = < l'_{hw+1}, \dots, l'_{n-hw} >$, where $l'_i = \sum_{j\in[-hw,hw]} l_{i+j}/w$. This moving window smoothing admits an incremental evaluation upon computingl'_i, with $l'_{i+1} = l'_i + (l_{i+hw+1} - l_{i-hw})/w$. This is illustrated in Figure 5-2.

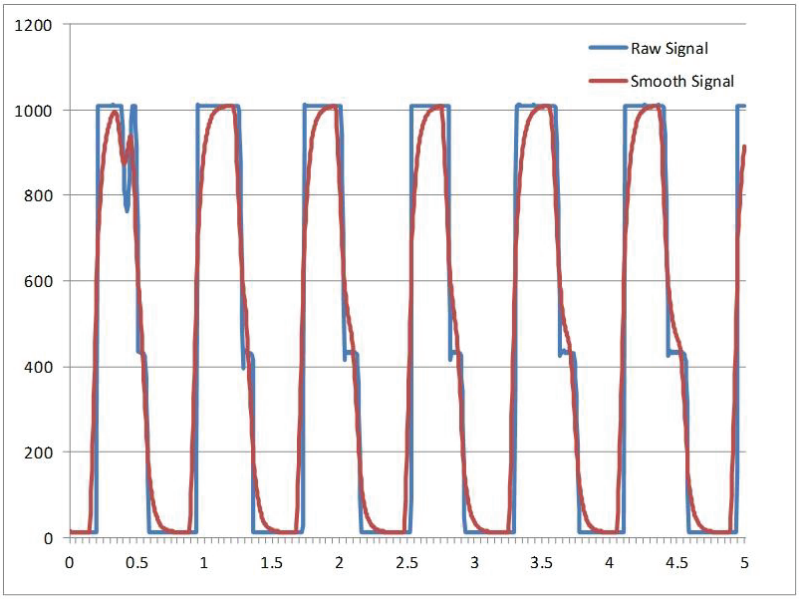
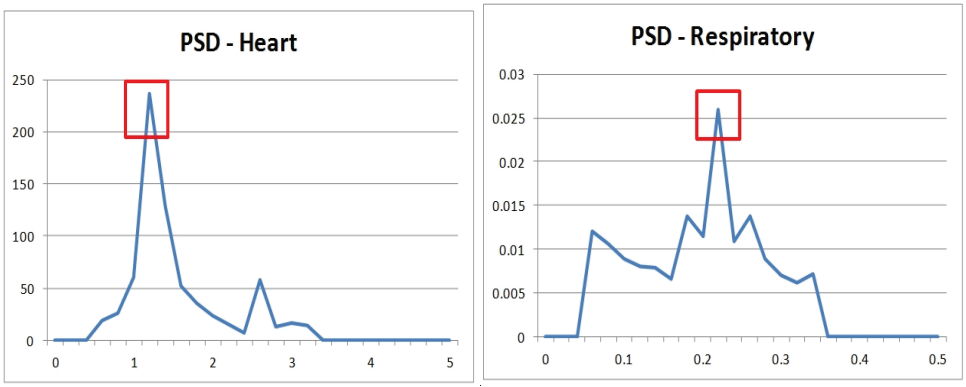


Figure 5-2 Signal smoothing.



(a) (b)

Figure 5-3 Frequency domain of heart beat (a) and inter-beat interval (b).

The *PPG* principle states that periodic changes in signal intensity are manifested by the cardiac cycle. As such, we need to extract the dominant frequency from the smoothened signal. We remove signals of extreme frequency induced by noise, and retain those signals of a proper frequency band that we are interested in, by employing a band-pass filter. The frequency band that we adopt for heart beat rate is $B_H = [0.5 \text{ Hz}, 3.5 \text{ Hz}]$, corresponding to 30 to 210 beats per

minute. Even for champion athletes, it is uncommon to record a heart beat rate below 30. Similarly, a heart beat rate above 210 is unlikely in humans.

To compute a continuous sequence of heart beat rates, we adopt a moving window of size $W_H = 5$ seconds on the smoothened signal L'. Upon acquiring $W_H \times f = 1000$ data points spanning 5 seconds, we start to evaluate the heart beat rate based on these 1000 data points. To perform filtering based on the desired band, we apply a Fast Fourier Transform (FFT) on the signal, transforming it from the temporal domain to the frequency domain. It is then easy to filter out unwanted frequency components. The raw FFT signal will be trimmed according to the passing band and in our situation, $B_H = [0.5 \text{ Hz}, 3.5 \text{ Hz}]$, as shown in Figure 5-3 (a).

Finally, we analyze the power spectral density of the smoothened signal and extract the one yielding the maximum power, via the Welch periodogram method [108]. This peak frequency, F_H^* is taken to be the heart beat rate (in Hz). The heart beat rate is then scaled for reporting as $60F_H^*$ (per minute). By sliding the moving window W_H, we are able to compute the heart beat rate throughout the experimental period in the form of a time series.

5.2.2 Measuring Respiratory Rate

Since respiration does not directly manifest itself in the periodic heart beat signal, we cannot directly extract respiratory rate by applying a band-pass filter corresponding to the potential range of respiratory rate on the raw *PPG* signal. However, since respiration corresponds closely to the high frequency component of the heart beat rate signal variation [12], we make use of heart beat rate variability, in the form of inter-beat interval (*IBI*) as the key signal to determine respiratory rate. *IBI* measures the timing difference between successive heart beats, and *IBI* fluctuation is known to be useful in characterizing respiratory sinus arrhythmia,

which is a cardiorespiratory phenomenon in phase with the inhalation and exhalation of the breathing process [12].

To compute *IBI*, we detect the peaks in the smoothened signal L' that represent the physical heart beats and measure the timing difference between successive peaks as *IBI* values, to generate the heart variability signal $H =< h_1, h_2, \dots, h_n >$. We then analyze H to extract the respiratory signal. Owing to the uneven distribution of the data points in this *IBI* signal H in the temporal domain, we perform interpolation to obtain an equally-spaced series $H' =< h'_1, h'_2, \dots, h'_n >$. When the respiration rate changes, the high frequency peak also shifts accordingly [12]. We are interested in the higher frequency component of H', which indirectly measures the respiratory rate. We make use of an appropriate band-pass filter: $B_R = [0.15\text{Hz}, 0.4\text{Hz}]$, to extract the higher frequency component of H', after applying a FFT on it. This band-pass filter represents a reasonable range of respiratory rates between 9 to 24 breaths per minute. Figure 5-3 (b) shows the peak power frequency obtained after the band-pass filter $B_R = [0.15\text{Hz}, 0.4\text{Hz}]$ has been applied on the frequency domain.

We then perform spectral analysis to locate the peak power frequency F_R^*. We adopt a moving window $W_R = 60$ instead of $W_H = 5$ to compute the respiratory rate. A longer window is required for computing respiratory rate, because respiratory rate is much slower than heart beat rate, which means that there is a longer latency in the detection of the respiratory rate, as compared to the heart beat rate. Due to the possibility of spectral power spreading, instead of directly locating for one single candidate with a peak power frequency, we identify a cluster of strong candidates and apply a smoothing filter to locate the dominating powerful group. This gives due credit to a cluster of neighboring frequencies of high power, which is more representative than a single frequency with an even

higher power but without any supporting neighbor with high enough power. To be precise, for each moving window W_R, we locate candidate frequencies corresponding to local maxima power. To qualify as a candidate C_i, the power of the frequency F_{C_i} must be a local maximum within a region N_T, which has power no less than a threshold P_T. In other words, it must possess a local maxima power around its neighborhood of size N_T:

$$\text{Power}(F_{C_i}) \geq \text{Power}(F_j), F_j \in [F_{C_i} - N_T, F_{C_i} + N_T] \quad 5.1$$

Upon locating a sequence of candidate peak power frequencies $F_C =< F_{C_1}, F_{C_2}, \dots, F_{C_n} >$ for each frame, we normalize it to the respiratory rate (per minute) by multiplying it by 60 to yield $R_C =< R_{C_1}, R_{C_2}, \dots, R_{C_n} >$, the respiratory rate candidates. As the moving window slides, we accumulate the different respiratory rate candidates into a histogram. Finally, we apply an average filter to obtain the peak candidate for respiratory rate, which reflects a dominating group of frequencies possessing the highest power. In our experiments, we adopted $P_T = 5\%$ of the total power and $N_T = 0.02$. Too high a value for P_T would limit the size of the candidate set, while too low a value would not be effective in discarding weak candidates. Too high a value for N_T would reduce the size of candidate set, while too low a value would have generated too many candidates. This is illustrated in Figure 5-4, where the peak frequency corresponding to the respiratory rate can be identified in the histogram.

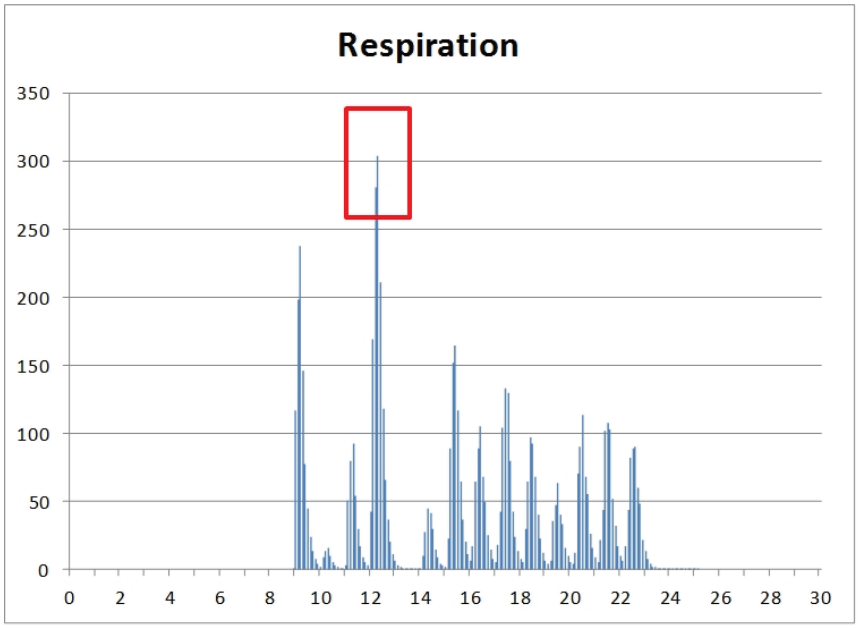


Figure 5-4 A histogram for respiratory rate candidates.

The standard algorithm for determining respiratory rate assumes the availability of readings for the whole period. In particular, the algorithm of adopting a histogram works well when all data have been captured and stored offline. It provides a more accurate estimate for the respiration rate over time. However, in real applications, it is important to generate continuous physiological signals online, even before sufficient data is available. We thus adapt the algorithms to produce continuous outputs once some data become available. There is a tradeoff made when adapting the algorithm to the real time setting in terms of the length of the warmup stage. During the initial warmup stage, the computed respiratory rate would be inaccurate and fluctuate much due to insufficient data. This can be illustrated in Figure 5-12, where a delay period of 3 seconds is adopted, so that we obtain our first reading after 3 seconds. The initial fluctuation is quite significant, potentially extending for a few more seconds. However, the signal will become more stable and can track the respiratory rate as time goes on, especially after the first real breath is taken by the human subject. For practical purpose, it

would be sufficient to remove the readings for the first 10 seconds and adopt the remaining physiological signal series for emotion analysis purpose, since one would normally not expect the system to respond instantaneously, as we would wait for a computer system to "boot", a smartphone to "turn on" or an application to "start". In our evaluation experiments, we remove the first 3 data points for a warmup stage of 9 seconds.

5.3 Evaluating Physiological Signals Computation

We conduct experiments to evaluate the accuracy of the physiological signals obtained when users use the physiological mouse. In the experiment, we invite 8 subjects, 5 males (Subjects 1 to 5) and 3 females (Subjects 6 to 8), to participate in our experiments using the mouse. The ages of the subjects range from 20 to 30 and they are all university undergraduate and graduate students who do not suffer from any underlying chronic illness. We would like to study the viability and validity of the concept and algorithms in deriving physiological signals from *PPG* signals captured via the simple attachment of a small LED and light sensor on to the mouse. We conduct three sets of experiments, the first one to study the performance of heart beat rate computation, and the other two the performance of respiratory rate computation. We present the study in more details in the followings.

5.3.1 Evaluating Heart Beat Rate

The goal of our first experiment is to study the accuracy of the computation of heart beat rate from *PPG* signals acquired by our physiological mouse prototype. To measure the heart beat rate, we use a iHealth Pulse Oximeter [45] that clips onto a finger of the subject. Figure 5-5 shows the heart beat measurements of using iHealth device. The heart beat readings from the iHealth sensor are taken as the

ground truth and compared with the signal returned by the physiological mouse to calculate the heart beat error.

Figure 5-5 iHealth device to measure heart beat rate.

In this first experiment, each subject is requested to use the mouse for 2 minutes. The reading from the iHealth device and the result coming from the mouse are recorded every 3 seconds. This gives us two data series for each subject i, one from iHealth device: $D_i =< d_{i,1}, d_{i,2}, \dots, d_{i,40} >$, the other one from the mouse: $M_i =< m_{i,1}, m_{i,2}, \dots, m_{i,40} >$. We wish to study the trend of sensor readings returned by the two devices. In particular, we compute an error series for each subject over the time period by Equation 5.2, where i is the i^{th} subject. The general trend for the error is computed as the average of all the subjects: Equation 5.3, where n_s is the number of subjects. The results for the average error over time and error for individual subjects over time are depicted in Figure 5-6 and Figure 5-7 respectively.

$$E_i =< e_{i,t} = \left|m_{i,t} - d_{i,t}\right|, t \in [1,40] > \quad 5.2$$

$$E = \sum E_i / n_s \quad 5.3$$

From Figure 5-6, we observe that there is an initial transient impact to the readings due to the warmup effect, and we discard the first three readings in order to remove the bias induced by this warmup effect. After that, the average error is normally below 3 across the board. Similarly, there are very few subjects displaying an error of more than 5 after the warmup stage. We also observe a relatively high error with Subject 6, but when we interviewed with the subject, she

mentioned that her hand was not always holding on to the mouse and it is apparent that part of the signal deviates much from the norm. If this subject were removed from our set of results, the error would drop from 2.90 to 2.65. However, we believe that this is an interesting case to report and do not proceed to request for additional data taking.

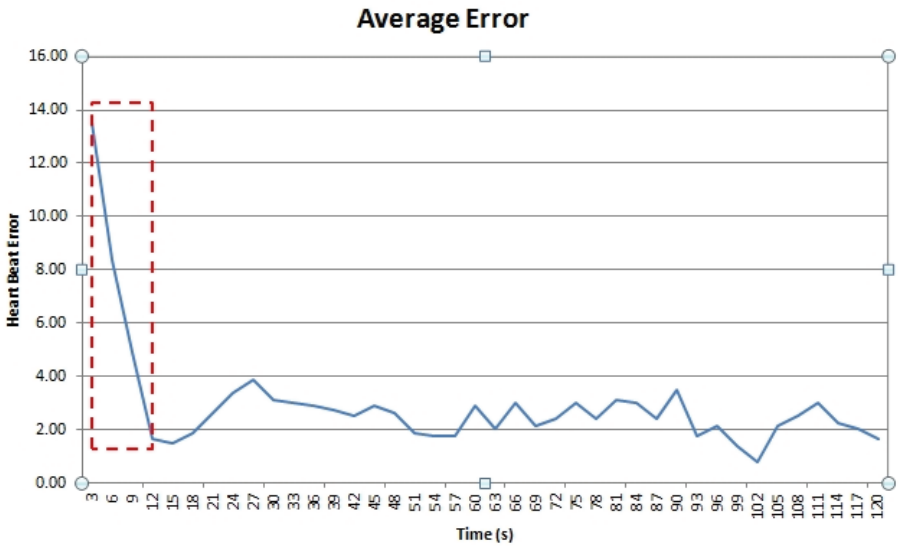


Figure 5-6 Average heart beat rate error for all subjects.

The physiological signals for heart beat rate obtained by our algorithms are compared with the ground truth in Table 5-1. We calculate the mean values of D_i, M_i over time and note their discrepancy, which reflects the error as an aggregate, for each subject, as depicted in Table 5-1. We compute two error metrics. The overall error reflects the error between the mean values of D_i and M_i, and the absolute error measures the average of deviations across all readings. It can be noted that the overall error is at most equal to the actual error, and this occurs when there is a systemic bias in which the reading of one device is consistently higher or consistently lower than the other. In our experiment, we do observe that the physiological mouse is returning a lower reading most of the time, making the actual error close to the overall error. Nevertheless, this error is very small. Besides

the error, we also measure the mean square error (MSE) for each subject, so as to quantify the variability of the errors. In general, we observe small MSE values of at most 10 (except for Subject 6), which means that the error seldom exceeds 3 or 4. We can conclude that the physiological mouse is able to attain a good accuracy for heart beat rate determination, sufficient for emotion recognition purpose.

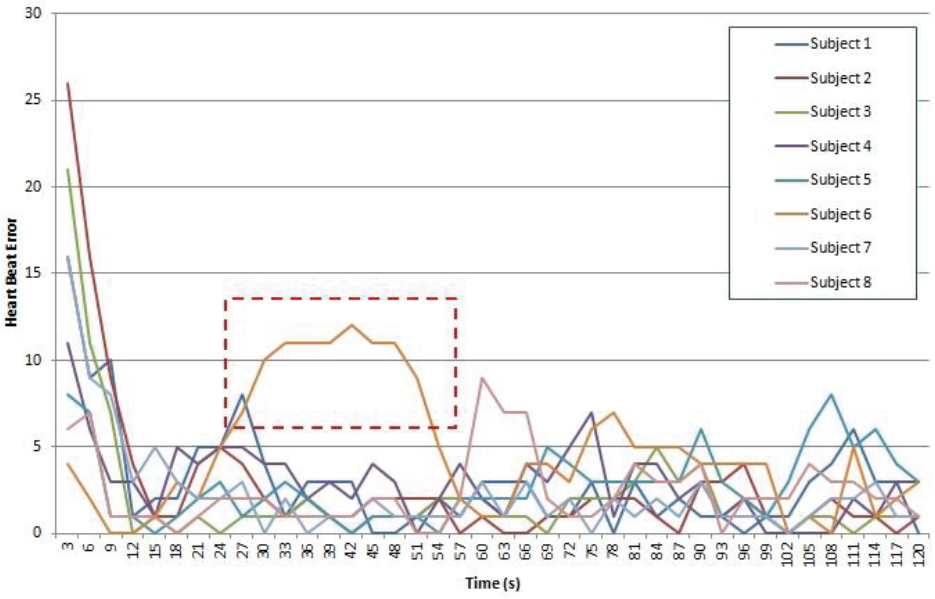


Figure 5-7 Heart beat rate error for individual subjects.

Subject	1	2	3	4	5	6	7	8
iHealth $\boldsymbol{d}$	103.97	65.84	82.92	68.11	86.84	66.27	66.81	80.86
Mouse $\boldsymbol{m}$	102.14	64.30	83.70	65.92	84.24	61.73	65.51	82.73
Overall error	1.77%	2.34%	0.95%	3.21%	2.99%	6.85%	1.94%	2.31%
Actual error	2.19%	2.49%	1.61%	3.89%	3.00%	7.18%	2.43%	2.77%
MSE	8.32	4.35	2.95	10.08	10.38	36.59	3.78	8.78

Table 5-1 Heart beat rate performance.

5.3.2 Evaluating Respiratory Rate

We then conduct experiments to evaluate the accuracy of the computation of

respiratory rate from *PPG* signals. This is more challenging, since the mechanism to derive respiratory rate from *PPG* signals is more complex and indirect, and there is more variation to the respiration pattern exhibited by human beings. Furthermore, unlike heart beat, which is more or less an involuntary mechanism, respiration is controllable by a human to a certain degree and breath holding is not an uncommon phenomenon.

We evaluate our respiratory rate measurement via two sets of experiments in line with [91]. The first set of experiments measures the respiratory rate under a controlled environment via a metronome. We implement a simple metronome by displaying an inhale and exhale indicator at a given frequency, and request our subjects to breathe according to the rhythm. The second set measures respiration in a more natural context via self-reporting. Subjects are asked to breathe naturally over a period of time. For the measurement, unlike [91], which makes use of an intrusive respiratory belt fastened around the chest of the subject to measure respiration, we request our subjects to press a key on the keyboard on every inhale and exhale. The timestamp of each keypress then gives us the actual respiration events and hence the respiration rate. While the respiration rate induced by the metronome is constant in the controlled experiment, the actual respiration rate in the natural experiment exhibits variations. This variation provides us with more room for experimental validation of the accuracy over time and future human affect recognition.

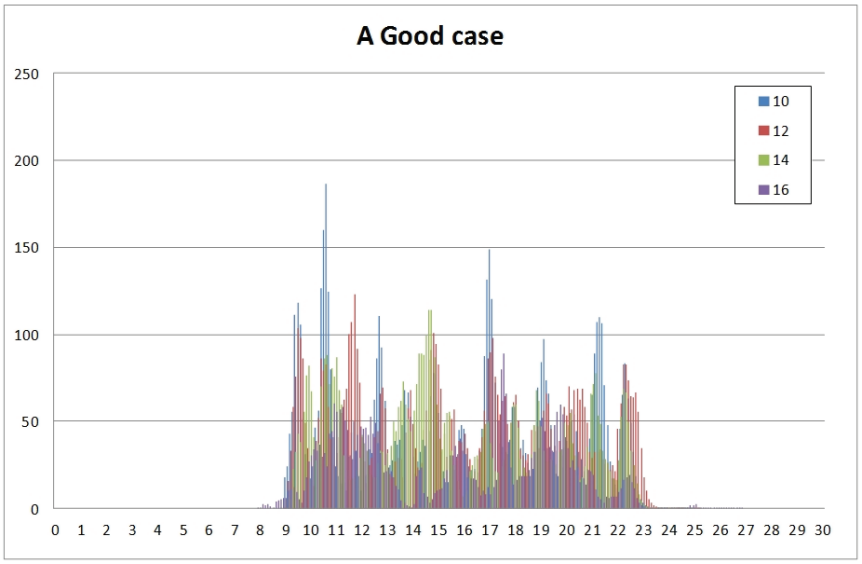


Figure 5-8 Example of candidate respiratory rates in controlled experiment (a good case).

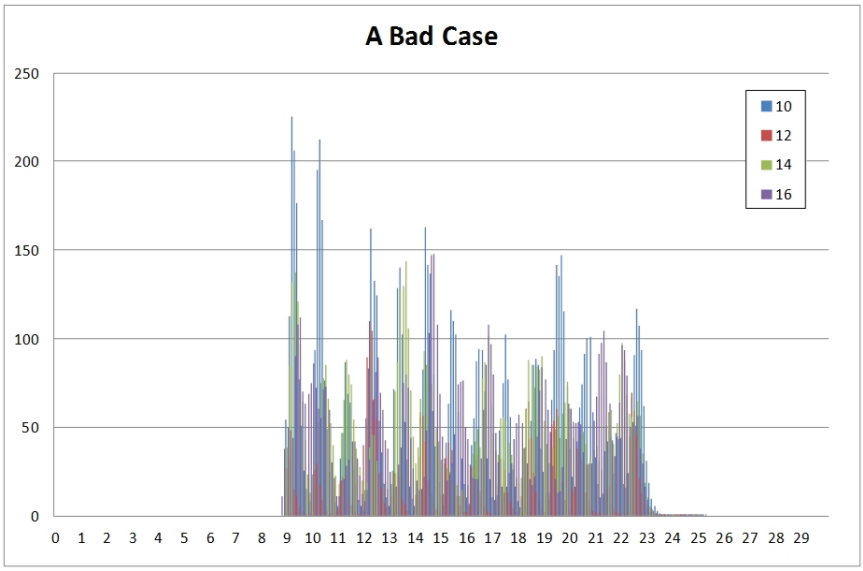


Figure 5-9 Example of candidate respiratory rates in controlled experiment (a bad case).

In our first set of controlled experiments, we ask the subjects to breathe according to the predefined rhythms of 10, 12, 14 and 16 breaths per minute for 2 minutes. The examples of resulting candidate respiratory rate histograms are demonstrated in Figure 5-8 and Figure 5-9, which show the "good" and "bad" scenarios among the subjects respectively. For each subject, the corresponding histogram summarizes the candidate rates for the four rhythms in different colors.

It can be observed that in general, the peak frequency for each rhythm is correct. In the good case scenario, there is just one non-negligible second peak at 14 per minute. The result for the bad case contains non-negligible peaks at 16 per minute and there are also a number of spikes at 12 per minute. The results are summarized in Table 5-2. The error for each specific rhythm is illustrated in Figure 5-10. It can be observed that the error is not high, only ranging from 2.5% to 6%, with an average value of 4.1%. The mean square errors (MSE) are also very low, all below 1.0. The small error rate is attributed to both the stable breathing rhythm and the spectral analysis performed on the collected offline data. We believe that this set of experimental results would indicate a potentially "best" case scenario for respiratory rate computation from *PPG* signals, as compared with natural breathing situation.

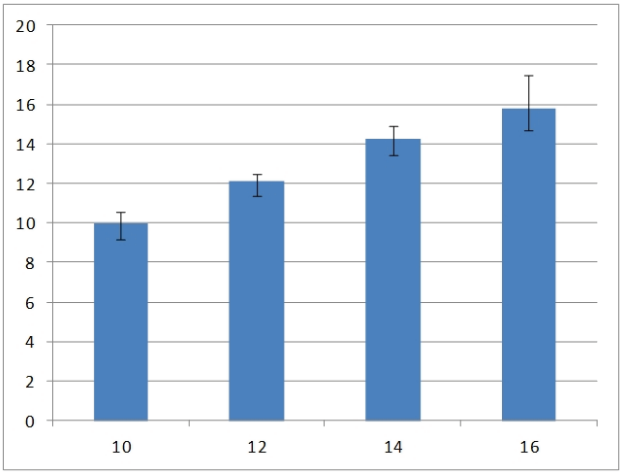


Figure 5-10 Average result with error bar for different respiratory rhythms.

Subject	1	2	3	4	5	6	7	8
Rhythm 10	10.6	10.2	10.6	10.3	9.6	10.3	9.2	9.3
Rhythm 12	12.4	12.5	11.7	12.3	11.8	12.5	12.2	11.4
Rhythm 14	14.4	14.4	14.6	14.3	14.9	14.4	13.6	13.4
Rhythm 16	16.4	15.5	17.5	15.6	14.9	16.4	14.7	15.6

Actual error	3.46%	3.08%	5.77%	2.50%	5.00%	3.08%	5.19%	4.42%
MSE	0.21	0.17	0.76	0.11	0.56	0.16	0.63	0.34

Table 5-2 Respiratory rate performance: controlled experiment.

Performance results from the second set of natural respiratory experiments are shown in Figure 5-11and Figure 5-12, as captured by the physiological mouse and the self-reported respiratory rate, as well as the average error for each subject. Figure 5-11 summarizes the overall average for all the subjects while Figure 5-12 highlights individual subjects. From these figures, we also observe that there is an initial transient impact to the readings due to the warmup effect. In general, the respiratory rate exhibits a larger error than heart beat rate, since the mechanism for its determination is more complicated.

To make a fair comparison with heart beat rate computation and knowing that the initial transient data would likely be incorrect, we remove the first three data points (corresponding to a warmup stage of 9 seconds) when presenting Table 5-3. The reported rate is computed by measuring the time difference between two consecutive respirations. Since the frequency of respiration is lower than the frequency implied by our 3 seconds reporting interval, there may not be a breath taken in an interval. We thus perform linear interpolation to the respiratory rate when computing for the ground truth. This sampling frequency may also exert some impact on the accuracy of our physiological mouse. Comparing with the heart beat rate, the overall error for respiratory rate is generally smaller than those of the actual error, indicating that the error can go both ways and there is no evidence of the presence of any systemic error. The mean square error (MSE) is also not very high across the board, with a highest value of 9 to 16, i.e., a deviation of 3 or 4 in the worst case.

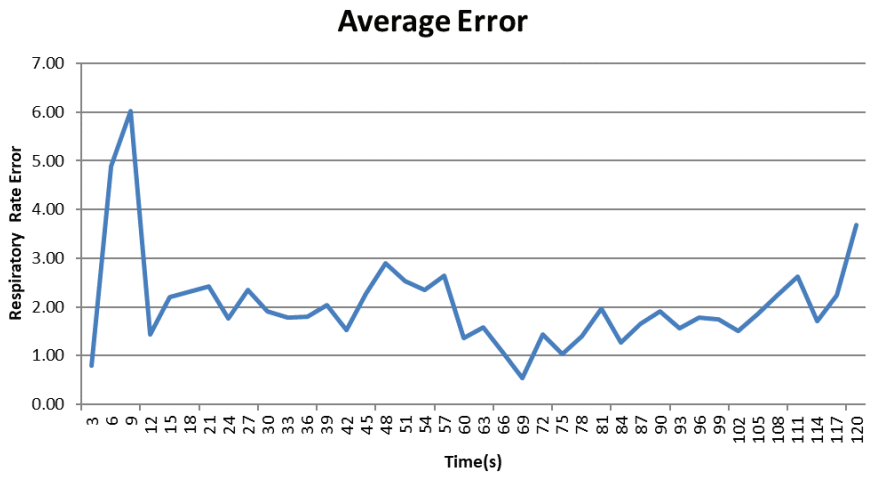


Figure 5-11 Average respiratory rate error for all subjects: Natural respiration.

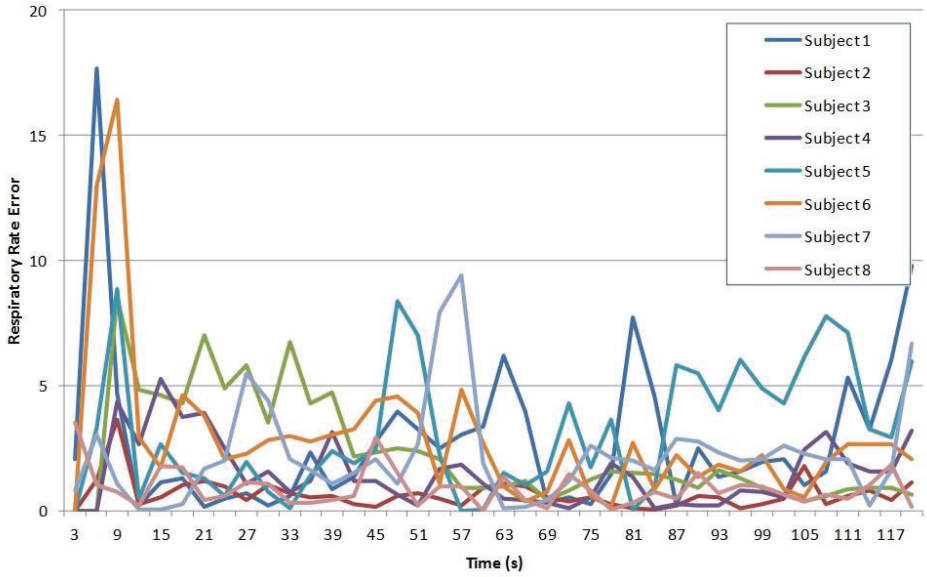


Figure 5-12 Respiratory rate error for individual subjects: natural respiration.

Subject	1	2	3	4	5	6	7	8
Ground rate	16.41	10.05	10.78	10.84	13.92	13.05	12.70	13.08
Mouse rate	18.10	9.84	8.52	10.89	10.97	15.26	12.11	13.26
Overall error	9.34%	2.22%	26.61%	0.51%	26.90%	14.46%	4.90%	1.32%
Actual error	12.69%	5.89%	27.81%	13.43%	28.71%	15.00%	19.01%	6.06%
MSE	10.84	0.47	8.61	3.71	15.09	6.84	9.42	1.04

Table 5-3 Respiratory rate performance: natural respiration.

Though we can obtain very accurate results for heart beat rate as well as respiratory rate under the controlled environment, it can be observed that the results for natural respiratory are not as good, with an average actual error rate of 16.9%, maximum reaching 28.7% and a minimum of 5.9%. This is likely due to the fact that respiratory rate is measured indirectly via some biological phenomenon, which allows noise to set in. The achievable accuracy also varies widely from subject to subject. The overall error rates, as with heart beat experiments, are in general lower and for respiratory rates, much lower. This also implies that the errors can sometimes cancel out each other, without clear presence of systemic errors. We believe that the measured respiratory rate can still be used for emotion recognition, especially when we compute for more aspects of the respiratory rate to form the list of features in recognition, e.g., the average rate over a past window and the rate variation, besides the instantaneous respiratory rate signal.

5.4 Correlating Physiological Signals with Human Emotions

Our pilot study to determine the relationship between physiological signals and human emotions involves two sets of experiments drawn from two frequently performed tasks in human-computer interaction: namely, watching videos and playing games. These experiments involve 8 subjects. The experimental setup is illustrated in Figure 5-13. The subject is requested to hold on to the mouse during the entire experiment.

Figure 5-13 Experimental setup for emotion-related experiments by using the physiological mouse.

In the first set of experiments, each subject is requested to watch two short videos, a funny video of about 2 minutes and a horror video of about 3 to 4 minutes. The funny video is extracted from the famous hidden camera comedy show: “Just For Laughs: Gags", Season 9, Episode 8, between 9'38" and 10'58", whereas the horror video is taken from the movie “Final Destination 5”, running from 22'05" to 26'35". The funny video presents a joke whereby a number of participants are invited to hold a big sign asking for kisses from passers-by. The participants are all men, and the passers-by happen to be attractive women. At the end of the video, a male passer-by appears and attempts to follow the instructions, which leads to a number of humorous moments. In the horror video, a number of teenagers are practicing gymnastics, with some workers repairing the electricity supply nearby. There are a number of “disasters destined to happen” in the scene, including a hanging fan that is about to break loose (scene showing a loosening screw), uneven bars starting to fall apart, and leaking water that is slowly making its way across the ground towards a live wire. The screw that drops from the hanging fan on the ceiling onto the balance beam happens to play an important role in this horror video. It has been shown repeatedly with several near misses by the girl on the

beam, hence building up the suspense. There are also camera pans to the ceiling and the floor to build the suspense. Finally, the girl as expected steps on the screw to trigger a cascading series of accidents which culminate in the killing of another girl by electric shock and a fairly horrifying scene highlighting the death of a third girl, breaking her neck in a miserable way. Indeed, one of our female subjects could not handle the horror scene in the video and opted to watch an alternative video involving a snake charmer (despite the fact that she has a phobia of snakes).

In the second set of experiments, we request our subjects to play a video game using the physiological mouse. The game we utilize is called "House of the Dead", which is a classic first-person-shooter game, set in a haunted mansion with zombies and monsters. The sequence culminates in a battle with a powerful "named mob", whom they have to beat before the experiment is considered complete. Since not all subjects possess the same skill level, the amount of time that they spend to complete the experiment varied from 5 minutes to 11 minutes.

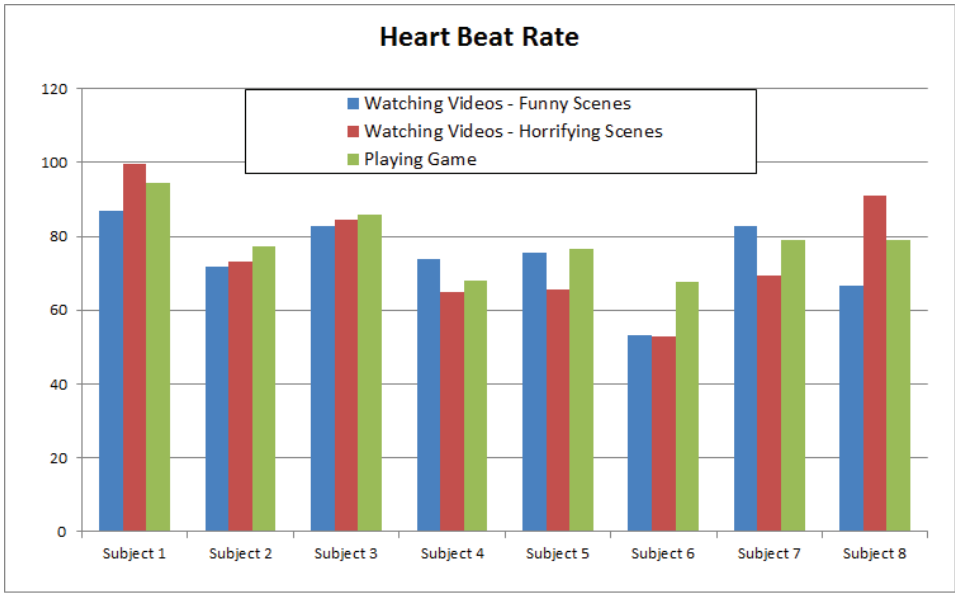


Figure 5-14 Average heart beat rates across subjects in different tasks.

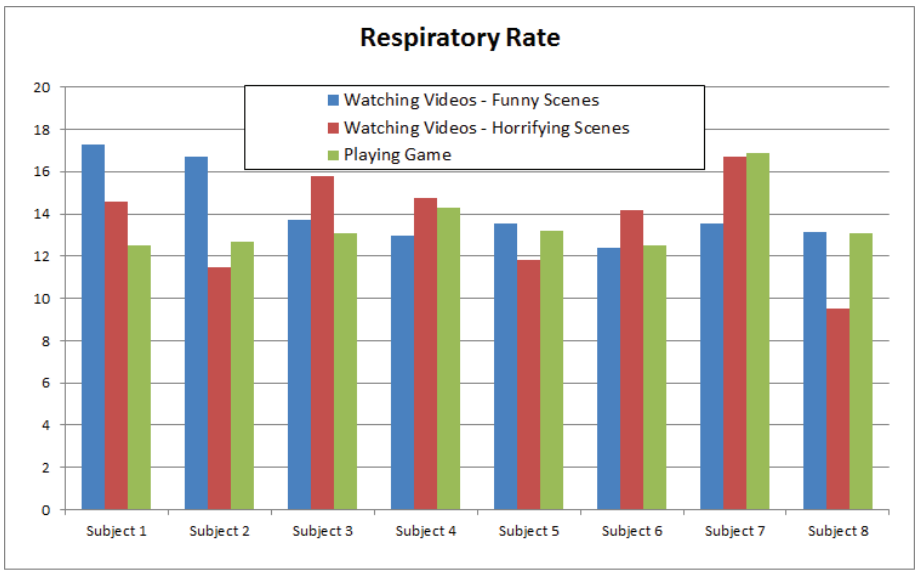


Figure 5-15 Average respiratory rates across subjects in different tasks.

We measure the two physiological signals for each subject and summarize the results in Figure 5-14 and Figure 5-15 for the two sets of experiments. This provides a comparison for a same subject across different activities. In general, we observe that the heart beat rate is highest for the game-playing activity, which is not surprising given the more interactive nature of the activity, demanding continuous attention from the subject. Between the two videos, with the exceptions of two subjects, the horror video does not result in a significantly faster average heart beat than the funny video, and in fact, some subjects' heart rates are actually higher during the watching of funny video than during the horror video. Upon post-experiment interviews, our subjects stated that they were not particularly nervous or tense during the watching of horror video, perhaps as a result of desensitization because they were used to watching such videos. Subject 8 is the subject who had snake phobia (and who chose to watch the video with the snake charmer instead of the horror video). Not surprisingly, her heart beat rate during that video watching session is significantly higher than normal. Subject 1 also does not watch videos often, which means that he is not as desensitized as the others

and more susceptible to physiological changes brought on by changes of emotion.

The respiratory rate, on the other hand, shows less of a pattern than the heart beat rate. It is interesting that a heightened heart beat rate does not automatically result in a faster respiratory rate (as evidenced especially for Subject 8). One potential cause of this is the breath holding phenomenon in the presence of suspense or threat for some human. This is an interesting issue that we intend to further investigate in future work.

For a more in-depth investigation, we measure the physiological signals and align the temporal changes with turning points in the plot-line of the videos. Similarly, in game playing, we align the changes in the heart beat and respiratory rate with events in the game playing. In game playing, we differentiate two different types of events. The first type is when the player is to be engaged, namely, either about to face the boss or about to suffer an imminent death, and the second type is the set of transition cut-scenes that help to narrate the underlying storyline behind the game. We understand that changes would occur in the vicinity of the moment for the key events, i.e., funniest or humorous part in the funny video, horrifying part in the horror video, and the moment of threat in the game playing. We call these the elicitation events. We consider the heart beat rate and respiratory rate 3 seconds before and 3 seconds after the event and measure the change. According to the result, there can be three possibilities: a certain increase in heart beat rate or respiratory rate, a certain decrease, and a negligible change (effectively no change).

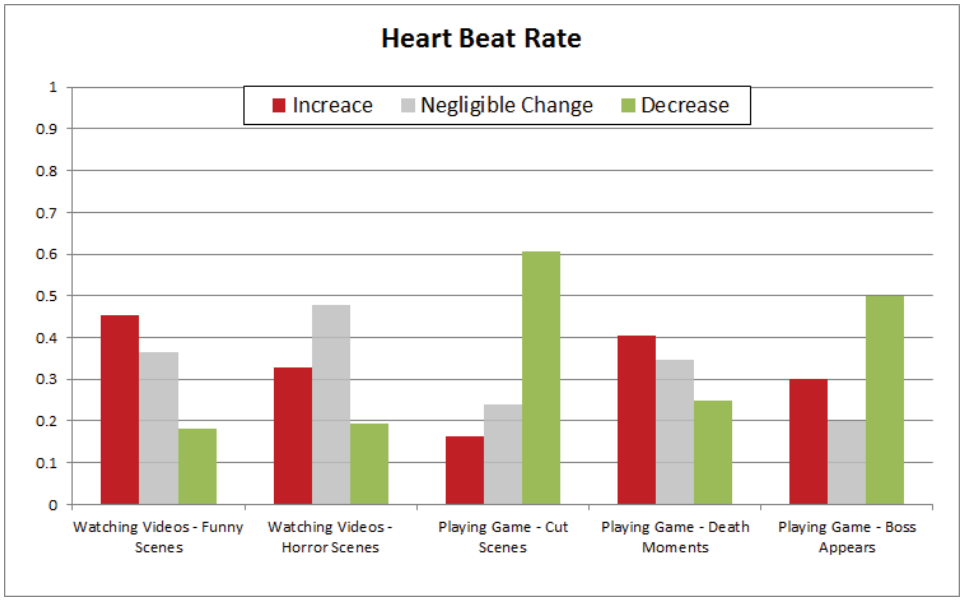


Figure 5-16 Average heart beat rates across tasks.

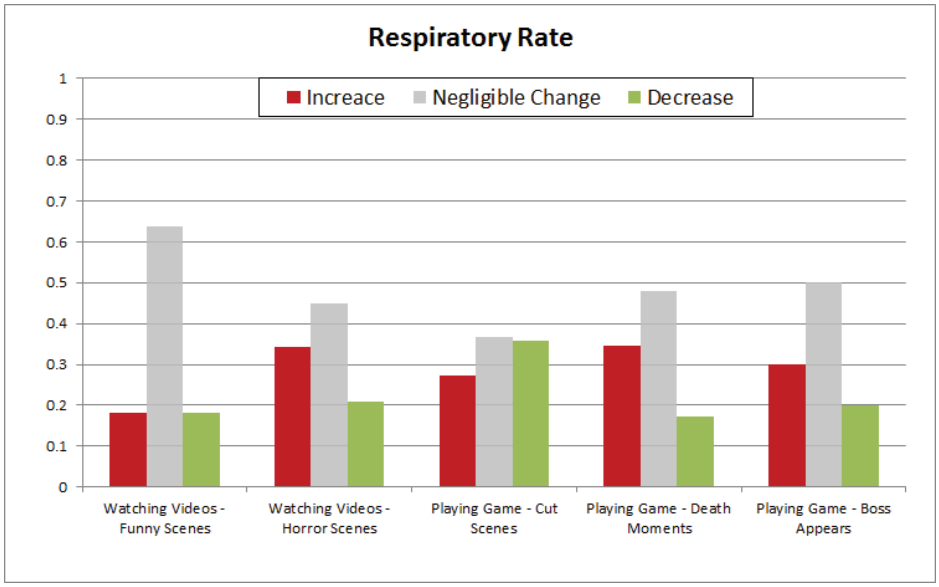


Figure 5-17 Average respiratory rates across tasks.

According to Figure 5-16 and Figure 5-17, unsurprisingly, the transition cut scenes in the game usually elicit a drop in the heart beat rate, which is expected since those scenes usually constitute a break from the constant interactive activity. Similarly, it is also expected that a player facing imminent death would generate a higher heart beat rate, and this seems also to be accompanied with an increase in respiratory rate. What is less expected is that the heart beat rate does not seem to

go up following the appearance of the boss, perhaps because the appearance of the boss is rather expected and that the image of the boss is not particularly scary. The funny scenes also appear to elicit an increase in the heart beat rate, perhaps because subjects often shift their position when laughing out loud, which then increases the heart beat rate momentarily. The respiratory rate does not seem to show much clear pattern.

It is also interesting to analyze the change in heart beat rate and respiratory rate for an individual subject as a function of time. Figure 5-18 demonstrates the example data for one representative subject playing the game, showing the heart beat rate (red curve) against the left *y*-axis and respiratory rate (blue curve) against the right *y*-axis. It can be seen that the transition cut-scenes elicit a drop in the heart beat rate, as evidenced in the overall data. The moments of imminent death (vertical blue lines) seem to be correlated with an increase in the heart rate, which is also expected. The appearance of the boss (vertical purple line), on the other hand, does not seem to result in an increase in the heart beat rate, perhaps due to the fact that the boss appearance is rather expected following the transition scene. The heart rate is highest towards the end of the experiment, perhaps as a consequence of increased tension upon seeing the end of the mission "within sight".

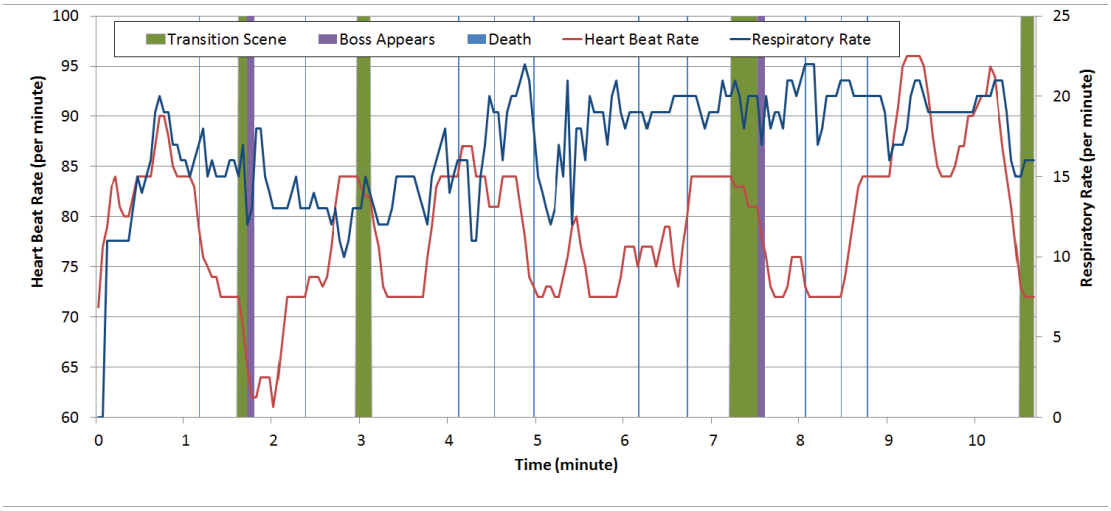


Figure 5-18 An example of temporal physiological signals captured in playing game.

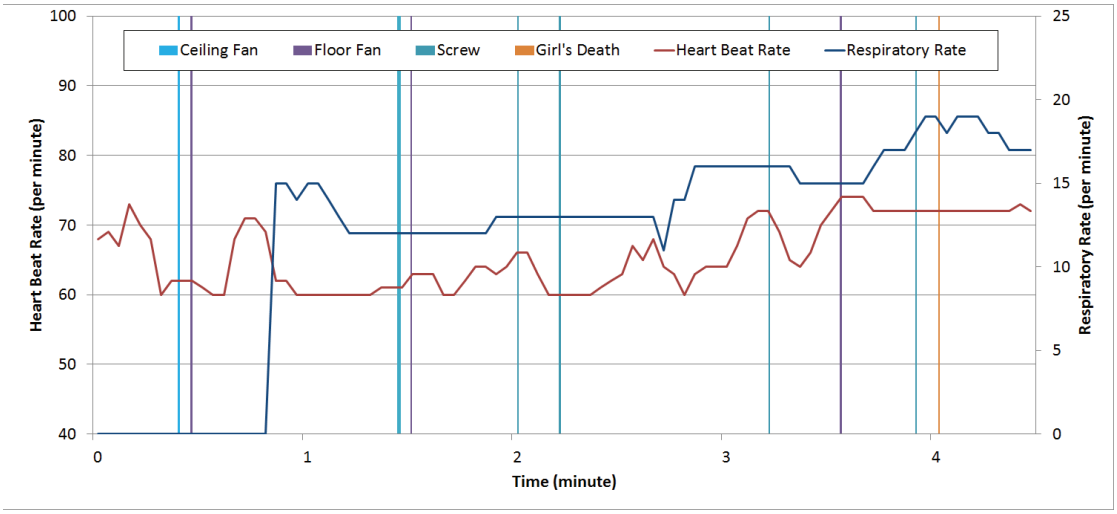


Figure 5-19 An example of temporal physiological signals captured in watching the horror video.

Figure 5-19 illustrates the heart beat rate (red curve) and respiratory rate (blue curve) for another representative subject watching the horror video. The key horrifying events are indicated in the timeline and it can be observed that the heart beat rate increases more often than staying the same when those events happen. Nevertheless, when it comes near the end of the video that the girl is going to get killed, the heart beat rate and respiratory rate both stay at a relatively high level, implying intense felt by the subject. We would be conducting more in-depth study

to correlate the different events with physiological signals and machine learning would be our next direction to pursue.

5.5 Summary

This chapter presents the physiological mouse to measure human heart beat rate and respiratory rate in a non-intrusive manner. We designed and built a prototype of the physiological mouse by enhancing a daily used mouse by some low cost devices.

To verify the performance of our physiological signals computation, we present the evaluation experiments by asking subjects to hold on to the mouse. Experiments demonstrate that our physiological mouse can effectively measure users' heart beat rate with an overall error below 3% and respiratory rate with an overall error below 5% and 10% for the situations of controlled respiratory and natural respiratory respectively.

With the ultimate goal towards developing an emotion-aware and even intention-aware mouse capable of determining the human emotions and intention, we also conduct a pilot study to investigate the relationship between physiological signals and human emotions, by requiring subjects to use the mouse when they are watching videos and playing games. Our future work will investigate that into more various emotions and diversified tasks.

We believe that our study can contribute to affective computing and human-computer interaction, as it provides a novel solution to measure physiological signals during daily computer interaction tasks in a non-intrusive manner, which could be further applied to enhance user experience in real applications in the future.

Chapter 6 Conclusion and Future Work

Understanding human intention has been gaining attention in recent studies in computer science. Once a computer can interpret human intention, it opens up the vast possibility to provide potential assistance to users in advance, in social events and daily human-computer interactions, such as automatically raising an alarm for fierce activities inside a bar, and automatically enlarging a potential target button for a user. This thesis investigates techniques for understanding human intention in different applications.

We investigate human intention in two aspects. We first study fight detection in human social interaction. There are some prior studies about fight detection. However, they are constrained by reliance on high level feature recognition and on simulated fight events. In addition, we also study on predicting user intention in daily computer interaction tasks, facing challenges on limited nature of prescribed tasks, ignoring useful modalities, and need of expensive and intrusive devices to capture signals.

In this thesis, we present a low cost motion analysis-based approach for fight detection. The experiments show that the proposed approach can effectively detect human fights in real surveillance scenarios and can even distinguish human real fights from simulated fights. We developed a multimodal approach to predict users' interaction intention in a nature web search task. The experimental results in this study demonstrate that our approach can achieve promising performance for user intention prediction. We also investigate measuring human physiological signals in a non-intrusive manner and the feasibility of correlating physiological signals to human emotions. This thesis concludes with a summary of the contributions and potential future work.

6.1 Contributions

6.1.1 Detecting Human Real Fight

In this study, we investigated the efficient approaches of using motion analysis for detecting fights in videos, without having to recognize complex behaviors, gestures or action events via video processing.

Our main contributions of this study are summarized as follows:

- We proposed and implemented low-cost and effective motion analysis-based approaches for automatic fight detection;
- We collected real surveillance videos containing real fight events from YouTube and annotated to produce a human real fight dataset;
- We evaluated our approach and compared the performance with the state-of-the-art studies, and the experiment results show that our approach could accurately detect human real fights;
- We conducted a new study in discriminating real fight events against simulated fight events, an issue often overlooked by prior studies, and the experiment results demonstrate that a computer could distinguish real from simulated fight events;
- We proposed a cross-species learning technique for cross-species fight detection;
- We collected a good set of animal fight videos from YouTube and annotated them to produce an animal fight dataset;
- We performed in-depth evaluations of our method, which sheds light on the appropriateness of feature representation for cross-species learning for human fights as well as the effectiveness of adaptation from different sources.

6.1.2 Modeling User Interaction Intention

In this study, we explored the multimodal approaches for predicting user interaction intention by using non-intrusive capturing of users' interaction and body signals. We focused our study on a nature web search task.

Our main contributions of this study are summarized as follows:

- We proposed two feature representations to encode interaction and body signals captured in daily computer interaction tasks;
- We proposed a user intention prediction approach for multi-step human-computer interaction task based on mouse, eye, head and body movements, as well as historical activity sequence;
- We conducted our study and collected user intention dataset in a common daily computer task: web search task;
- We evaluated our user intention prediction model, and results show that our proposed approaches could achieve reasonable accuracy;
- We performed in-depth evaluations of our approaches to investigate the appropriateness of feature representation for modeling user intention;
- We conducted a pilot study to investigate user selection slips detection based on the multimodal intention prediction model.

6.1.3 Non-intrusively Measuring Physiological Signals

In this study, we developed a prototype of physiological mouse to investigate the feasibility of capturing human heart beat rate and respiratory rate in a non-intrusive manner.

Our main contributions of this study are summarized as follows:

- We designed and built a prototype for the novel physiological mouse by

making use of low-cost optical components;

- We proposed to capture the *PPG* signal by using the physiological mouse for measuring human physiological signals;
- We provided algorithms to compute the physiological signals: heart beat rate and respiratory rate;
- We conducted experiments to evaluate the physiological signals computation algorithms for accuracy;
- We conducted a pilot study to investigate the relationship between captured physiological signals and human emotions that we drove the experimental subjects into, via video watching and gaming.

6.2 Limitations

This thesis investigates techniques for understanding human intention. The experimental results of the current studies are promising. However, there are still some limitations.

The first limitation of our studies is the relatively small size of datasets, especially for the study about physiological mouse. As a pilot study towards physiological, emotion and even intention-aware mouse, our experiments only involve 8 subjects. The size of this dataset is reasonable for evaluating the accuracy of physiological signals, but it is not sufficient to gear towards emotion recognition. We would need to increase the number of subjects as well as the size of our datasets for further study of detecting human emotion and intention by using the physiological mouse. For our fight detection and user interaction prediction tasks, our datasets have moderate sizes. However, if we want to develop our approaches by applying a hybrid deep learning approach, we would need to increase the size of our datasets along with our future work, since well-performing

deep learning approaches require a great amount of training data.

In addition, our studies are constrained by using hand-crafted features. Some data-driven approaches such as deep learning have not yet been involved in the current study. As the very first step towards understanding human intention, we developed hand-crafted features not only due to the constraint of small dataset, but also because of their interpretability with physical meanings such as acceleration of moving body parts, which helps us more for the purpose of this thesis. After gaining a fundamental understanding through our studies, we will try to investigate the approaches by integrating data-driven approaches in the future work.

Finally, our studies can be further improved by investigating more diversified tasks. For the fight detection study, in addition to human fight events, we can also study on other fierce activities, such as arguments, etc. While for the study of predicting user interaction intention, we can also study on other common daily computer interaction tasks, such as reading, writing, playing game, and crowdsourcing, etc. Moreover, the current study of applying our intention prediction model to detect user slips is only conducted in simulated scenarios. We would try to conduct this study in multiple real tasks in the future.

6.3 Future Work

With respect to the limitations described above, there are some future works that are worth further investigation, based on the current studies.

6.3.1 Detecting Intention to Fight

Our current study is focused on detecting fight events when they are happening. In the future, we would like to extend our algorithm with the ability to detect the intention to fight, that is, the fight detection model should have the ability to predict fight events ahead before they happen. Such a preventive

application is highly useful in practice. For instance, when a system detects fight intention it will raise an alarm for warning, which may well deter the potential fight from happening, when the warning is perceived by the potentially offending parties.

To this end, these kinds of models may need to detect the angry moments, arguments or other potential aggressive behaviors. Cross-species learning could also be an interesting direction of these works, in linking the affective states and actions between human and animals.

6.3.2 Investigating on Diversified Real Tasks

In real usage, a user intention prediction model needs to be able to predict user's intention a few seconds ahead of the interaction event. Although the current model can achieve reasonable performance for predicting a few seconds ahead, it can still be further improved. In future work, we plan to experiment with more sophisticated models to improve the prediction performance, such as taking mouse-gaze movements coordination, gaze-head movements coordination, etc. into consideration. Moreover, our current study only focuses on web search task. In the future, we would try to study user intention prediction on more diversified tasks. For instance, we can study on writing tasks, where keyboard inputs dominate the interactions. We can then utilize users' typing behaviors to model user intention, which are not considered in our current studies.

In addition, in our current study, we only investigate user slips detection in some toy experiments. In the future, we plan to develop real time user slips detection models and investigate their suitability in real applications. Further experiments will also be conducted to evaluate the feasibility and usability of the models.

6.3.3 Applying Physiological Mouse to User Intention Prediction

In the future, we plan to characterize more precisely the relationship between physiological signals and human emotions by using the physiological mouse, through more diversified tasks that are designed to elicit different emotions. Besides, we will try to investigate the feasibility of correlating human physiological signals with human intentions in different tasks. In addition, more sensors to measure skin temperature and skin conductivity could be augmented onto the mouse, making it a fully functional physiological mouse capable of returning multiple useful signals in a non-intrusive manner for physiological signal computation. In this aspect, different machine learning approaches would then be studied to better recognize the user emotion and interaction intention based on the various signals.

Finally, we would like to expand the collection of human interaction and body signals in a multimodal setting, to more accurately detect human intention with the increased dimension of inputs, which include the physiological signals captured by the physiological mouse. In our user intention prediction study, we have successfully utilized mouse interaction, gaze interaction, head as well as body movements to build user intention prediction model. The physiological signals may also be one useful type of complementary signals for our ultimate user intention prediction model. More extensive experiments would need to be conducted to investigate multimodal user intention prediction approaches. Moreover, we would also like to explore building up a general model for generic users, and a learning module that will adapt to a specific user over time.

6.3.4 Integrating with Deep Learning Approaches

Current approaches described in this thesis rely on the hand-crafted features. Compared with deep learning approaches, hand-crafted features carry more

physical meanings such as the acceleration of fighting actions. That can help us understand more about human intention. However, it is also interesting to explore other features based on deep auto-encoders or features embedded within general event detection CNN [98, 107], etc. Therefore, in the future, we will increase the size of our datasets, and then try to apply and integrate deep learning approaches in our studies.

6.4 Other Relevant Contributions

In addition to the main contributions previously described, the following describes other relevant contributions arising from my thesis project.

6.4.1 Using LSTM for User Intention Prediction

We extend our study of user intention prediction by an initial attempt to explore with a deep learning approach. In this study, we adopt the long short-term memory (LSTM) [32, 38] network to build our model. LSTM is a recurrent neural network architecture, which is appropriate for the recognition of sequential pattern. Specifically, we propose a dual-stream LSTM framework to predict users' intention as depicted in Figure 6-1. Our proposed network framework utilizes the historical activity information as well as the mouse interaction information to predict user's intentions. Instead of modeling the two types of features in one LSTM network, we model them via two separate LSTM networks and adopt a weighted average approach to fuse the mouse interaction and historical activity networks.

Both networks are modeled as an LSTM with two layers and 64 neurons on each layer. We use the mini-batch stochastic gradient descent algorithm to learn the network parameters, where the batch size is set to 128 and momentum set to 0.8. The learning rate is initialized as 0.001 and decays by 10 times every 10,000

iterations. The whole training procedure stops after 30,000 iterations.

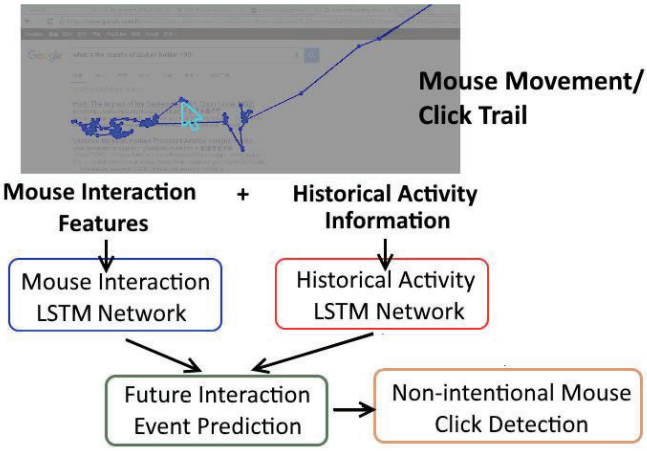


Figure 6-1 A dual-stream LSTM for user intention prediction.

To evaluate, we consider two scenarios in daily computer usage: a more structured crowdsourcing annotation task and a more free-form, open-ended web search task. Our results indicate that we could predict the next interaction event with reasonable accuracy.

Reference

Tiffany C. K. Kwok, **Eugene Yujun Fu**, Erin You Wu, Michael Xuelin Huang, Grace Ngai, and Hong Va Leong. 2018. “Ev’ry Little Movement Has a Meaning of Its Own: Using Past Mouse Movements to Predict the Next Interaction,” in *Proceedings of the 23th International Conference on Intelligent User Interfaces (IUI)*, 2018, pp. 397-401.

6.4.2 Using LSTM for Fight Detection

In this thesis, we have already attempted to encode the temporal information of local motion signals by statistical features for human fight detection, which lead to promising performance. We then extend our study by using deep learning

approach to encode temporal information. Specifically, we extract the local motion signals that describe motion amplitudes and accelerations within different regions of video frames. We then apply a long short-term memory (LSTM) network [32, 38] to learn the temporal motion representation from each video segment. In general, our method encodes both the spatial and temporal motion features, which are essential for fight detection. Figure 6-2 demonstrates the framework of our approaches.

We utilize the LSTM network to analyze the temporal information from the extracted local motion sequences. As a fight action generally presents a very unique pattern of speed and acceleration change along with time, learning an appropriate temporal motion representation is highly valuable for fight recognition. Furthermore, a video of a natural fight scene may contain multiple basic motion subsequences, which can be rather diverse and hard to delineate with human-designed features. We, therefore, use an LSTM to capture the useful temporal feature representation directly from the fight data.

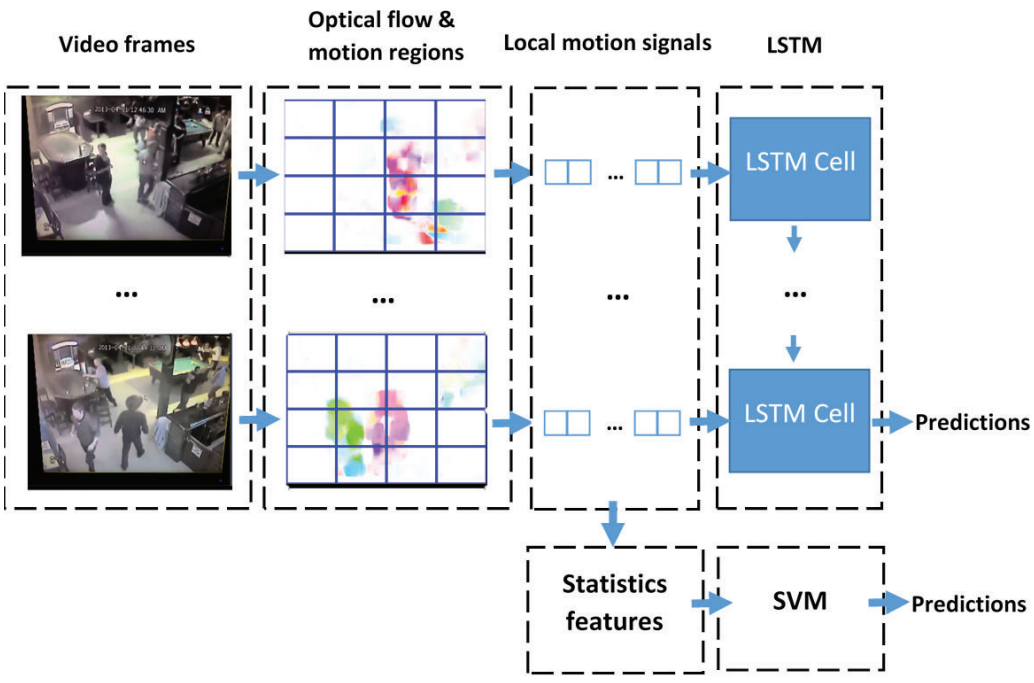


Figure 6-2 Learning fight detection model from local motion signals using SVM and LSTM.

In our method, a single-layer LSTM is deployed to accept the local motion features as input and output the binary recognition result of fight and non-fight. The input of LSTM includes the $n_c \times n_r$ motion magnitude features and the $n_c \times n_r$ acceleration features. The outputs of the LSTM first input layer are then fed into the second layer, which contains 64 nodes. Each step of the LSTM represents a time frame in the video.

To generate more data for training and to prevent severe over-fitting, data augmentation is commonly adopted to generate diverse training samples. In some related works, random cropping and horizontal flipping are employed to augment training samples. In this paper, we exploit cropping augmentation technique to augment our training samples, by adapting the methods introduced in [92, 107]. In the cropping technique, the extracted regions are focusing on not only the center area of an image but also the corner area of an image. Besides, the width and height of cropped region are randomly selected to avoid only selecting regions of fixed size. Finally, the selected cropped regions will be resized to the original size of the frame for feature extraction and network training.

In our experiment, we investigate the proposed *LMF* with LSTM model for fight detection on human real fight dataset introduced in Section 3. We evaluate this approach and compare with state-of-the-art methods in fight or violence detection. The experiment results show that this approach can outperform the state-of-the-art approaches including the approach of combining our *LMF* with SVM. This also indicates that *LMF* is a good representation for fight actions. More importantly, it can be well generalized across different classifiers. The finding of this study is being polished for publication.

Reference

Eugene Yujun Fu, Hong Va Leong, Grace Ngai. 2019. "Automatic Fight Detection from Local Motion Signals with LSTM". **– In Preparation**

6.4.3 Modeling Mouse and Gaze Interaction for Stress Detection

Detecting mental stress has gained attention in recent studies, as mental stress can lead to anxiety, depression, and various mental illnesses. However, most of the prior techniques applied for stress detection are constrained by the reliance on the interface layout or other information obtained from the graphical user interface (GUI), making these kinds of approaches hard to generalize across different interfaces and be applied to detect human mental stress in real applications.

To address this challenge, we proposed a GUI-agnostic stress detection approach in this study, which can utilize the information about the correlation between mouse and eye gaze movements to build stress detection model, without considering the actual information of GUI. We evaluated the approach in two different computer interaction tasks with two different kinds of GUI. The experiment results suggest that our approach can effectively detect users' mental stress in both of the tasks, in a GUI-agnostic manner.

Reference

Jun Wang, **Eugene Yujun Fu,** Grace Ngai, Hong Va Leong. 2019. "Detecting Stress from Mouse-Gaze Attraction". To appear in *Proceedings of ACM/SIGAPP Symposium on Applied Computing*, 2019.

References

[1] Ahsan, G.M.T., Gani, M.O., Hasan, M.K., Ahamed, S.I., Chu, W., Adibuzzaman, M. and Field, J. 2017. A Novel Real-Time Non-invasive Hemoglobin Level Detection Using Video Images from Smartphone Camera. *Proceedings of the 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. (2017), 967–972.

[2] Alexander, J., Cockburn, A., Fitchett, S., Gutwin, C. and Greenberg, S. 2009. Revisiting read wear: analysis, design, and evaluation of a footprints scrollbar. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), 1665–1674.

[3] Allen, J. 2007. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*. 28, 3 (2007), R1–39.

[4] Arapakis, I. and Valkanas, G. 2014. Understanding Within-Content Engagement through Pattern Analysis of Mouse Gestures. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (2014), 1439–1448.

[5] Asano, T., Sharlin, E., Kitamura, Y., Takashima, K. and Kishino, F. 2005. Predictive interaction using the delphian desktop. *Proceedings of the 18th annual ACM symposium on User interface software and technology* (2005), 133–141.

[6] Balakrishnan, G., Durand, F. and Guttag, J. 2013. Detecting pulse from head motions in video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2013), 3430–3437.

[7] Baltrušaitis, T., Robinson, P. and Morency, L.-P. 2016. Openface: an open

source facial behavior analysis toolkit. *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on* (2016), 1–10.

[8] Banovic, N., Grossman, T. and Fitzmaurice, G. 2013. The effect of time-based cost of error in target-directed pointing tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), 1373–1382.

[9] Blakemore, S.-J. and Decety, J. 2001. From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*. 2, 8 (2001), 561–567.

[10] Blunsden, S. and Fisher, R.B. 2010. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*. 4, 1–12 (2010), 4.

[11] Brown, E.T., Ottley, A., Zhao, H., Lin, Q., Souvenir, R., Endert, A. and Chang, R. 2014. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*. 20, 12 (2014), 1663–1672.

[12] Brown, T.E., Beightol, L.A., Koh, J. and Eckberg, D.L. 1993. Important influence of respiration on human RR interval power spectra is largely ignored. *Journal of Applied Physiology*. 75, 5 (1993), 2310–2317.

[13] Burgos-Artizzu, X.P., Dollár, P., Lin, D., Anderson, D.J. and Perona, P. 2012. Social behavior recognition in continuous video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012), 1322–1329.

[14] Calvo, R.A. and D'Mello, S. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*. 1, 1 (2010), 18–37.

[15] Chang, C.J., Amick, B.C., Menendez, C.C., Katz, J.N., Johnson, P.W., Robertson, M. and Dennerlein, J.T. 2007. Daily computer usage correlated with undergraduate students' musculoskeletal symptoms. *American journal of industrial medicine*. 50, 6 (2007), 481–488.

[16] Chen, J., Liu, X., Tu, P. and Aragones, A. 2013. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*. 34, 15 (2013), 1964–1970.

[17] Chen, M.Y. and Hauptmann, A. 2009. *Mosift: Recognizing human actions in surveillance videos*.

[18] Cheng, W.H., Chu, W.T. and Wu, J.L. 2003. Semantic context detection based on hierarchical audio models. *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval* (2003), 109–115.

[19] Chuda, D., Kratky, P. and Tvarozek, J. 2015. Mouse Clicks Can Recognize Web Page Visitors! *Proceedings of the 24th International Conference on World Wide Web* (2015), 21–22.

[20] Clair, A.S., Mead, R., Matarić, M.J. and others 2010. Monitoring and guiding user attention and intention in human-robot interaction. *ICRA-ICAIR Workshop, Anchorage, AK, USA* (2010), 1025.

[21] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C. 2004. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV* (2004), 1–2.

[22] Dai, W., Yang, Q., Xue, G.-R. and Yu, Y. 2007. Boosting for transfer learning. *Proceedings of the 24th International Conference on Machine Learning* (2007), 193–200.

[23] Debnath, P.P., Rashidul Hasan, A.F.M. and Das, D. 2017. Detection and

controlling of drivers' visual focus of attention. *ECCE 2017 - International Conference on Electrical, Computer and Communication Engineering*. (2017), 301–307.

[24] Duan, L., Xu, D. and Tsang, I.W.-H. 2012. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*. 23, 3 (2012), 504–518.

[25] Elkahky, A.M., Song, Y. and He, X. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. *Proceedings of the 24th International Conference on World Wide Web* (2015), 278–288.

[26] Emotiv 2014. EEG System / Electroencephalography.

[27] Evans, A. and Wobbrock, J. 2012. Taming wild behavior: the input observer for obtaining text entry and mouse pointing measures from everyday computer use. *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), 1947–1956.

[28] Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. *Image Analysis*. (2003), 363–370.

[29] Fisher, R.B. 2004. The PETS04 surveillance ground-truth data sets. *Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance* (2004), 1–5.

[30] Fitchett, S. and Cockburn, A. 2012. Accessrank: predicting what users will do next. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), 2239–2242.

[31] Gao, Y., Liu, H., Sun, X., Wang, C. and Liu, Y. 2016. Violence detection using oriented violent flows. *Image and vision computing*. 48, (2016), 37–41.

[32] Gers, F.A., Schmidhuber, J. and Cummins, F. 1999. Learning to forget: Continual prediction with LSTM. (1999), 850–855.

[33] Giannakopoulos, T., Kosmopoulos, D., Aristidou, A. and Theodoridis, S. 2006. Violence content classification using audio features. *SETN* (2006), 502–507.

[34] Google 2012. A Google A Day.

[35] Guo, Q. and Agichtein, E. 2010. Ready to buy or just browsing?: detecting web searcher goals from interaction data. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), 130–137.

[36] Guo, Q., Jin, H., Lagun, D., Yuan, S. and Agichtein, E. 2013. Mining touch interaction data on mobile devices to predict web search result relevance. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*. (2013), 153.

[37] Hassner, T., Itcher, Y. and Kliper-Gross, O. 2012. Violent flows: Real-time detection of violent crowd behavior. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2012), 1–6.

[38] Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*. 9, 8 (1997), 1735–1780.

[39] Hongeng, S., Brémond, F. and Nevatia, R. 2000. Representation and optimal recognition of human activities. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2000), 818–825.

[40] Hu, B., Zhang, Y., Chen, W., Wang, G. and Yang, Q. 2011. Characterizing search intent diversity into click models. *Proceedings of the 20th international conference on World wide web* (2011), 17–26.

[41] Huang, J. and White, R. 2012. User See, User Point: Gaze and Cursor Alignment in Web Search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (2012), 1341–1350.

[42] Huang, M.X., Kwok, T.C.K., Ngai, G., Leong, H.V. and Chan, S.C.F. 2014. Building a self-learning eye gaze model from user interaction data. *Proceedings of the 22nd ACM international conference on Multimedia* (2014), 1017–1020.

[43] Huang, M.X., Li, J., Ngai, G. and Leong, H.V. 2016. StressClick. *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*. (2016), 1395–1404.

[44] Huang, M.X., Li, J., Ngai, G., Leong, H.V. and Hua, K.A. 2017. Fast-PADMA: Rapidly Adapting Facial Affect Model from Similar Individuals. *IEEE Transactions on Multimedia*. (2017).

[45] iHealth 2014. iHealth Pulse Oximeter.

[46] Iwashita, Y., Takamine, A., Kurazume, R. and Ryoo, M.S. 2014. First-person animal activity recognition from egocentric videos. *2014 22nd International Conference on Pattern Recognition (ICPR)* (2014), 4310–4315.

[47] Jatupaiboon, N., Pan-Ngum, S., Israsena, P., Chen, B.-W., Hsieh, S. and Wu, C.-H. 2013. Real-Time EEG-Based Happiness Detection System. *The Scientific World Journal*. 2013, (2013).

[48] Jayagopi, D.B., Hung, H., Yeo, C. and Gatica-Perez, D. 2009. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*. 17, 3 (2009), 501–513.

[49] Jiang, J. and Allan, J. 2016. Reducing click and skip errors in search result

ranking. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (2016), 183–192.

[50] Kato, Y., Kanda, T. and Ishiguro, H. 2015. May I help you?: Design of Human-like Polite Approaching Behavior. *HRI '15 Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (2015), 35–42.

[51] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*. 13, 3 (2001), 637–649.

[52] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*. 13, 3 (2001), 637–649.

[53] Kim, B.S. and Yoo, S.K. 2006. Motion artifact reduction in photoplethysmography using independent component analysis. *IEEE transactions on biomedical engineering*. 53, 3 (2006), 566–568.

[54] Kim, J. and André, E. 2008. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*. 30, 12 (2008), 2067–2083.

[55] Kim, S., Valente, F., Filippone, M. and Vinciarelli, A. 2014. Predicting Continuous Conflict Perception with Bayesian Gaussian Processes. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*. 5, 2 (2014), 187–200.

[56] Kusk, K., Nielsen, D.B., Thylstrup, T., Rasmussen, N.H., Jørvang, J., Pedersen, C.F. and Wagner, S. 2013. Feasibility of using a lightweight context-aware system for facilitating reliable home blood pressure self-measurements. *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare* (2013), 236–239.

[57] Ladha, C., Hammerla, N., Hughes, E., Olivier, P. and Ploetz, T. 2013. Dog's life: wearable activity recognition for dogs. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2013), 415–418.

[58] Lagun, D. and Agichtein, E. 2015. Inferring searcher attention by jointly modeling user interactions and content salience. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015), 483–492.

[59] Lallé, S., Toker, D., Conati, C. and Carenini, G. 2015. Prediction of users' learning curves for adaptation while using an information visualization. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (2015), 357–368.

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

[61] Laufer, L. and Németh, B. 2008. Predicting user action from skin conductance. *Proceedings of the 13th international conference on Intelligent user interfaces* (2008), 357–360.

[62] Lee, B. and Oulasvirta, A. 2016. Modelling error rates in temporal pointing. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 1857–1868.

[63] Lee, B., Savisaari, O. and Oulasvirta, A. 2016. Spotlights: Attention-Optimized Highlights for Skim Reading. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 5203–5214.

[64] Li, J., Ngai, G., Va Leong, H. and Chan, S. 2016. Multimodal Human Attention Detection for Reading. *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16* (2016), 187–192.

[65] Lin, C.-J., Wu, C. and Chaovalitwongse, W.A. 2015. Integrating human behavior modeling and data mining techniques to predict human errors in numerical typing. *IEEE Transactions on Human-Machine Systems*. 45, 1 (2015), 39–50.

[66] List, T., Bins, J., Vazquez, J. and Fisher, R.B. 2005. Performance evaluating the evaluator. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on* (2005), 129–136.

[67] Lu, Y., Mahmoud, M. and Robinson, P. 2017. Estimating Sheep Pain Level Using Facial Action Unit Detection. *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on* (2017), 394–399.

[68] Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y. and Kameyama, K. 2015. Measuring Photoplethysmogram-Based Stress-Induced Vascular Response Index to Assess Cognitive Load and Stress. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. April (2015), 857–866.

[69] Mandayam Comar, P. and Sengamedu, S.H. 2017. Intent Based Relevance Estimation from Click Logs. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), 59–66.

[70] Maniu, S., O'Hare, N., Aiello, L.M., Chiarandini, L. and Jaimes, A. 2013. Search behaviour on photo sharing platforms. *Multimedia and Expo (ICME), 2013 IEEE International Conference on* (2013), 1–6.

[71] Marcos-Ramiro, A., Pizarro-Perez, D., Marron-Romera, M., Nguyen, L. and Gatica-Perez, D. 2013. Body communicative cue extraction for conversational analysis. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. (2013).

1–8.

[72] Masciocchi, C.M. and Still, J.D. 2013. Alternatives to eye tracking for predicting stimulus-driven attentional selection within interfaces. *Human--Computer Interaction*. 28, 5 (2013), 417–441.

[73] Mazur-Milecka, M. and Rumiński, J. 2017. Automatic analysis of the aggressive behavior of laboratory animals using thermal video processing. *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE* (2017), 3827–3830.

[74] Mikkelsen, S., Vilstrup, I., Lassen, C.F., Kryger, A.I., Thomsen, J.F. and Andersen, J.H. 2007. Validity of Questionnaire Self-Reports on Computer, Mouse and Keyboard Usage during a Four-Week Period. *Occupational and environmental medicine*. 64, 8 (2007), 541–547.

[75] Mok, R.K.P., Chang, R.K.C. and Li, W. 2017. Detecting Low-Quality Workers in QoE Crowdtesting: A Worker Behavior-Based Approach. *IEEE Transactions on Multimedia*. 19, 3 (2017), 530–543.

[76] Mollaret, C., Mekonnen, A.A., Ferrane, I., Pinquier, J. and Lerasle, F. 2015. Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (2015), 1–6.

[77] Navarathna, R., Lucey, P., Carr, P., Carter, E., Sridharan, S. and Matthews, I. 2014. Predicting movie ratings from audience behaviors. *IEEE Winter Conference on Applications of Computer Vision (WACV)* (2014), 1058–1065.

[78] Negulescu, M. and McGrenere, J. 2015. Grip change as an information side channel for mobile touch interaction. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), 1519–1522.

[79] NeuroSky 2014. NeuroSky / MindSet.

[80] Nguyen, L.S., Marcos-Ramiro, A., Marrón Romera, M. and Gatica-Perez, D. 2013. Multimodal analysis of body communication cues in employment interviews. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (2013), 437–444.

[81] Nievas, E.B., Suarez, O.D., Garcia, G.B. and Sukthankar, R. 2011. Violence detection in video using computer vision techniques. *International Conference on Computer Analysis of Images and Patterns* (2011), 332–339.

[82] Pan, S.J., Tsang, I.W., Kwok, J.T. and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*. 22, 2 (2011), 199–210.

[83] Pan, S.J. and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*. 22, 10 (2010), 1345–1359.

[84] Pantic, M. and Rothkrantz, L.J.M. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*. 91, 9 (2003), 1370–1390.

[85] Pasqual, P.T. and Wobbrock, J.O. 2014. Mouse pointing endpoint prediction using kinematic template matching. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 743–752.

[86] Perronnin, F., Sánchez, J. and Mensink, T. 2010. Improving the fisher kernel for large-scale image classification. *European Conference on Computer Vision* (2010), 143–156.

[87] Rainville, P., Bechara, A., Naqvi, N. and Damasio, A.R. 2006. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International journal of psychophysiology*. 61, 1 (2006), 5–18.

[88] Ramseyer, F. 2013. Synchronized movement in social interaction. *Proceedings of the 2013 Inputs-Outputs Conference: An Interdisciplinary Conference on Engagement in HCI and Performance* (2013), 2.

[89] Sangineto, E., Zen, G., Ricci, E. and Sebe, N. 2014. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. *Proceedings of the 22nd ACM International Conference on Multimedia* (2014), 357–366.

[90] Schwarz, J., Marais, C.C., Leyvand, T., Hudson, S.E. and Mankoff, J. 2014. Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 3443–3452.

[91] Scully, C.G., Lee, J., Meyer, J., Gorbach, A.M., Granquist-Fraser, D., Mendelson, Y. and Chon, K.H. 2012. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Transactions on Biomedical Engineering*. 59, 2 (2012), 303–306.

[92] Simonyan, K. and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. (2014).

[93] Slanzi, G., Balazs, J. a. and Velásquez, J.D. 2017. Combining Eye Tracking, Pupil Dilation and EEG Analysis for Predicting Web Users Click Intention. *Information Fusion*. 35, (2017), 51–57.

[94] Soleymani, M., Lichtenauer, J., Pun, T. and Pantic, M. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*. 3, 1 (2012), 42–55.

[95] Soomro, K., Zamir, A.R. and Shah, M. 2012. {UCF}101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*. (2012).

[96] Sun, Y., Yuan, N.J., Xie, X., McDonald, K. and Zhang, R. 2016. Collaborative Nowcasting for Contextual Recommendation. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. (2016), 1407–1418.

[97] Toker, D., Steichen, B., Gingerich, M., Conati, C. and Carenini, G. 2014. Towards facilitating user skill acquisition: identifying untrained visualization users through eye tracking. *Proceedings of the 19th International Conference on Intelligent User Interfaces* (2014), 105–114.

[98] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015), 4489–4497.

[99] Trung, P., Giuliani, M., Miksch, M., Stollnberger, G., Stadler, S., Mirnig, N. and Tscheligi, M. 2017. Head and shoulders: automatic error detection in human-robot interaction. *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (2017), 181–188.

[100] Vandewynckel, J., Otis, M., Bouchard, B., Bouzouane, A. and others 2013. Towards a real-time error detection within a smart home by using activity recognition with a shoe-mounted accelerometer. *Procedia Computer Science*. 19, (2013), 516–523.

[101] Vinciarelli, A., Dielmann, A., Favre, S. and Salamin, H. 2009. Canal9: A database of political debates for analysis of social interactions. *International Conference on Affective Computing and Intelligent Interaction and Workshops* (2009), 1–4.

[102] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F. and Schroeder, M. 2012. Bridging the gap between social animal and

unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*. 3, 1 (2012), 69–87.

[103] Wacharamanotham, C. 2014. Making bare hand input more accurate. *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (2014), 307–310.

[104] Waluyo, A.B., Yeoh, W.-S., Pek, I., Yong, Y. and Chen, X. 2010. Mobisense: Mobile body sensor network for ambulatory monitoring. *ACM Transactions on Embedded Computing Systems (TECS)*. 10, 1 (2010), 13.

[105] Wang, H. and Wanga, L. 2017. Cross-Agent Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. (2017).

[106] Wang, J., Huang, M.X., Ngai, G. and Leong, H.V. 2017. Are You Stressed ? Your Eyes and the Mouse Can Tell. *International Conference on Affective Computing and Intelligent Interaction* (2017), 222–228.

[107] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. *European Conference on Computer Vision* (2016), 20–36.

[108] Welch, P. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*. 15, 2 (1967), 70–73.

[109] Wickens, C.D., Hollands, J.G., Banbury, S. and Parasuraman, R. 2015. *Engineering psychology & human performance*. Psychology Press.

[110] Wobbrock, J.O., Cutrell, E., Harada, S. and MacKenzie, I.S. 2008. An error model for pointing based on Fitts' law. *Proceedings of the SIGCHI conference on human factors in computing systems* (2008), 1613–1622.

[111] Xiao, T., Li, H., Ouyang, W. and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1249–1258.

[112] Xu, P., Sugano, Y. and Bulling, A. 2016. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 3299–3310.

[113] Yan, K., Kou, L. and Zhang, D. 2016. Domain Adaptation via Maximum Independence of Domain Features. *arXiv preprint arXiv:1603.04535*. (2016).

[114] Yang, Z. and Rothkrantz, L.J.M. 2010. Automatic aggression detection inside trains. *2010 IEEE International Conference on Systems Man and Cybernetics (SMC)* (2010), 2364–2372.

[115] Yao, Y. and Doretto, G. 2010. Boosting for transfer learning with multiple sources. *2010 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (2010), 1855–1862.

[116] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. 2014. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems* (2014), 3320–3328.

[117] Zelnik-Manor, L. and Irani, M. 2001. Event-based analysis of video. *Computer Vision and Pattern Recognition, 2001. Proceedings of the 2001 IEEE Computer Society Conference on CVPR* (2001), II--II.

[118] Zen, G., Porzi, L., Sangineto, E., Ricci, E. and Sebe, N. 2016. Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia*. 18, 4 (2016), 775–788.

[119] Zeng, Z.H., Pantic, M., Roisman, G.I. and Huang, T.S. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. 31, 1 (2009), 39-58.

[120] Zhang, X., Sugano, Y., Fritz, M. and Bulling, A. 2017. It's written all over your face: Full-face appearance-based gaze estimation. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), 51-60.

[121] Zheng, N., Paloski, A. and Wang, H. 2011. An efficient user verification system via mouse movements. *Proceedings of the 18th ACM conference on Computer and communications security* (2011), 139–150.

[122] Zhou, P., Ding, Q., Luo, H. and Hou, X. 2017. Violent Interaction Detection in Video Based on Deep Learning. *Journal of Physics: Conference Series* (2017), 12044.

[123] Zhou, X., Cao, X. and Ren, X. 2009. Speed-accuracy tradeoff in trajectory-based tasks with temporal constraint. *IFIP Conference on Human-Computer Interaction* (2009), 906–919.

[124] Ziebart, B., Dey, A. and Bagnell, J.A. 2012. Probabilistic pointing target prediction via inverse optimal control. *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (2012), 1–10.