

A game-based approach towards human augmented image annotation.

Seneviratne, Attgalage Lasantha Gunathilaka

For additional information about this publication click this link.

<http://qmro.qmul.ac.uk/jspui/handle/123456789/2445>

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk

**QUEEN MARY, UNIVERSITY OF LONDON
SCHOOL OF ELECTRONIC ENGINEERING
AND COMPUTER SCIENCE**

**A GAME-BASED APPROACH
TOWARDS HUMAN AUGMENTED
IMAGE ANNOTATION**

Thesis submitted to University of London in partial
fulfilment of the requirement for the degree of
Doctor of Philosophy

**ATTYGALAGE LASANTHA GUNATHILAKA
SENEVIRATNE**

London 2011

I dedicate this work to my beloved family and uncle.

Acknowledgements

Firstly I would like to thank my supervisor, Prof. Ebroul Izquierdo, for his excellent supervision, his knowledge, his belief and interest in the work and encouragement, guidance and motivation throughout. Also, I will be forever thankful to him for providing an opportunity to pursue a PhD in Multimedia and Vision Group. I would like to extend my thanks to all the staff members of the School of Electronic Engineering and Computer Science for their support in particular to Dr. Yannis Patras, Dr. John Bigham, Mr. Kok Ho Huen and Mrs. Melissa Yeo.

I thank all the members of MMV group for making my stay at Queen Mary's and in London a very memorable experience. I am grateful for having the pleasure of closely working with Dr. Krishna Chandramouli, Dr. Naeem Ramsan, Dr. Qianni Zhang, Dr. Ivan Damnjanovic and Dr. Tomas Piatrik. Also I am thankful to Navid, Akram, Virginia and Karthike for their constant support throughout my stay.

A special thanks to all my friends, Luka, Antony, Berni, Aris, Kapila, Eric, Romesh, Dharshana and Niroshan for their friendship and support during my stay in London.

Finally, I thank all my family and relatives for their ongoing love and support. I am very thankful to my parents, sisters, wife and my loving uncle for their unlimited love and guidance, for always believing in me and encouraging me throughout. Words cannot express how much I love you all.

Abstract of Thesis presented by Attygalage Lasantha Gunathilaka Seneviratne to
University of London.

A GAME-BASED APPROACH TOWARDS HUMAN AUGMENTED IMAGE ANNOTATION

ABSTRACT: Image annotation is a difficult task to achieve in an automated way. In this thesis, a human-augmented approach to tackle this problem is discussed and suitable strategies are derived to solve it. The proposed technique is inspired by human-based computation in what is called “human-augmented” processing to overcome limitations of fully automated technology for closing the semantic gap. The approach aims to exploit what millions of individual gamers are keen to do, i.e. enjoy computer games, while annotating media.

In this thesis, the image annotation problem is tackled by a game based framework. This approach combines image processing and a game theoretic model to gather media annotations. Although the proposed model behaves similar to a single player game model, the underlying approach has been designed based on a two-player model which exploits the player’s contribution to the game and previously recorded players to improve annotations accuracy. In addition, the proposed framework is designed to predict the player’s intention through Markovian and Sequential Sampling inferences in order to detect cheating and improve annotation performances. Finally, the proposed techniques are comprehensively evaluated with three different image datasets and selected representative results are reported.

List of Abbreviations

Abbreviations

AI	- Artificial Intelligent
ANOVA	- Analysis of Variance
AQL	- Acceptance Quality Level
CBIR	- Content Based Image Retrieval
CLD	- Colour Layout Descriptor
CRR	- Correct Rejection Rates
CSS	- Curvature Scale Space
CTW	- Content Tree Weighting
DCD	- Discrete Cosine Transform
DCT	- Discrete Cosine Transform
DTMC	- Discrete Time Markov Chain
EHD	- Edge Histogram Descriptor
GLCM	- Grey-Level co-occurrence Matrix
GT	- Game Theory
GWAP	- Games with a Purpose
HMM	- Hidden Marker Model
HTD	- Homogenous Texture Descriptor
KKB	- KissKissBan
MM	- Markov Models
MPEG	- Moving Picture Experts Group
MSM	- Manhattan Story

Abbreviations (cont.)

NE	- Nash Equilibrium
NN	- Neural Networks
OC	- Operating Characteristic Curve
PBD	- Perceptual Browsing Descriptor
PDF	- Probability Density Function
PPM	- Prediction by Partial Match
RBF	- Radial Basis Function
RQL	- Rejectable Quality Level
SD	- Standard Deviation
SPRT	- Sequential Probability Ratio Test
SS	- Sequential Sampling
SVM	- Support Vector Machines
TPM	- Transition Probability Matrix
VC	- Vapnik- Chervnonenkis
WWW	- World Wide Web

Contents

Chapter 1.....	1
INTRODUCTION.....	1
1.1 Research Objectives.....	5
1.2 Contribution of the Thesis	6
1.3 Structure of the Thesis	7
Chapter 2.....	10
GAMES WITH A PURPOSE: A SURVEY OF RELATED WORK.....	10
2.1 Introduction.....	10
2.2 Towards human augmented image annotation	12
2.3 Existing game-base frameworks.....	14
2.3.1 Output agreement games	14
2.3.2 Inversion-Problem games	15
2.3.3 Input-agreement games.....	16
2.4 GWAP - Popular games and their strategies	17
2.5 Summary.....	22
Chapter 3.....	24
PROPOSED FRAMEWORK FOR IMAGE ANNOTATION	24
3.1 Introduction.....	25
3.2 System Overview.....	26
3.2.1 Payoff Calculation and Decision making Unit.....	29
3.2.2 Player Outcome Prediction Unit.....	30
3.3 Implementation of the framework	31
3.3.1 Graphical User Interface.....	32
3.4 Experimental Setups	34
3.5 Summary.....	35
Chapter 4.....	37

PAYOFF CALCULATION AND DECISION MAKING	37
4.1 Introduction.....	37
4.2 Introduction to Game Theory and Decision making	38
4.2.1 Interactive decision problems and static games	39
4.2.2 Cooperative and Non-cooperative games.....	40
4.2.3 Nash Equilibrium based decision making	43
4.2.4 The Problem of multiple Equilibria.....	44
4.3 Applying the Nash Equilibrium based Decision Making in Image Annotation	45
4.3.1 Nash Equilibrium representation	50
4.4 Weighting player 2's Payoff by Image Classification.....	55
4.4.1 Analysis of Low-level Features	57
4.4.2 Fusing MPEG-7 Visual Descriptors for Image Classification	65
4.5 Summary.....	66
Chapter 5.....	67
PLAYER'S OUTCOME PREDICTION AND DECISION MAKING.....	67
5.1 Introduction.....	67
5.2 Prediction by Markov Chains	69
5.2.1 Hidden Markov Models.....	72
5.3 Decision making by Sequential Sampling.....	73
5.3.1 Methods used in product sampling	74
5.3.2 Sequential Probability Ratio Test	77
5.3.3 OC curve and probability distribution	80
5.4 Applying the Markov based and Sequential Sampling based Prediction Techniques.....	83
5.4.1 Player's outcome prediction by Markovian based inference.....	83
5.4.2 Player's outcome prediction by Sequential Sampling	84
5.5 Summary.....	88
Chapter 6.....	90
EXPERIMENTAL EVALUATION	90
6.1 Introduction.....	90
6.2 Experimental Evaluation	93
6.2.1 Performance measure in image classification.....	94

6.2.2 Measure of Usability	97
6.2.3 Measure of Efficiency	104
6.2.4 Measure of Precision	106
6.3 Result Obtained for Different Configurations of the System	115
6.3.1 Two-player game model with no prediction mechanisms installed .	117
6.3.2 Two-player game model followed by the Markovian prediction	121
6.3.3 Two-player game model followed by the proposed sampling prediction mechanism.....	126
6.4 Summary.....	128
Chapter 7.....	129
CONCLUSION AND FUTURE WORK.....	129
List of Author's Publications	147
Appendix A	148
Support Vector Machines (SVM)	148
A.1 Introduction.....	148
Appendix B.....	155
Image Databases	155
B.1 ESP Image Dataset	155
B.2 Caltech 101 Image Dataset	155
B.3 Corel Image Dataset	156
Appendix C	161
Questionnaire on Usability Test.....	161
C.1 Template of the Usability Test	161
Appendix D	163
Outcomes of Analysis of Variance (ANOVA)	163
D.1 Excitement	163
D.2 Addiction	164
D.3 Enjoyability.....	165
D.4 Game difficulty level	166

List of Figures

Figure 2.1: Output agreement game mechanism.	14
Figure 2.2: Inversion problem game mechanism.	15
Figure 2.3: Input agreement game mechanism.	16
Figure 2.4: Overview of “GWAP”.....	17
Figure 2.5: Visual representation of current GWAP approaches.	22
Figure 3.1: A complete block diagram of the framework.	27
Figure 3.2: Some screenshots of INT-1.....	33
Figure 3.3: Some screenshots of INT-2.....	33
Figure 4.1: Segmenting player's outcome into set of tags.....	46
Figure 4.2: Player's probability distribution in gaming.	46
Figure 4.3: Payoff outcome representation.	54
Figure 5.1: A diagrammatic overview of a Markov Model.	72
Figure 5.2: A basic structure of a HMM.	72
Figure 5.3: An example of Sequential Sampling plan.....	77
Figure 5.4: Operating characteristic curve.	81
Figure 5.5: Player's probability distribution in gaming.	83
Figure 5.6: Operating characteristic curve.	86
Figure 5.7: Proposed Sequential Sampling plan.	87
Figure 6.1: Age distribution of the players.....	99

Figure 6.2: Excitement level of games.	100
Figure 6.3: Addiction of games.	101
Figure 6.4: Enjoyability.	102
Figure 6.5: Difficulty in game play.	103
Figure 6.6: Average labels per minute.	105
Figure 6.7: Average precision for ESP dataset.	108
Figure 6.8: Average precision rates for Caltech dataset.	109
Figure 6.9: Average precision rates for Corel dataset.	110
Figure 6.10: Payoff outcome for IA-GTMM	114
Figure 6.11: Payoff outcome for IA-GTSS	115
Figure 6.12: Performance measure for classical players, example - 1.	117
Figure 6.13: Performance measure for classical players, example - 2.	118
Figure 6.14: Performance measure for random players, example – 1.	119
Figure 6.15: Performance measure for random players, example – 2.	119
Figure 6.16: Performance measure for true players, example -1.	120
Figure 6.17: Performance measure for true players, example -2.	121
Figure 6.18: Performance measure for classical players, example -1.	122
Figure 6.19: Performance measure for classical players, example -2.	122
Figure 6.20: Performances measure for random players, example -1.	123
Figure 6.21: Performances measure for random players, example -2.	124
Figure 6.22: Performances measure for genuine players, example - 1.	125
Figure 6.23: Performances measure for genuine players, example - 2.	125
Figure 6.24: Outcome measure for classical players (Prediction by sampling).	126
Figure 6.25: Outcome measure for random players (Prediction by sampling).	127

Figure 6.26: Outcome measure for genuine players (Prediction by sampling).....	127
Figure A.1: Binary classification problem.....	149
Figure B.1: Representative images for different categories taken from the ESP Image dataset (part 1).....	156
Figure B.2: Representative images for different categories taken from the ESP Image dataset (part 2).....	157
Figure B.3: Representative images for different categories taken from the Caltech Image dataset (part 1).....	158
Figure B.4: Representative images for different categories taken from the Caltech Image dataset (part 2).....	159
Figure B.5: Representative images for different categories taken from the Corel image dataset.....	160

List of Tables

Table 4.1: Truth table for all possible actions	51
Table 4.2: Payoff representation for all actions.	51
Table 6.1: Performances of the SVM classifier (Precision).....	95
Table 6.2: Performances of the SVM classifier (CRR).....	95
Table 6.3: Precision when trained with 500 images.....	95
Table 6.4: CRR when trained with 500 images.....	96
Table 6.5: Performances of the proposed frameworks for ESP Dataset (part 1). ..	112
Table 6.6: Performances of the proposed frameworks for ESP Dataset (part 2). ..	113
Table A.1: Commonly used kernel functions.....	154
Table D.1: ANOVA results related to excitement levels across different age categories	164
Table D.2: ANOVA results related to excitement levels across the four games tested	164
Table D.3: ANOVA results related to addiction levels across different age categories.....	164
Table D.4: ANOVA results related to addiction outcomes across all four games	165
Table D.5: ANOVA results related to perceived enjoyment reported by different age categories.....	165
Table D.6: ANOVA results related to perceived enjoyment reported for the four games tested.....	166

Table D.7: ANOVA results related to game difficulty level across different age categories.....	166
Table D.8: ANOVA results related to game difficulty level across all four games.....	166

Chapter 1

INTRODUCTION

It is said that “a picture is worth a thousand words”. This refers to the idea that complex scenarios can be represented by just a single image. Human beings are all capable of obtaining a majority of information in the real world by visual sense and this includes entities that can be visualized, such as images and videos. Recent developments in social networks and an increasing number of portable electronic devices, such as cameras and camera embedded mobile phones, have contributed to the already large quantity of digital multimedia content on the World Wide Web (WWW). As a consequence, the following question arises, do people label the content? If so, how often do they do so? With the increase of digital media, problems of automated classification, annotation, indexing, retrieving, and aggregating become critical for the provision of useful and user friendly multimedia systems. Reacting to these and other similar questions, researchers around the world have designed a considerable number of algorithms and frameworks with the capabilities of automated image annotation.

1. INTRODUCTION

Image annotation can be divided into two broad classes: automated annotation and manual annotation. The traditional automated framework uses multi-class image classification techniques with a large number of classes, as large as the vocabulary size. These techniques extract feature vectors from images and use machine learning techniques to assign annotation words automatically to new images. The advantages of automatic image annotation versus content-based image retrieval (CBIR) [1] are that queries can be more naturally specified by the user. CBIR generally requires human attention. Here, users have to search by image concepts such as colour and texture or finding example queries. One of the main problems in CBIR is that certain image features in example images may override the concept that the user is really focusing on. The traditional methods of image retrieval such as Flickr [2] have relied on manually annotated images, which is expensive and time-consuming, especially given the large and constantly-growing image databases in existence. Although the literature is full of automated tagging techniques, it is still not truly perfect. As a consequence, there is a huge gap between the outcomes of automated tagging and manual tagging and this is because of the existence of the semantic gap.

Over the last decade, a number of research directions have been explored addressing the semantic gap problem. One such approach is crowdsourcing (or manual annotation), which has been successfully used for harvesting multimedia annotations. For instance, very promising results have been reported for the well-known ESP game [3]. It has been shown that this particular game can be modified to annotate different types of multimedia materials or features [4] [5]. As a consequence, games like this are called “Games with a Purpose” (GWAP). Since the ESP game was introduced, a number of similar approaches to address the semantic gap issue have been proposed. Including the ESP game, most of the other approaches use humans in image tagging. Among them, the ASAA (Application for Semi-Automatic Annotation) [6] and “Manhattan Story Mash-up” [7] are two

1. INTRODUCTION

different game strategies introduced in the literature. These strategies have extended the crowdsourcing paradigm into another era by introducing two different methods of harvesting human brainpower. The most primitive approach in engaging human attention is designing interactive frameworks with multiplayer game strategies. It has been shown to be fun and entertaining. As a result, public attention is drawn into playing the game and its real purpose, image annotation, goes largely unnoticed. However, multiplayer game strategies present their own challenges in practice. For example, the ESP game is used to annotate images using two similar key words given by two unseen players. This approach is highly effective if players do not cheat by entering unrelated keywords such as “cat” for every image [8], leading the system to generate information that is not useful. Most of the games introduced in the literature use at least two-players interacting remotely through the Internet to prevent cheating and control a potential flow of misleading annotations into the metadata base. This important requirement make these games unsuitable for applications where only single isolated players are available, e.g. for gadgets with no Internet connectivity. Moreover, this phenomenon is encouraged by the survey conducted by the Mobile Marketing Watch [9], which showed that as of 2010, only 24% of people in the UK use their mobile phone to access the Internet. In addition, [10] shows only 27% of teens are interested in internet gaming and 82% of teens are more likely to play games alone, where GWAP have not taken these factors into consideration when designing games.

Although the literature is full of game-based approaches, little research has been conducted on the use of standalone games and Game Theory (GT) based approaches for image annotation. Recent research aimed at image annotation is strongly influenced and inspired by social aspects of the human condition [11], and as a result, a number of game based approaches are introduced [5] [7] [8]. Since millions of people like to play games on a daily basis, there is no doubt of the efficiency of such systems. Game-based approaches are not only attractive to derive

1. INTRODUCTION

practical annotation techniques, but also derive strategies aimed at improving existing methods.

The work presented in this thesis builds on the theory of Economics, namely Game Theory. Furthermore, statistical inference such as Markovian and Sampling theories are also used in this work. Similar GT based approaches have been successfully applied as alternative methods for the purposes of decision making and aggregating different information for multiplayer games [12]. However, to the best of our knowledge, no studies have been undertaken so far on the application of these models for image annotation when using standalone games. Game Theory has become successful in recent years because it fits so well into the new methodology of Economics. Nowadays, all economists start with primitive assumptions about the utility functions, production functions and endowments of the actors in the models [13]. The reason is that it is usually easier to judge whether primitive assumptions are sensible than to evaluate high-level assumptions about behaviour and consequently it is widely used in decision making and aggregating information in competitive environments [14]. The Markov chain is a characterization of a system that transits from one state to another. It concerns any random process given with the Markov property, i.e. the property, simply said, that the next state depends only on the current state and not on the past [15]. As a fact, Markov models (MM) are mostly used in statistical modelling and for outcome predictions of human behaviour [16] [17]. Sequential Sampling (SS) is the part of statistical practice concerned with the selection of a subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference [18]. Both MM and SS methods feature some of the most desired characteristics of prediction. For instance, Markov chains predict an outcome based on the present state outcome and that strongly represents the human behaviour in practice [19]. In the other hand, SS uses all available historical data for decision

making. Furthermore, it knows the risk of making a wrong decision and that makes a short coming of the other well-establish approaches [20] [21]. In this work, we investigate the application of the aforementioned algorithms for game-based annotations of images.

1.1 Research Objectives

Image annotation is the first step towards the semantic based indexing of multimedia content. The main goal of the proposed work is to annotate images based on human perception using a framework of standalone games and to reject bad annotations from cheating oriented players. Addressing this problem, the thesis focuses on the following specific objectives:

- To investigate the application of human based computation models, in particular standalone game-based approaches for image annotation.
- To study and develop a standalone game for image annotation whilst exploiting the player's contributions in gaming, previously recorded player's contribution and image classification outcomes to improve annotation accuracy.
- To study and enhance the performance of Game Theory based decision making mechanism for image annotation with single player models.
- To develop an approach for predicting the player's intention prior to exposing non-annotated images based on Markov and Sequential Sampling techniques.
- To evaluate the usability of the gaming system.

1. INTRODUCTION

- To evaluate the proposed system for image annotation by using real-world examples.

1.2 Contribution of the Thesis

The thesis provides significant technological contributions to the following areas:

- A standalone game for image annotation is implemented concerning the player's interaction and the use of Game Theories and strategies.
- A Game-based framework is implemented for aggregating the player's contribution, previously recorded players contribution and image classification outcomes for obtaining useful annotations.
- A Game Theory based decision making technique is introduced for the purpose of concluding player outcomes, i.e. to accept or to reject a player's annotation in a fair manner.
- Prediction based on Markovian inference and Sequential Sampling techniques are introduced to minimise the risk of having bad annotations.

The research described in the thesis and improvements of conventional approaches have been presented in a number of author's publications, which are given at the end of this thesis.

1.3 Structure of the Thesis

The thesis gradually introduces the model for developing a game-based image annotation system. The association of game strategies and theories are exploited to filter out the obtained annotations. In the following, the general structure of the thesis is presented.

- **Chapter 2:** In this chapter, an overview of the state of the art game-based image annotation techniques is discussed. Moreover, the concept of game-based image annotation and its key role in image indexing and retrieving is examined. The overview is a result of the literature reading and investigation into state of the art techniques in the area of game-based image annotation.
- **Chapter 3:** This chapter presents the central contribution of the thesis. It gives a comprehensive overview of the proposed game-based annotation approach; in particular, this introduces the functions implemented for the aggregating player's contribution, previously recorded player's contribution and image classification outcomes to improve the annotation accuracy. Moreover, this introduces the proposed outcome prediction models, i.e. Markov model based influencing technique and Sequential Sampling techniques for improving the annotation performances.
- **Chapter 4:** Game Theory models and strategic situations, in which an individual's success in making choices depends on the choices of the other participants. It is widely used in economics, politics and social psychology. It was initially introduced to analyse competitive environments. However, nowadays, Game Theory is used as a general theory for the rational aspect of social science, where it is broadly used in a certain way to allow participation of human as well as non-human players. In this chapter, an overview of

1. INTRODUCTION

motivations and subsequent related Game Theory inspired techniques, focusing on the Nash Equilibrium (NE) based decision making and aggregating techniques based on payoff functions are introduced. This section will also introduce the use of the Nash Equilibrium based decision making concept for the proposed game based framework. To use Nash's concept, two payoff functions are proposed in this section.

- **Chapter 5:** An important characteristic of a prediction algorithm is the ability to learn from previous experience in order to predict the future outcomes. The need for learning the process has led to vast amounts of research into the construction of prediction algorithms. Typically, prediction of human behaviour is the most difficult task to achieve in practice. The reason for this arises from the fact human behaviour is random and dynamic. In this chapter, we have introduced two different player prediction methods. One method is based on well known Markov chains and the other method is based on sampling theory, in particular Sequential Sampling. Both these techniques are extensively evaluated and selected results are given in the next chapter.
- **Chapter 6:** In this chapter, an extensive experimental evaluation of the game based annotation framework is presented. The challenge of developing an efficient game-based annotating framework involves capturing human attention. The first section of this chapter is dedicated to evaluating the performance of the image classification process. While the second section evaluates the usability of the game, in particular excitement, addiction, enjoyment and the difficulty in game play. These factors are been compared with two well known games, ESP and Phetch. The third section evaluates the efficiency of the proposed framework and finally a comprehensive evaluation of the proposed technique for image annotation is evaluated using three real

1. INTRODUCTION

image datasets. Here, selected representative results are reported.

- **Chapter 7:** This chapter discusses the introduced contributions of game-based annotation and closes the thesis with a relevant conclusion and an outlook to future work.

Chapter 2

GAMES WITH A PURPOSE: A SURVEY OF RELATED WORK

2.1 Introduction

The amount of visual information (images and videos) in digital format has grown exponentially in the last decades. This information is stored everyday to make huge databases and to distribute through the internet. However, most of this information is unstructured and as a result it is hard to search for a particular content. The goal of the field of image annotation is to develop new technologies to index visual contents and summarises them in an efficient way.

Automatic image annotation is the process in which computer systems automatically assign a key-word to visual content. Regardless of the popularity and need for automated image tagging, the field is still very much an open problem and this can be attributed to the existence of the semantic gap. This existence makes it hard to find a relationship between two things; first, the image representation, which is often called the low-level features and secondly, the visual object, which is often called the high level concept. In addition to these reasons, there are many

2. RELATED RESEARCH

other factors which make it harder to find a solution to the semantic gap problem. They include occlusion, background clutter, scale variations, view point etc. Since the early 90's, a significant amount of research related to image annotation has been conducted. Some initial efforts have recently been dedicated to automatically annotate images [22] [23] [24], image understanding and statistical learning [25] [26], visual templates [27], support vector machines (SVM) for image classification [28], context models [29], feedback learning [30]. Most of these approaches tackle the semantic gap problem by using machine learning techniques and using mainly two categories dependent on the scale of image analysis, namely Global feature based image tagging (scene-based approaches) and Block/region-based image tagging.

Global feature based image tagging approaches utilizes the properties of global image features such as colour and texture distributions. The key-idea is to somehow find a feature representation that is separable enough to distinguish between different classes of scenes. In [31], the author has used a SVM classifier [32] on a global HSV colour histogram to find the image of interest, while [33], employs a classification tree to model spatial correlation on colours, which both are popular approaches in the literature. The main disadvantage of using global features is that the features used are often insufficiently representative of the prominent objects that are used to represent the image or the scene.

Block/region-based image tagging approaches use object based image tagging. In the region based approach, the region of interest is extracted from the image. Namely, this process is called image segmenting. It identifies real world objects within the image. The general assumption is that feature extraction is based on a strong segmentation that better describes the visual object. However, limitations in automated segmentation make it harder to obtain a promising result in image classification.

2. RELATED RESEARCH

Problems in automated annotations and text-based access to images have driven interest in the development of image-based solutions. This is most often referred to as CBIR. Content-based image retrieval relies on the characterization of primitive features such as colour, shape and texture that can be automatically extracted from the images themselves. Queries to CBIR systems are most often expressed as visual exemplars of the type of image or image attribute being required. For example, users may submit a sketch, click on a texture palette or select a particular shape of interest [34]. The system then identifies those stored images with a high degree of similarity to the requested feature. In [35], various technologies for image indexing and retrieval based on shape, colour, texture and spatial location are discussed.

2.2 Towards human augmented image annotation

Numbers of online applications such as search engines require accurate image descriptions. However, there is no way to provide accurate textual descriptions for the millions of images which are online and in private databases. Manual labelling is the only method for obtaining correct image descriptions, but this process is very expensive and labour-intensive. Many tasks are trivial for humans but may be challenging to the most complicated computer programs. In general, such problems are solved by using artificial-intelligence. Addressing the semantic gap in the computer vision community, it is still hard to find a complete solution by using artificial intelligence. However, GWAP addresses this problem by constructing an environment for the channelling of human brain power through computer games [36].

Though previous research recognizes the utility of human cycles and the motivational power of the game like interfaces, earlier approaches were unsuccessful in harnessing human attention through computer games [36]. Some of the earliest examples of collaborative work can be dated back to the 1960's, where

2. RELATED RESEARCH

open source software development projects were introduced with network individuals accomplishing work online. The collaborative efforts by large numbers of individuals accomplished tasks that are impossible to achieve by ordinary computers and their programs. This collaborative work would save time and effort of an individual person by splitting work amongst a large group of individuals. Amazon Mechanical Turk system [37], is an example for such a collaborative work. Here, large computational tasks are split into smaller chunks and divided among a large group of individuals.

In the United States alone, 200 million hours are spent each day playing computer and video games [36]. By the age of 21, an average American has spent more than 10,000 hours on playing computer and video games which is equivalent to five years of full-time working. Addressing this fact, the GWAP is designed to channel time and effort towards solving computational problems and to improve the outcome of artificial intelligent algorithms. Unlike computer processors, humans require some incentives to become a part of a collective computation. In order to tackle this, GWAP was designed to target the online collaborative approaches, such as the multiplayer environment; a source that encourages people to participate in the process. Implementing a GWAP is much like designing an algorithm which has to be proven addictive to the players and providing correct outcomes.

In the literature, it has been shown that GWAP is used to annotate different types of multimedia materials or features [4] [38] [39]. The ESP was probably the first game designed to harvest image annotations and has led to a number of related approaches including: Squigl¹, Hot or Not², Google Image Labeller³, Verbosity [5], ASAA [6], Manhattan Story Mashup [7], KissKissBan [8], Phetch [39], Matchin [40] and Peekaboom [41]. The GWAP approach is characterized by a number of monitoring factors; an increasing number of internet users; people spending a lot of

¹ <http://www.gwap.com/gwap/gamesPreview/squigl/>

² <http://www.hotornot.com/>

³ <http://images.google.com/imagelabeler/>

2. RELATED RESEARCH

time on computers playing games; more accurate information on multimedia contents can be collected; human attention could be collected easily for no cost.

2.3 Existing game-base frameworks

Having published many GWAPs, authors in [36] have listed three game structural templates to generalize successful instances of human computation games; output agreement games, inversion problem games and input-agreement games.

2.3.1 Output agreement games

Output agreement games [42] are a generalization of the ESP game. Here, two players are chosen randomly among a large group of players and will be given the same content for both players as the input. Players are asked to provide outputs based on the given input.

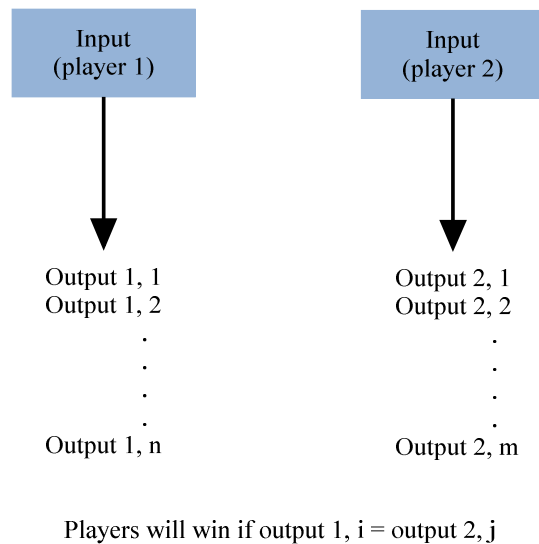


Figure 2.1: Output agreement game mechanism.

Here, players are indirectly forced to produce the output based on the input content because the players are restricted from communicating with one another. To

2. RELATED RESEARCH

win the game, both players must produce the same output which does not have to be produced at the same time. Since both players are restricted from communicating, they do not know anything about the other player's output. Therefore, the easiest way for both players to produce the same output is by entering something related to the input content. This game strategy forces players to produce outputs related to the input which is the only thing that both players have in common. In Figure 2.1 an example of the output-agreement game is shown.

2.3.2 Inversion-Problem games

Inversion-Problem games [42] choose players randomly from a large set of players. Here, one player is assigned the role of “describer” and the other player is assigned as the role of “guesser”.

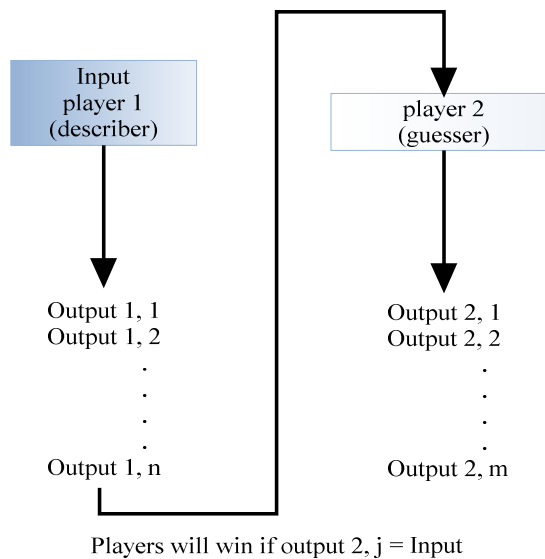


Figure 2.2: Inversion problem game mechanism.

The game chooses the input content and gives it to the describer. The describer produces output (in many games a single word or sentence) based on this input. The objective of the describer is to help the guessers to produce the original input. In these types of games, partners are successful when only guessers describe the input

2. RELATED RESEARCH

content correctly. If the describer's outputs are incorrect or incomplete, the guesser will not be able to produce the original input. Peekaboom, Phetch and Verbersity are some inversion games introduced in the literature. In Figure 2.2 an example of the inversion-problem is shown.

2.3.3 Input-agreement games

Input-agreement games [43] choose players randomly. In each round, both players are given inputs that are to be the same or different, known by the game itself but not by the players.

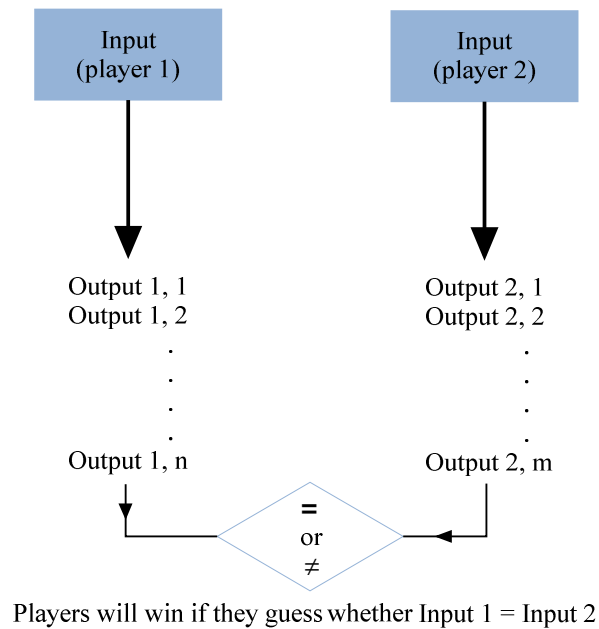


Figure 2.3: Input agreement game mechanism.

The players are told to describe their inputs, so their partners are able to assess whether their inputs are the same or different. Both players will win if they correctly determine whether they have been given the same or different inputs. "TagATune" [4] is an example game for input agreement game. Here, players are given a sound clip as the input and they have been asked to verify whether both

have been given the same input. Because players want to achieve a winning condition, they both want their partner to be able to describe the correct information. In Figure 2.3, the mechanism of the input-agreement game is shown.

2.4 GWAP - Popular games and their strategies

There are number of GWAP introduced in the literature. However, we are mainly concerned with their workings which have contributed largely to the multimedia community by addressing the semantic gap problem. In Figure 2.4, overviews of existing GWAP's are shown.

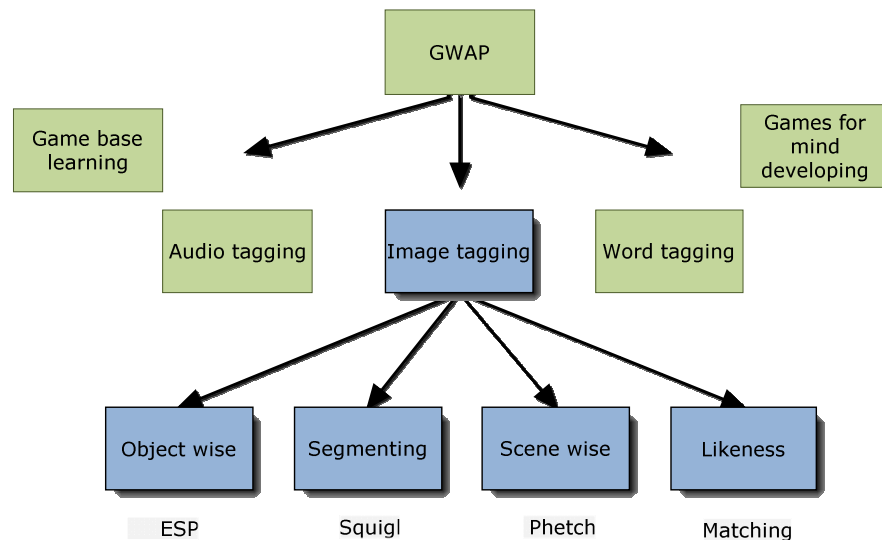


Figure 2.4: Overview of “GWAP”.

Games for object annotation

There are a number of games introduced in the literature for object wise image annotation. Among them, the ESP is the first and the most popular game that annotates images based on human perception. This game was introduced in 2003 and was played by 13,630 individuals [3] within the first four months. The game is

2. RELATED RESEARCH

designed using a java applet and the applet is connected to a main server for the purposes of data handling and monitoring. This game is designed to be played by two partners and is meant to be played online by a large number of pairs at once. Partners are randomly selected from among all the people playing the game. Players are not told who their partners are, nor are they allowed to communicate with their partners. The only thing partners have in common is an image they can both see. From the player's point of view, the goal of the ESP game is to guess what their partner is typing for each image. Once both players have typed the same key-word or string, they move on to the next image. Here, both players do not have to type the string at the same time, but each must type the same string at some point while the image is on the screen. Every time the players agree on an image, they will be rewarded with a certain number of points, encouraging them to play more in gaming. This game uses numerous techniques to prevent cheating. The IP addresses of players are recorded and allocated differently from that of their partner to make it difficult for players to be paired with themselves. To prevent global agreement of a strategy such as typing 'a' for every image, the game use pre-recorded game-play. If a massive agreement strategy is detected, the game insets a large number of bots to make it harder for cheating.

KissKissBan (KKB), for image annotation, is a different game from other human computation games. Here, the game is designed to be played by three online players. One of the players is called the "blocker" and the other two players are called the "couple". With the same image presented, the couples try to match (Kiss) with each other by typing the same word and the blocker tries to stop couples from matching (Ban). The blocker is only given seven seconds to act in each round and he/she is able to see every word the couples are typing during the game. Monitoring the actions of the couples not only makes the waiting process fun, but provides the blocker with an opportunity to stop the couples from achieving some unified strategy. For example, the blocker could give "a" as the blocked word if he/she found the couples trying to match on "a" in every round. The objective of the

2. RELATED RESEARCH

couples is to guess what the partner is typing. However, unlike the players in the ESP Game, the couples in KKB cannot see what the blocked words are. Therefore, the couples are encouraged to guess harder words to avoid guessing the word in the blocked words list.

Annotation by key-sentences

Image describing by key-sentence is another method to tackle the semantic gap problem. In [7] and [39], authors have addressed the benefits of this initiative by introducing 2 different games, namely Manhattan Story Mashup (MSM) and Phetch. Manhattan Story Mashup is a large-scale pervasive game, which combines the web, mobile phones and one of the world's largest public displays in Times Square. Here, the web players used a storytelling tool at the game website to mash up stories, either by writing new sentences or by re-using already given sentences. A noun from each new sentence was sent to a street player for illustration. The street player had to shoot a photo which represents the word within 90 seconds. The photo was then sent to two other street players who had to guess what the photo depicts amongst four nouns, including the correct one. If the photo-noun pair was guessed correctly, the original sentence was illustrated with the new photo and it was turned into an ingredient for new stories. Here, players will be rewarded by displaying the best story on the Reuters Sign in Times Square in real-time. This game was deployed as a part of SensorPlanet project at Nokia Research Centre to examine the player's creativity by exploiting ambiguity and how the players were engaged in a fast-paced competition.

Phetch is an online game played by three to five players. Initially, the game chooses one of the players as the "Describer" while the others are "Seekers." The Describer is given an image and helps the Seekers find it by textually describing it. Only the Describer can see the image and communication is one-sided: the Describer can broadcast a description to the Seekers but they cannot communicate back. Given the Describer's text, the Seekers can find the image using an image

2. RELATED RESEARCH

database which contains a large number of images. However, they are not cued as to how to extract a search query from the given text. The first Seeker to find the image obtains points and becomes the Describer for the next round. The Describer also gains points if the image is found. Unthinkingly, by observing the Describer's text, a collection of natural language descriptions of images are obtained. Here, the main disadvantage is that the Describer's text could contain unrelated textual descriptions, which is being posted among the related descriptions to generate false annotations.

Games for image segmentation

PeekaBoom and Squigl are two different games introduced in the literature for image segmentation. These games are designed to be played by two players that are randomly chosen. In PeekaBoom, players are named as 'Peek' and 'Boom'. Initially, Peek starts out with a blank screen, while Boom starts with an image and a word related to it. The goal of the game is for Boom to reveal parts of the image to Peek so that Peek can guess the associated word. Boom reveals circular areas of the image by clicking. A click reveals an area with a 20-pixel radius. Peek, on the other hand, can enter guesses of what Boom's word is. Boom can see Peek's guesses and can indicate whether they are hot or cold. For example, if the image contains a car and a dog and the word associated to the image is "dog," then Boom will reveal only those parts of the image that contain the dog. Thus, given an image-word pair, data from the game yield the area of the image pertaining to the word. If Peek managed to guess the correct key-word, both players will be given some points that encourage them to play further.

Squigl is another type of GWAP introduced for image segmentation. This game is designed to be played by two players, where players are given the same image associated with a keyword. Here, players are supposed to draw the contour of the key-object. Based on both player outcomes, the similarities are analysed by the framework. Depending on the similarity, players are assigned game points that

2. RELATED RESEARCH

encourage them to play the game further. The main purpose of this game is to generate a database of segmented objects that can be used in machine learning.

Games for Gesture based Image Tagging

The ASAA is the first game designed for gesture based image annotation. The game consists of a combination of manual and automatic image annotation, with interaction by means of gestural signs in front of a camera. Here, the game interface provides a three dimensional game, where people move tags and images, using a motion detection algorithm applied to the captured (user) image. The ASAA game uses semantic image annotation by means of a set of concepts previously trained for image classification. This information is used to calculate the score and the annotated images are used to refine the semantic concept models.

Games for image rating

Matchin [40] is another GWAP used to annotate images based on the likeness. This game is a two player model that gives both players two images for voting. If the players vote for the same image, they will be given some points encouraging further engagement in gaming. The objective of this game is to create a large database of images based on image likenesses. In [44] another approach is introduced for rating people based on pictures. The approach is called “HOT or NOT” which is a social entertainment website launched in the year 2000 and has been successfully subscribed by millions of members.

2. RELATED RESEARCH

In Figure 2.5, a number of screen shots for GWA^P are illustrated.

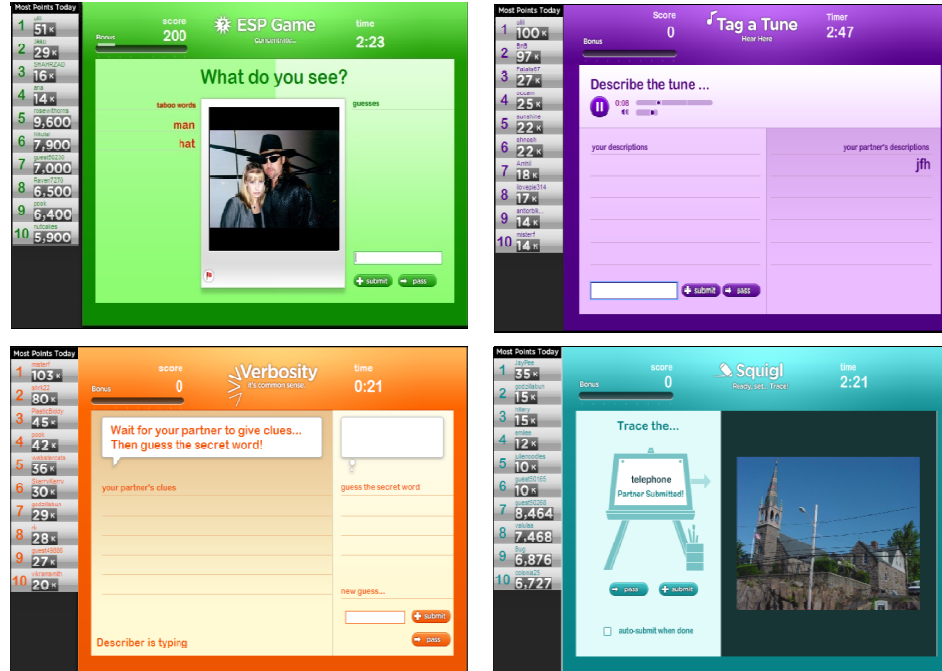


Figure 2.5: Visual representation of current GWA^P approaches.

2.5 Summary

This chapter brought an overview of current techniques for game-based image annotation and their strategies, principles, advantages and disadvantages. It is known that the semantic gap between low-level features and high-level concepts remain as the biggest challenge to the research community. Addressing this problem, a number of techniques have been proposed in the literature. The game-based approaches are among the best that give promising results in image annotation. A number of games that have already introduced image annotation can be categorised into three different structural templates, as output agreement, input agreement and inversion problem games. These templates have been used to design games for deferent purposes, such as object-wise image annotation, key sentence-wise image annotation, image segmentation, image rating etc. In the next chapter,

2. RELATED RESEARCH

the proposed framework for image annotation is presented. Here, decision making strategy based on Nash Equilibrium is used to aggregate different information. In this thesis, we used this technique as a source of inspiration for the design of novel approaches for image annotation.

Chapter 3

PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

The major contribution of this thesis is the design and implementation of a framework combining several paradigms for archiving more realistic image annotation. The proposed framework is a standalone game (single player), used to annotate images purely based on the players intention. However, the underlying approach is designed based on a two-player game model. It gives the independence to combine a number of different paradigms, such as player outcome prediction algorithms, Game theory based decision making concepts and the players overall contribution in annotation. In this section, a novel approach for image annotation based on Game Theory is presented. GT and its driven mathematical models are introduced to make decisions on the player's outcome, i.e. to accept or to reject the player's annotation.

3.1 Introduction

Many methods have been proposed for game based image annotation, but almost all these methods refine annotations using multiplayer game strategies (see Section 2.3). Furthermore, only a small amount of research has been undertaken for image annotation using Game Theories. In [3] [12], it is shown that game approaches provide more encouraging results in image annotation than automated approaches. In [6], the potential of obtaining promising results using standalone games are shown. However, the approach is designed based on a classifier and that forms limitations in this approach.

The proposed approach follows the well-known crowdsourcing paradigm, in which a given problem is tackled by exploiting the power of users in a widely distributed way. In our case, the aim is to harvest the power of millions of computer gamers for the purpose of annotation digital multimedia as in Flickr⁴, Facebook⁵ and Dailybooth⁶. Crowdsourcing has been successfully used for harvesting multimedia annotations. A commonality of all these games introduced in Section 2.3 is the use of at least two-players interacting remotely through the Internet so as to prevent cheating and control a potential flow of misleading annotations into the metadata base. A more critical issue related to cheating prevention in ESP-like games is the latent possibility of remote gamers agreeing on a strategy that can be used to provide quick useless annotations but yet obtaining high scores in the game. This and other drawbacks of ESP-like games are discussed in [8]. In contrast to ESP-like strategies, the approach proposed here can be instantiated as a standalone game or be deployed over the internet as well. Considering problems and limitations in multiplayer game approaches, we decided to explore the problems in the standalone framework because we are interested in finding the usefulness of

⁴ <http://www.flickr.com/>

⁵ <http://www.facebook.com/>

⁶ <http://dailybooth.com/>

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

standalone frameworks in image annotation.

The proposed framework can be initiated by single standalone games and is based on a two-player model. In this model, the gamer (user) takes the role of Player 1 while the machine takes the role of Player 2. The underpinning model considers two different types of gamers: rationally minded and malicious or deceptive players. It uses an outcome prediction mechanism to expose the player to the most suitable multimedia material, i.e. fully annotated (images that are fully annotated by a paid human operator), partially annotated (images that have obtained one or more annotations) or non-annotated (images that have no annotations at all) contents. For comparative purposes two different prediction techniques are proposed, one based on Markovian Model based inferences [15][45] and the other based on Sequential Sampling plans [5]. The proposed framework uses a pair of profile payoff functions to refine the player's outcome by finding its unique Nash Equilibrium [46]. It is further shown that the Nash Equilibrium of the model is equivalent to a fair solution and leads to a win-win situation. The player's reward is calculated according to a suitable scoring mechanism which encourages the user to carry on playing. The score is measured based on the player's dedication to the game, i.e. the player's potential to provide correct annotations and player's cost, i.e. the effect of incorrect or misleading annotations. Although the proposed model is suitable for many kinds of games targeting diverse types of media annotation, the game presented in this thesis is designed to annotate only still images.

3.2 System Overview

A diagrammatic overview of the proposed approach is given by Figure 3.1. The system relies on a small number of previously annotated images and a transitional database for storing partially annotated images. That is, the entire image database consists of three subsets: fully annotated, partially annotated and non-annotated images. Initially, the fully annotated subset would consist of a small number of images previously annotated by a human operator and the set of partially annotated

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

images is empty. Once the game is deployed, it is expected that both fully and partially annotated image sets start to grow as semantic metadata is obtained through the game. Thus, the complete framework comprises two main modules. The first module (right in Figure 3.1) handles fully-annotated multimedia units, while the second module (left in Figure 3.1) deals with partially and non-annotated multimedia units. The first module is used to understand the player's behaviour, confirm results from statistical inference, as well as estimate model parameters and the shape of its payoff functions. The second module is the actual annotation engine providing semantic metadata for non-annotated content.

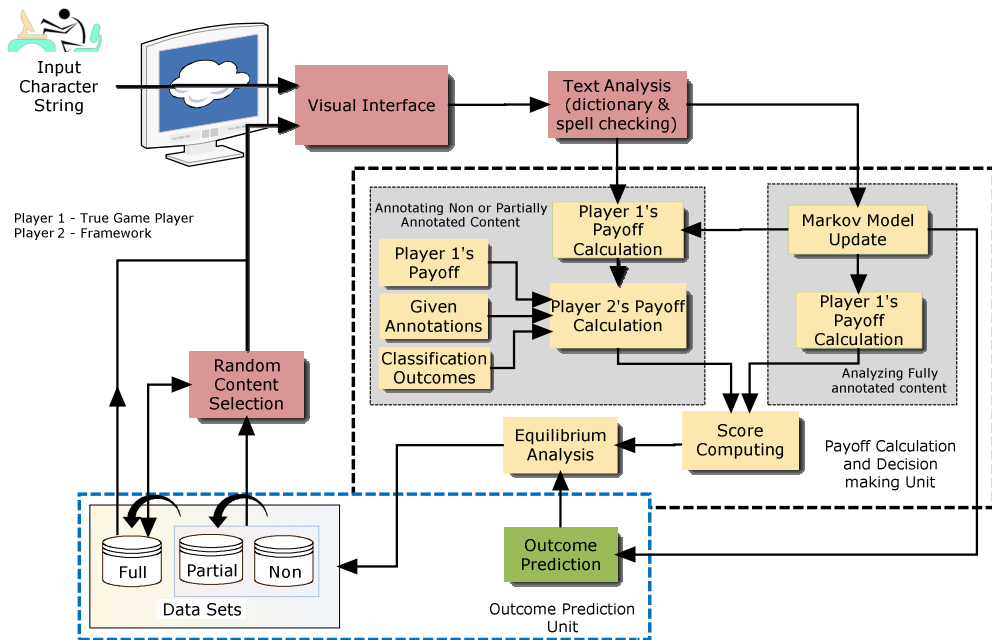


Figure 3.1: A complete block diagram of the framework.

In our case, two generic types of gamers are considered: rationally minded and malicious or deceptive players. The first type of player plays the game in a fair way trying to achieve high scores by correctly annotating content. This type of player is called “rational” in the sequel. The second type of player contains all those who try to achieve high scores by cutting corners and cheating. These types of player are called “malicious” in the sequel. Clearly, there will be users that change behaviour

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

while playing. Thus, the system assumes that a rational player can become malicious and vice-versa. At the start, a small set of fully annotated content is fed to the game to initiate the process of learning player's behaviour and model parameters. Here, a transition matrix (the one used by the proposed Markov model prediction) is used to measure player's contributions to the game, i.e. the player's potential to provide correct and meaningful annotations; and the player's cost, i.e. the effect of incorrect or misleading annotations. Next, content is extracted from one of the three available databases (fully, partially or non-annotated) and uploaded into the system.

Database selection for content extraction depends on the predicted player's behaviour as detailed in Chapter 5. In subsequent steps, this prediction is done by taking into account the previous outcomes of the player for a series of fully annotated contents. When the player prediction module expects an incorrect annotation with high probability, then it exposes a fully annotated unit to the player. On the other hand, when it predicts a valid annotation with high probability, it loads a partially annotated or a non-annotated piece of content, based on the outcome of equilibrium analysis module. However, a module referred to as a Random Content Selection module that forces extraction of content from the fully annotated multimedia database at random time intervals is also used. Given that, in practice, players change their behaviour often and rational-minded players could thus become malicious, the Random Content Selection module addresses this problem by exposing the player to a number of fully annotated contents at random time intervals. The outcomes for these images are used to update the state of the MM, with the aim to assist MM in representing the player's latest behaviour in gaming.

The visual interface is the window of the instantiated game. Its design depends on the game strategy. However, it has two fundamental tasks: to expose content to the player and to enable input of character or strings to be associated with the exposed content. Two different game strategies are described in Section 3.3.1, to illustrate the corresponding visual interfaces and test the performance of the

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

proposed model.

Although Figure 3.1, shows that the proposed framework is developed using a number of modules, the entire process can be summarised using two major units, namely, the Payoff Calculation and Decision making unit and the Player Outcome Prediction unit.

3.2.1 Payoff Calculation and Decision making Unit

Payoff calculation and decision making is one of the important units in this framework. This unit mainly relies on the player's outcome, i.e. accuracy of the player annotations. The fundamental algorithm is designed to measure the player's contribution in gaming in order to expose them to the most suitable, i.e. fully annotated, partially annotated or non-annotated content, as well as to decide whether to accept or to decline the player's outcome. In order to do so, two payoff functions are constructed and represent both players' contribution in gaming. In the beginning of each game, a player will be exposed to a number of fully annotated images. Outcomes are then used to form a transition matrix. This matrix is used to measure the Player 1's overall contribution in gaming as well as the cost. Here, player's overall payoff is measured by subtracting the player's bad contribution from the good one. Player 2 in this game is a virtual player and therefore his contribution is measured based on a number of different aspects. Here, the payoff function used to measure Player 2's payoff is designed to aggregate number of different key factors; Player 1's payoff, previously recorded players contributions and image classification outcomes. Since, in this game, players are not fully independent, and given that the objective of the machine (that takes the role of Player 2) is to encourage Player 1 to produce correct annotations, it is fair to use Player 1's information, i.e. payoff or any other information to measure the Player 2's payoff. More formally, if the machine motivates Player 1, it can be assumed that the probability of entering a previously recorded annotation by Player 1 would increase. This further confirms the suitability of using Player 1's information to

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

assign correct weights the Player 2's payoff. In addition, the Player 2's payoff is weighted by a SVM classifier, whereby the classifier selects the most optimal trained concept from a set of pre-trained concepts based on the player's input keyword. The probabilistic outcome (the probability of an image being relevant to the trained concept) from the classifier is used as a factor for weighting the Player 2's payoff.

For partially annotated and non-annotated contents, the Player 2's cost is calculated based on the number of different annotations that have been obtained by an image. In practice, if the framework performs well, annotations from cheating oriented players will be recognised. As a result, the framework would accept a few different annotations, i.e. those from trustworthy players. Thus, using the number of different annotations assigned to an image for calculating the Player 2's cost is the most optimal solution.

The proposed framework uses game theories, in particular Nash Equilibrium based decision making techniques for exposing the player to the most suitable image content, i.e. fully annotated or non-annotated. Since game theories postulate that decisions should be based on primitive actions, two different primitive, yet influential, game actions are introduced to the system. One of these actions represents the short-term contribution of the Player 1 in gaming and the other action represents the long-term contribution of Player 2 in gaming. More information on game theory based decision making is given in Chapter 4.

3.2.2 Player Outcome Prediction Unit

Player outcome prediction is the second most important unit in this framework. It is used to predict the player's outcome prior to exposing images, fully annotated, partially annotated or non-annotated. As a result, it is used to improve performance of this framework. In this thesis, we have introduced two different outcome prediction algorithms. One is based on the well known Markov chains and the other one is based on Sequential Sampling plans. Since human outcomes are dynamic and

do not follow any sequence, it can say that prediction based on present intention is the most practical approach to predict future outcomes. This dilemma is highly encouraged by the research conducted in [19]. Here, it says that there is a high potential that human behaviour depends on current intention and is not based on past performances. Since Markov chains predict future events based on the outcome of the present event, we used Markov chains in this thesis to predict human outcomes. To compare and evaluate the Markov prediction approach, we also used sampling algorithms to predict human outcomes, in particular Sequential Sampling, where the prediction and decision making is influenced by examining the entire distribution, not only based on the present outcome. Unlike the Markov approach, SS is well known and the involved risk of accepting a defective sample is what makes it admired when compared to the Markov approaches. More information on this unit is given in Chapter 5.

3.3 Implementation of the framework

During the process of developing the two main units, each of which implements some algorithm studied and proposed are given in this thesis. In order to build a successful system, those two units have been carefully integrated. The objective was to get the user's attention; a graphical interface that can satisfy the goal of the proposed approach is also implemented. The main goal of the approach is to build up a simple framework that can satisfy a large number of game players, thus could obtain a large number of valid annotations for a given set of images. Bearing in mind the above objectives, the following major features were constructed in devising the process:

- An easy to use and attractive human-machine interface.
- Support for the storage of metadata.
- An easily adopted method for annotating various type of multimedia content.

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

Currently backbone structure and basic modules of the targeted system have been implemented. Based on these, an experimental environment has been composed for image annotation. The proposed framework is mainly implemented using C++. Implementation of GUI (Graphical User Interface) employs the OpenGL Application Programmable Interface (API) development environments. This API is the interface implemented for the game application which allows the other applications to communicate with it. OpenGL is an open source toolkit designed to provide efficient, portable access to the user interface facilitated by the operating systems on which it is implemented. It is a premier environment for developing portable, interactive 2D/3D graphic applications. OpenGL has become the industry's most widely used and supported 2D and 3D API [47]. OpenGL fosters innovation and speeds application development by incorporating a broad set of rendering texture mapping, special effects and other powerful visualization features.

3.3.1 Graphical User Interface

For testing purposes, we developed two graphical interfaces. The first interface, (denoted by INT-1) is designed based on a scenario where the players are asked to create a keyword by picking characters from a series of dropping characters. This interface displays 4 to 5 characters at a time. Players have to collect each character by using arrow keys on the keyboard. For example, if a player wanted to enter the keyword "CAT", he/she would collect each character "C","A" and "T" in a sequential order.

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION



Figure 3.2: Some screenshots of INT-1.



Figure 3.3: Some screenshots of INT-2.

This interface additionally displays a number of magic characters that can be changed into any character which is demanded by the player. To fulfil a player requirement, this game allows players to change the speed of the spinning characters. The second interface, (denoted by INT-2), was based on a design of a simple game scenario. Players were asked to annotate images by typing a key word. Here, the image subject for annotation is randomly displayed in one of 6 displays

and the player is asked to steer himself towards the image and enter a keyword. In the case of a player bumping into given obstacles or unable to complete annotations in a given time frame, they will be given a life penalty. Here, a set of screenshots are shown in Figure 3.2 and Figure 3.3, for INT-1 and INT-2 respectively.

3.4 Experimental Setups

A general experimental environment is constructed by implementing frameworks as mentioned earlier in this chapter. It contains different setups for specific experiments on proposed algorithms. However, some of these setups are common to all the experiments that have been conducted in this thesis. Since this framework depends on human players, we used a number of different players to evaluate the framework. Moreover, we used three natural image databases for these experiments, namely, ESP, Caltech and Corel image datasets. Details of these datasets are given as follows.

ESP dataset

The first dataset is a small set containing 200 images selected from the ESP dataset, which is referred as the ESP dataset in this thesis. Here, manual labelling of the ground-truth for 100 images were conducted prior the experiments. These images contain complex scenes and scenarios with large numbers of objects present, such as busy streets, seaside, landscape, office environments etc. Therefore, they cannot be categorised into a particular semantic category. Since the ground-truth for this dataset is known, player outcomes from these images have been used to measure the player confidence in image annotation.

Caltech 101 dataset

The second dataset is a small set containing 200 images selected from the Caltech 101 dataset, which is referred as the Caltech dataset in this thesis. Here, manual labelling of the ground-truth for 100 images were conducted prior to the experiments. This dataset contains a higher level of ground truth based on semantic

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

meaning. Images belonging to the same class illustrate the same concepts, however, their visual appearance is different. This dataset consisted of 101 object categories which do not overlap with any other concepts.

Corel dataset

The third dataset is a small set containing 200 images selected from the Corel dataset, which is referred as the Corel dataset in this thesis. Here, manual labelling of the ground-truths for 100 images were conducted prior the experiments. This dataset contains a higher level of ground truth based on semantic meaning. Images belonging to the same class illustrate the same concepts, however, their visual appearance is different in practice. The dataset consists of seven concepts, namely, Car, Lion, Tiger, Cloud, Elephant, Building and Vegetation.

3.5 Summary

In this chapter, an overview of the proposed framework was given. Although the proposed framework is developed using a number of modules, the entire process can be summarised by using two major units, namely, payoff calculation and decision making unit and player outcome prediction unit. The payoff calculation unit is designed to measure the player's contribution in gaming in order to expose the player to the most suitable content. In other words, the optimal fully annotated, partially annotated or non-annotated content is selected based on the Nash Equilibrium based decision-making process. The player outcome prediction unit, on the other hand, enhances the performance of the payoff calculation and decision making unit by predicting the player's outcome. Here, two different graphical user interfaces were developed for testing purposes. However, in practice, the annotation is achieved by offering the image subject to the player through the interface and prompting the players to comment on it using a string of characters. This string is subsequently analysed by the dictionary analysis module to establish whether the player has entered a valid keyword. Following the keyword search, the payoff

3. PROPOSED FRAMEWORK FOR IMAGE ANNOTATION

calculation and equilibrium analysis unit will measure each player's payoff and finally the score computation module will calculate the scores of both players. This process will continue until a game session ends. Two major units in this framework form the backbone of the proposed research and are elaborated in the next two chapters, where the practical aspects of the framework will be discussed.

Chapter 4

PAYOFF CALCULATION AND DECISION MAKING

4.1 Introduction

Payoff calculation relies on correct decisions to distinguish between rational and malicious players. That is the dilemma faced here. A simplistic way to approach this problem is to state the desired outcome and to behave in a way that leads to attaining that outcome. However, one should ask is it always possible to achieve the desired outcome? Taking this dilemma into consideration, an alternative approach is given in [48]. This approach uses the causes of actions that are available for a problem and determines the outcome for each of these actions; where one of these outcomes is preferred because it is the outcome that maximises something, i.e. payoffs in our case. The causes of actions that lead to a preferred outcome are then picked from the available action set. Whenever players attain this profile of actions, their outcomes are taken as valid outcomes. This technique is called the “making an

optimal decision”. Here, we used this technique to distinguish between rational and malicious players in the proposed framework. In this chapter, an overview of Game Theory for decision making is presented. The chapter mainly focuses on techniques that are closely related to the proposed research in this thesis, without presenting the whole literature.

4.2 Introduction to Game Theory and Decision making

Game Theory is the study of the choice of strategies by interacting rational agents. The main criterion of a game theoretic analysis is to discover which strategy is a person’s best response to the strategies chosen by the other agents. It is defined as [49] the best response for a player as the strategy that gives maximum outcome or a so-called payoff, given the strategy that the other player has chosen or is expected to choose. In general, Game Theory is based on a scientific metaphor, where most of the interactions we do not usually think of as games, such as the share market, investments and insurance companies, can be treated and analysed as we would analyse games. Nowadays, Game Theories treat all kinds of human choices as if they were strategies of a game. In general, Game Theory studies the rational choice of strategies. Human beings are absolutely rational in their choices, especially when they are involved in rewards such as profits, incomes or benefits etc. This hypothesis narrows the range of possibilities, which is that absolutely rational behaviour is more predictable than irrational behaviour. The key idea in Game Theoretic analysis is to discover which strategy is a person’s best response to the strategies chosen by the others. Classical models treat players as inanimate objects and therefore fail in interdependent decision making. Those models neglect the fact that people make decisions and are strongly influenced by what others decide. Game Theory models, on the other hand, are constructed around the strategic choices available to players where the preferred outcomes are clearly defined and known [50].

4. PAYOFF CALCULATION AND DECISION MAKING

It is said that Game Theory was conceived in the seventeenth century by mathematicians attempting to solve gambling problems [13]. However, it is considered to have begun with the publication of Emile Borel in 1921. Since then, a number of papers have been published. Among them, von Neumann and Morgenstern's "The Theory of Games and Economic Behaviour in 1944" [51] presents a basic blend of economics and Game Theory. This introduced the idea that conflict could be mathematically analyzed and provide the most suitable answer. The "Prisoner's Dilemma" [13] and Nash's papers on the existence of equilibrium [52] gives the preliminary essentials of the modern non-cooperative game theory. Around the same time, Shapley [53] introduced rich information about cooperative game theories. Since then, large numbers of works related to Game Theory have been undertaken.

Most of the decision making models that exist nowadays need to make a decision as to which modes to use. In general, there are rational models, intuitive models, rational-intuitive models etc. When considering a decision making problem, one approach to the problem is to determine the desired outcome and then to behave in a way that leads to that result [54]. This approach leaves open the question of whether it is always possible to achieve the desired outcome. Addressing this problem an alternative approach is introduced, where it lists the courses of action that are available and determines the outcome of each of those behaviours [48]. This approach selects one of the outcomes that is preferred because it is the one that maximizes something of value, i.e. payoff, money, profit, etc. The course of action that leads to the preferred outcome is then picked from the available set. This approach makes an optimal decision for the problem of decision making.

4.2.1 Interactive decision problems and static games

In a game-based environment, most of the decision problem involves two or more individuals. Making a decision in such situation is tricky as the payoff to each

4. PAYOFF CALCULATION AND DECISION MAKING

individual depends on what every individual decides. Players in a static game make decisions in isolation. As a result, each player has no knowledge of the decision made by the other players before making their own decision. These games are referred to most of the time as simultaneous decision games because there is no order in which the decisions are made. Simultaneous games are represented by the “normal form”, where the game is shown as a table of numbers with different strategies and solved using the concept of a Nash Equilibrium (NE) [48]. It is so called that the game is in an “extensive form” when games are represented as tree diagrams. In an extensive form, each decision about how that game has been designed to perform is represented as a branch point in the tree diagram.

Rational Behaviour

The rational behaviour is the action made by individuals as they try to maximize the benefits and minimize their costs. In practice, humans make decisions on a problem by comparing the costs and benefits of different actions. The rational behaviour depends on the costs and benefits of certain actions and is easy to explain using the economic theory. As an example, people make decisions about how they act by comparing the costs and benefits of different courses of action.

4.2.2 Cooperative and Non-cooperative games

A **cooperative game** is a game which the players have complete freedom of pre play communication to make a joint binding agreement. These agreements can be used to share payoffs or to coordinate game strategies between players. One can say that this sharing property can simplify the analysis of a game. However, it is not true in a cooperative game as partial agreements may complicate the issue to such an extent that n -person cooperative game theory is neither as elegant nor as cohesive as the non-cooperative case. In [55], an explanation of the complexity of cooperative games is given. It clearly mentions that sharing is not possible or non transferable in some cases such as “years in prison” or “early payroll”, which is practically true. An alternative in this situation is to include the possibility of side

4. PAYOFF CALCULATION AND DECISION MAKING

payments in some transferable unit such as money.

Non-cooperative games are ones in which absolutely no communication is allowed between players and in which players are awarded their due profit according to the rules of the game [55]. Moreover, it is forbidden for players to share payoffs or any information regarding game plan/strategy etc. However, this is not to assert that transitory strategic cooperation cannot occur in a non-cooperative game if permitted by the rules of the game.

The fundamentals of a game are the players, actions, payoffs and information. These are known collectively as the rules of a game. The modeller's objective is to describe a situation in terms of the rules of a game so that it explains what will happen in a situation. Trying to maximize their payoffs, the players will plan strategies that pick actions depending on the information that has arrived at each moment. The combination of strategies chosen by each player is known as the equilibrium. Given an equilibrium, the modeller can see what actions come out of the combination of all the players' plans and this tells the outcome of the game [13].

The basic element of any game is its participants who are independent decision makers called *players*. They may be individual persons or organisations who make decisions. In general, a game must have two or more players. The total number of players may be large, but must be finite and known. Each player must have more than one choice, because a single choice can have no strategy and therefore cannot alter the outcome of a game. Thus, the players' goals are to maximize their utility through a choice of actions. An action is most often represented by the variable a , and the action space of a player is represented by A . A set of actions available to the player can thus be represented by

$$A = \{a_1, a_2, a_3, \dots \dots a_m\},$$

where m is the number of actions available to the player.

4. PAYOFF CALCULATION AND DECISION MAKING

An *outcome* is the result of a complete set of strategies selected by a group of players in a game. If the player is indifferent to the difference between two or more outcomes, then those outcomes are assigned the same numeric payoff. A payoff is a function $\pi : A \rightarrow \mathbb{R}$ that gives a numeric value with every action $a \in A$. An action a^* becomes an optimal action if:

$$\pi(a^*) \geq \pi(a) \quad \forall a \in A \quad (4.1)$$

Where the optimal decision is to choose $a^* \in A$, that maximizes the payoff. In practice, two actions may lead to the same maximal payoff, therefore either will represent the optimal decision. The outcome of the game is a set of appealing elements that the modeller picks from the values of actions, payoffs and other variables after the game is played out.

The *pure strategy* of a player is the movement plan for the game instructing in advance what the player will do in reaction to every event. It is said a player's outcome is a choice if the player selects a strategy without knowing the strategy of the other players. On the other hand, the player knows that players follow a pure strategy when they know about the other strategies of the other players. When players know all the information in a game, i.e. their own strategies and payoff functions and those of the others, it is called a game with *complete information*. Whenever a player knows the rules of a game and their own choices, but not the payoff functions of the other player, this is called a game with *incomplete information*. It is called "A game of perfect information" when players know how the other players move or game strategies and that could influence the result of his own choice. "A game of imperfect information" is a game where players sometimes do not know the move that other players have made, either because those choices are made simultaneous or they are hidden.

A *zero-sum game* [56] is widely used when two players are considered. This is a mathematical representation of a game where a player's gain or loss is balanced by the losses or gains of the other players. If the total gains are added up and the total

losses are subtracted, they would sum to zero. On the other hand, mixed-motive games [57] are extensively used in interpersonal decision-making. In matrix form, two people choose between two alternatives, a cooperative or competitive act and though each person makes his choice separately, both choices jointly determine the payoff to each subject. Thus, the game is designed such that each person's payoff is not only dependent upon his own choice, but also upon the choice of the other person.

4.2.3 Nash Equilibrium based decision making

Something that has always been a source of curiosity is what action will be chosen by the players in a strategic game? In general, it has been assumed in a static game that each player chooses the best available action [58]. Addressing this dilemma, John F. Nash [59], has introduced the following strategy which is called the theory of Nash equilibrium. Since then, the concept of Nash Equilibrium has become a major topic in Game Theory, economics and other social sciences. Here, each player chooses the action according to the model of rational behaviour, given by the player's belief about the other player's action. If every player's belief about the other player's action is correct, it will form the Nash equilibrium.

A Nash Equilibrium is an action profile a^* with the property that no player i can do better by choosing an action different from a_i^* , given that every other player j adheres to a_j^* . In other words, neither player could do better by adopting another strategy when the strategy adopted by the other player is given. The Nash Equilibrium for a two player game is a pair of actions (a_i^*, a_j^*) such that:

$$\pi_1(a_i^*, a_j^*) \geq \pi_1(a_i, a_j^*) \quad \forall a_i \in A_1 \quad (4.2)$$

and

$$\pi_2(a_i^*, a_j^*) \geq \pi_2(a_i^*, a_j) \quad \forall a_j \in A_2 \quad (4.3)$$

4. PAYOFF CALCULATION AND DECISION MAKING

It is clear from Equation 4.2 and 4.3 that, even though the Nash Equilibrium might not include strictly dominated strategies, it may include weakly dominated ones. In game theory, a player's strategy will determine the action the player will be taking at any stage of the game. Thus, in this context, the strategy S_i is strictly dominated by strategy S_j if, for every choice of strategies of the other players the payoff from choosing S_j is strictly greater than that obtained by S_i . If that is the case, the strategy S_i is the dominated strategy and S_j the dominant strategy. Moreover, the strategy S_i is weakly dominated by strategy S_j if, for every choice of strategies of the other players, the payoff from choosing S_j is greater or equal to the payoff from S_i .

In the literature, some experimental work [60] supports the concept that agents in repeated games do learn to form Nash equilibrium, however, there is no theoretical explanation that is given for this phenomenon. In practice, the action of a player depends on the other player's action and whenever choosing an action, the player takes into consideration the actions that the opposing players would choose. This makes the players believe in the other player's action in gaming. This belief can derive from the past experience in playing the game and their experience is sufficiently extensive that a player knows how the opponents will behave. No one tells a player the action that the opponent will choose, but a players previous involvement in gaming leads players to be sure of these actions [60].

4.2.4 The Problem of multiple Equilibria

Every game that has a finite strategy forms at least one Nash equilibrium. However, some games have multiple equilibriums and that leads them to have more than one possible solution. From a mathematical point of view, this multiplicity of equilibria is a problem when we want one answer, not a family of answers. And many economists would also regard it as a problem that has to be solved by some restriction of the assumptions that would rule out the multiple equilibria [49]. But, from a social scientific point of view, there is another interpretation. Many social

scientists believe that coordination problems are quite real and important aspects of human social life. From this point of view, we might say that multiple Nash equilibria provide us with a possible "explanation" of coordination problems. That would be an important positive finding, not a problem [49]. However, the existence of multiple equilibria illustrates a common difficulty for modellers in practice. In such cases, modellers add more details to the rules of the game or use an equilibrium refinement, adding conditions to the basic equilibrium concept until only one strategy profile satisfies the refined equilibrium concept. There is no single way to refine Nash equilibria and therefore modellers should insist on a strong equilibrium, rule out weakly dominated strategies or use iterated dominance [13].

4.3 Applying the Nash Equilibrium based Decision Making in Image Annotation

This section will introduce the use of the Nash Equilibrium based decision making concept for the proposed game based framework. To use Nash's concept, two payoff functions are proposed. In this game, the payoff of Player 1 plays the main role and is always measured based on the historical data of the player's performances in image annotation. In order to do so, initially the framework feeds players with a number of fully annotated images; it analyzes the player comment in order to measure player confidence, thus, the transition probabilities. This is been done by using a Markovian model [15]. The two states of the Markov Model (MM) are: a "correct" and an "incorrect" tag or annotation is entered, and they are represented by the variable C and I , respectively. Here, the player outcomes for fully annotated images are sequentially ordered and segmented into sets of tags for the purpose of calculating conditional probabilities in the transition matrix. For an example, the probability of $P(C_{t+1}|I_t)$ is estimated by dividing the number of sets in which the label 'correct' occurs after 'incorrect' by the total number of tag sets containing 'incorrect'. According to [19], human behaviour is governed by the current intentions, rather than being based on the previous experiences. However,

4. PAYOFF CALCULATION AND DECISION MAKING

given that current player intentions cannot be anticipated, they are best approximated by assuming that the current outcome is dependent on his previous outcome, i.e. based on the correct or incorrect annotation the player has entered. In Figure 4.1, an overview of the segmenting process is given.

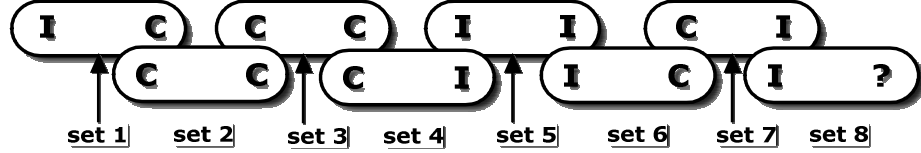


Figure 4.1: Segmenting player's outcome into set of tags.

Given by the player outcome at step t on preceding multimedia content, the probabilistic outcome at step $t + 1$ is estimated by using the transition matrix M . This matrix gives the change of behaviours of players in the Markovian chain.

$$M = \begin{bmatrix} P(C_{t+1}|C_t) & P(I_{t+1}|C_t) \\ P(C_{t+1}|I_t) & P(I_{t+1}|I_t) \end{bmatrix}$$

where, $P(C_{t+1}|C_t)$ denotes the probability of obtaining a correct annotation at step $t + 1$, when the player is given a correct annotation at step t . Similarly, other probabilities $P(C_{t+1}|I_t)$, $P(I_{t+1}|C_t)$ and $P(I_{t+1}|I_t)$ are measured using players historical data, i.e. using segmented outcomes. A diagrammatic overview of the proposed MM is given in Figure 4.2.

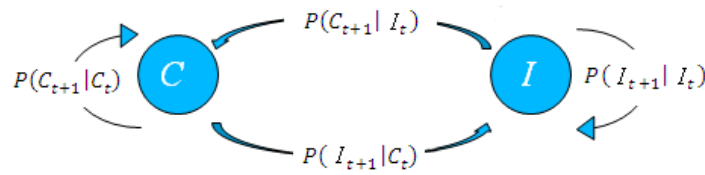


Figure 4.2: Player's probability distribution in gaming.

The initial idea is to measure player performances in image annotation. To do so, two payoff functions are implemented, where it measures good (G_i) and bad (B_i) contributions of the players. Here, Player 1's good contribution is measured by (4.4)

4. PAYOFF CALCULATION AND DECISION MAKING

using the MM and its associated transition probabilities.

$$P(G_1) = (P(C_{t+1}|C_t)P(C_t) + P(C_{t+1}|I_t)P(I_t)) \quad (4.4)$$

where $P(C_{t+1}|C_t)P(C_t)$ is the overall probability of obtaining a correct annotation at $t + 1$, when the state ‘correct’ considered, and $P(C_{t+1}|I_t)P(I_t)$ is the overall probability of obtaining a correct annotation at step $t + 1$, when the state ‘incorrect’ considered. Here, $P(C_t)$ is the player’s overall probability of entering a correct annotation, which is measured by dividing the number of correct annotations given by the player by the number of fully annotated contents the player has been exposed to. Similarly, $P(I_t)$ is the overall probability of entering an incorrect annotation, measured by dividing the number of incorrect annotations given for fully annotated contents by the number of fully annotated contents provided to the player. Player 1’s bad contribution is measured by (4.5) using the MM and its associated transition probabilities.

$$P(B_1) = (P(I_{t+1}|C_t)P(C_t) + P(I_{t+1}|I_t)P(I_t)) \quad (4.5)$$

where $(I_{t+1}|C_t)P(C_t)$ is the probability of having incorrect annotations at $t + 1$ when the state ‘correct’ considered and $P(I_{t+1}|I_t)P(I_t)$ is the probability of having incorrect annotations at $t + 1$ when the state ‘incorrect’ considered. When Player 2 is considered, $P(G_2)$ in gaming is estimated as follows:

$$P(G_2) = (\pi_1(a_1, a_2) + P(K) + P(F)) / k_1 \quad (4.6)$$

$$k_1 = \begin{cases} 3, & \text{if } P(K) \text{ and } P(F) \text{ available} \\ 2, & \text{if } P(K) \text{ or } P(F) \text{ available} \\ 1, & \text{otherwise} \end{cases}$$

where $\pi_1(a_1, a_2)$ is the payoff of Player 1 in gaming, which is calculated by subtracting bad contributions ($P(B_1)$) from the good ones ($P(G_1)$), see Equation 4.8 for more information; $P(K)$ is the probability of entering a given annotation, which is calculated by dividing the number of given annotations that are similar to player’s

4. PAYOFF CALCULATION AND DECISION MAKING

input keyword by the total number of annotations obtained by the image; $P(F)$ is the outcome of low-level feature classification that indicates that the probability of an image is relevant to the trained concept (see Section 3.2.1 for a justification on using $\pi_1(a_1, a_2)$, $P(K)$ and $P(F)$ for calculating Player 2's contribution); and k_1 is the normalising constant that defines the availability of $P(K)$ and $P(F)$. In practice, classification outcomes are not entirely accurate and should thus only be used when greater than the threshold of the F-measure of the given concept. When considering Player 2 in gaming, $P(B_2)$ is estimated as:

$$P(B_2) = N * T_1 \quad (4.7)$$

where N is the number of dissimilar annotations that has been assigned to an image. Thus, if framework performs well in game play, the annotations entered by players inclined to cheat would be identified and rejected. Consequently, the number of dissimilar annotations assigned to an image would be smaller. This logic confirms the suitability of using the number of dissimilar annotations to calculate cost of Player 2 in gaming. Here, T_1 is the allocated cost per annotation, which is used for limiting the maximum number of dissimilar annotations per image. Since players are asked to annotate images based on the main object or character, and it can be assumed most of the rationally minded players will do so, only a few dissimilar annotations will be obtained for an image. Consequently, restricting the number of dissimilar annotations that can be assigned to an image will not result in any performance loss of the system.

Whenever non-annotated or partially annotated content is exposed, the profiles of actions are estimated as follows. Let's assume that the action of player i taken at each round is to be a_i . Action a_1 indicates that annotations of Player 1 are good or bad in a game round and is observed by the outcome prediction unit. Here, a_1 is assigned 1 whenever the prediction unit says the player is trustworthy and would enter a correct annotation. Therefore, it can be assumed that this process represents the short term contribution of the player in gaming.

4. PAYOFF CALCULATION AND DECISION MAKING

$$a_1 = \begin{cases} 1, & \text{if prediction says the player will enter a good annotation} \\ 0, & \text{otherwise} \end{cases}$$

Similarly, a_2 is Player 2's action property and is calculated using a threshold score. In practice, the outcome of the score computation module is used to measure Player 1's total score achieved before the onset of ongoing game round (m_1) (a detailed explanation about score calculation is given under the heading 'Score Computation', later in this section). Thus, when m_1 is less or equal to a certain threshold score ($m_1 \leq \text{threshold score}$), action a_2 is assigned the value of 0. Similarly, it is equal to 1 when the m_1 is greater than the threshold score ($m_1 > \text{threshold score}$). Although Player 1 increases his total score by submitting 'correct' annotations, the framework keeps the record of the difference in game points between the m_1 and the threshold score, defined as T_2 . Thus, whenever a player cheats, his total score will be reduced based on the calculation performed by the score computation module, while keeping the threshold score unchanged. In addition, whenever m_1 falls below the threshold score, the threshold score will be kept unchanged until m_1 improves sufficiently and becomes greater than the threshold score with a lead of T_2 . Therefore, it can be assumed that this process represents the long term contribution of the player in gaming.

$$a_2 = \begin{cases} 1, & \text{if } m_1 > \text{threshold score} \\ 0, & \text{otherwise} \end{cases}$$

For each round, given all the information including the action profile (a_1, a_2) , a general function for calculating Player 1 and 2's payoff can be defined as follows:

Payoff of Player 1:

$$\pi_1(a_1, a_2) = a_1 P(G_1) - a_2 P(B_1) \quad (4.8)$$

Payoff of Player 2:

$$\pi_2(a_1, a_2) = a_2 P(G_2) - a_1 P(B_2) \quad (4.9)$$

4. PAYOFF CALCULATION AND DECISION MAKING

The above payoff functions consisted of two terms. First term, i.e. $a_1 P(G_1)$ and $a_2 P(G_2)$ denotes the gain of good contribution of the players in respect to their action. The second term $a_2 P(B_1)$ and $a_1 P(B_2)$ demonstrates the cost or bad contribution in gaming with respect to interaction of the opponent player. One can say that it is not fair to measure the cost of the player based on the opponent player's action. However, it is fair to use action a_1 over Player 2 and vice versa because the Player 2 in this game is not a fully independent player of Player 1.

Here, (4.8) and (4.9) are slightly modified when analysing fully annotated contents. In practice, Player 2 is well aware of all fully annotated contents and their associated metadata. Hence, the player is capable of correctly examining fully annotated contents. As a consequence, $P(G_2)$ of Player 2 is assigned to 1 and $P(B_1)$ is assigned to 0. Also, action property a_2 is given a value of 1.

For each round, given all the information including the action profile (a_1, a_2) , both player payoffs are calculated as follows:

Player 1 payoff:

$$\pi_1(a_1, a_2) = a_1 P(G_1) - a_2 P(B_1) \quad (4.10)$$

Player 2 payoff:

$$\pi_2(a_1, a_2) = 1 \quad (4.11)$$

4.3.1 Nash Equilibrium representation

Nash Equilibrium is a solution concept of a game involving two or more game players in which each player is assumed to know equilibrium strategies of the other players, i.e. the users strategic profile where every player is unilaterally optimum, in the sense that no player is willing to change its own strategy as this would cause a performance loss [61]. Due to the nature of this game, there exists an infinite number of equilibriums. In terms of accuracy, not all produce correct annotations.

4. PAYOFF CALCULATION AND DECISION MAKING

Table 4.1: Truth table for all possible actions

a_1	a_2	Player 1 payoff $\pi_1(a_1, a_2)$ - Human	Player 2 payoff $\pi_2(a_1, a_2)$ - Machine
0	0	0	0
0	1	$-P(B_1)$	$P(G_2)$
1	1	$P(G_1) - P(B_1)$	$P(G_2) - P(B_2)$
1	0	$P(G_1)$	$-P(B_2)$

Here, Table 4.1 shows the player's outcome for all possible strategic actions. It shows the feasible region is inside a convex hull of: 1. $(0, 0)$, 2. $(-P(B_1), P(G_2))$, 3. $(P(G_1) - P(B_1), P(G_2) - P(B_2))$, 4. $(P(G_1), -P(B_2))$. In Figure 4.3, a graphical representation of all possible equilibriums is shown. However, the shape of this graph could be varying according to the changers of the internal variables of the payoff functions.

Nash Equilibrium refinement

Presupposing that the players are rational, from (4.8) it is reasonable to assume that $\pi_1(a_1, a_2) > 0$, because players do not play games unless they obtain positive game points.

Table 4.2: Payoff representation for all actions.

Actions		Player 2's long term contribution	
		Long term bad ($a_2 = 0$)	Long term good ($a_2 = 1$)
Player 1's short term contribution	Short term bad ($a_1 = 0$)	(0,0)	(-, +)
	Short term good ($a_1 = 1$)	(+, -)	(+, +)

If players are cooperative, Table 4.2 shows that action pair *Short good*, *Long good* forms the unique Nash equilibrium. Given that the framework chooses action “*Long good*”, players are better off choosing “*Short good*” than “*Short bad*” as that

4. PAYOFF CALCULATION AND DECISION MAKING

can significantly increases player payoff (from the right column of the table it shows *Short good* yields a positive payoff where *Short bad* yields a negative payoff). Given that Player 1 chooses “*Short good*”, Player 2 is better off choosing *Long good* than *Long bad* (from the bottom row of the table we see *Long good* yields a positive payoff where *Long bad* yields negative). The action profile *Short bad, Long bad* is not a unique Nash equilibrium, because if Player 2 chooses the action *Long bad*, player payoff to *Short good* exceeded the payoff to *Short bad* (the first components of the entries in the left column of the table). The action profile *Short good, Long bad* is not a unique Nash equilibrium; this is because if Player 1 chooses the action *Short good*, payoff of Player 2 for action *Long good* exceeds the payoff to action *Long bad* (second components of the entries of the bottom row of the table). The action profile (*Short bad, Long good*) is also not a unique Nash equilibrium. Because if Player 2 chooses *Long good*, Player 1 payoff to *Short good* exceeds the payoff to *Short bad* (first component of the entries in the right column of the table).

According to the above theory, rational players may decide to work together in order to maximise their payoffs. However, as in this game, Player 2 is not active, thus this strategy is not applicable. Nonetheless, given that action profile *Short good, Long good* forms a unique Nash Equilibrium, it is fair to accept a player keyword as a valid annotation when players meet this condition, and $P(G_1) > P(B_1)$, i.e. $\pi_1(a_1, a_2) > 0$. This logic is valid, as whenever $P(G_1) > P(B_1)$, the probability of entering a valid annotation by Player 1 increases. However, new keywords, i.e. a keyword new to the image, will be accepted only when $P(G_2) > P(B_2)$, i.e. whenever $\pi_2(a_1, a_2) > 0$. Consequently, whenever $P(B_2)$ increases, high good contribution levels $P(G_2)$ are expected from the player, which will increase the probability of obtaining a valid keyword, leading to annotations that are more accurate. However, when $P(G_2) \leq P(B_2)$, i.e. $\pi_2(a_1, a_2) \leq 0$, a keyword will be accepted as a valid annotation only if it has been previously described by other players. This, again, is a valid logic, as it has been previously established that the probability of entering an incorrect similar annotation by two malicious players for

4. PAYOFF CALCULATION AND DECISION MAKING

a particular image is very low. However, there is a high probability that the rationally minded players would enter an annotation described before by some other player. Therefore, whenever $P(G_2) \leq P(B_2)$ occurs, a keyword is accepted as correct only if it has been described before by some other player.

In practice, the framework measures action profile (a_1, a_2) prior to exposing the contents. Therefore, it exposes players to non-annotated or partially annotated images only when both actions are equal to 1 and $P(G_1) > P(B_1)$, i.e. $\pi_1(a_1, a_2) > 0$. More formally, whenever action property $a_1 = a_2 = 1$, players will be exposed to non-annotated images when the overall probability of entering a correct annotation ($P(C_t)$) is less than a given threshold T_3 , i.e. $P(C_t) < T_3$. Moreover, player will be exposed to a partially annotated content whenever $P(C_t) \geq T_3$. In practice, as the framework only accepts annotations when $a_1 = a_2 = 1$, the probability of accepting an incorrect annotation by the framework can be assumed to be low. Thus, under these conditions, exposing a non-annotated or a partially annotated content would not affect the outcome of annotations. However, given that partially annotated images contain annotations from previous game plays, they already assign a cost to Player 2 (see Equation 4.7). In fact, when players are exposed to a partially annotated content, there is a risk that an annotation could be rejected by the framework whenever $P(G_2) < P(B_2)$ occurs. To address this problem, players are exposed to partially annotated images when $P(C_t)$ is greater or equal than a given threshold T_3 , i.e. $P(C_t) \geq T_3$, which minimises the risk of rejecting a correct annotation by the framework.

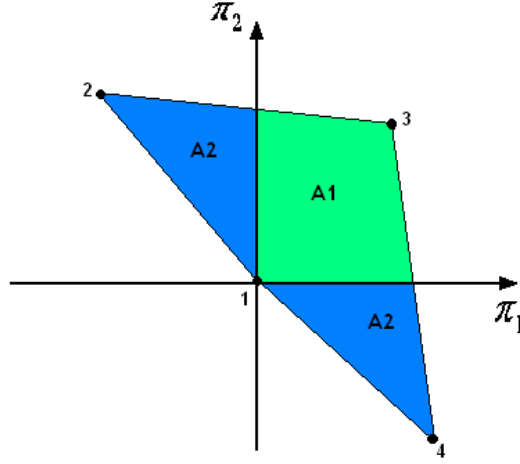


Figure 4.3: Payoff outcome representation.

Figure 4.3, shows the feasible and enforceable regions of the proposed game. Here, A1 is the enforceable region in this game, where it represents most of the rational players. Region A2 is the area that most of the less-rational players are located in this game. In practice, the outline of this graph can vary based on the following variables, $P(G_1)$, $P(B_1)$, $P(G_2)$ and $P(B_2)$.

Score Computation

The main purpose of the score computation unit is to reward players for their contribution in gaming, thus, to yield game points. Additionally, this game uses the score computation algorithm for another purpose, which is to measure the action property of Player 2 (a_2).

$$\text{Player } i\text{'s total score} = m_i + (\text{Player } i\text{'s payoff} * 100) \quad (4.12)$$

where m_i is the player i 's total score achieved before the onset of the ongoing game round. In practice, when Player 1 enters a keyword, the total score will be calculated based on the payoff functions and the outcome will be displayed as a reward for his contribution in gaming. Once the scores are displayed, they will be used as the m_i scores for the next game round. In fact, the framework will be able

to measure the action property a_2 before exposing any new image to the player.

4.4 Weighting player 2's Payoff by Image Classification

Machine learning aims to understand the fundamental principles of learning on a particular problem as a computational process. The area of machine learning deals with the design of tools that can learn from observed data, adopt changes and improve performance with experience. Typically machine learning has become an essential tool that is expected to solve increasingly complex problems. In this section, an introduction to machine learning is briefly introduced. In the literature, due to the existence of complex problems, an extensive research work for developing many efficient learning algorithms has been carried out. In this thesis, we used image classification techniques just to weigh the player's contribution in gaming. In order not to lose the generality, a brief summary on existing categories of machine learning is given in this section.

Image classification algorithms help to invent ways to classify data into meaningful categories. Thus, it is widely used for the purpose of image indexing and retrieving. Classification enhances not only the accuracy in indexing and retrieving, but also the speed. Therefore, a large image dataset can be organised according to the classification rule, within a short amount of time. Typically, image classification relies on either low-level features or heuristic structures [62]. In general, there are two types of classification schemes, supervised and unsupervised classification. In supervised classification, some degree of human attention is required to make a correction in classification. On the other hand, unsupervised classification does not need human attention. In unsupervised, the main goal is to partition a given set of data into groups. Namely, this process is called clustering, where the data points in a cluster are more similar to each other than the points in different clusters (this can be still regarded as a problem of one of the learning

4. PAYOFF CALCULATION AND DECISION MAKING

factions).

In 1956, Artificial Intelligence (AI) began its days. The initial idea was to make a machine behave like a human being. Since then, the research on machine learning eventually grew as a sub-field of AI. In 1958, Rosenblatt proposed the first ever learning machine which is called the Perceptron [63]. This method uses a weighted sum of inputs followed by a threshold binary output where the weights could be adjusted to learn different tasks. However, it had its limitations being that it could learn certain non linear mapping. Addressing this problem in 1974, Werbes proposed an algorithm for learning weights in a multi-layer network, also called neural networks (NN). Since then, NN went on to be successful and has been used for learning representations, classification and regression mappings in many applied domains [64]. In recent years, NN has undergone many extensions [65]. Since, NN originated, similar concepts in statistics also introduced additional extensions. Like AI, statistics were also concerned with tasks such as estimating models from observations. One of the important key features in statistics is the Bayes rule [66]. Bayesians are statistics that use probabilities for measuring prior beliefs. Generative modelling, (a model for generating random observable data), is often Bayesian and that uses the Bayes rule extensively. Typically, Bayes are different from frequentists that only use probabilities from frequencies of observable data.

One key development of the 1990's was the popularization of generalisation bounds on learning machines. This brought both applied and theoretical interest to classifiers and complexity tools such as the Vapnik-Chervonenkis (VC) dimension [67]. The VC- dimensions are broadly used in statistical learning to guarantee the generalisation. This motivated the large margin decision boundaries and the support vector machines were introduced. Since then, it is been widely used in many domains for data classification.

SVM has attracted a lot of interest because of its unique features, such as the capability of dealing with high-dimensional input feature vectors. Because SVM is well documented and predefined executables are available, we used SVM to weight

Player 2's payoff in the proposed game (See Appendix A for an extended description concerning SVM's). Here, we used the Acemedia⁷ toolbox to extract low-level features from images. For classification purposes, we used the LIBSVM executable tool [68] with default parameters. The classification process here has been completed offline to reduce the computation power required for the game. The linear combination of three low-level descriptors colour layout (CLD) [69], dominant colour (DCD) [70] and edge histogram (EHD) [71] descriptors are used for image classification. These descriptors were merged through a fusion way to combine a number of descriptors into a single feature vector as in Equation 4.20. As a consequence, it obtains high performances in image classification [72]. The fusion strategy used here is called the Merging Fusion Method [73]. A detailed description of Merging Fusion Method is given in Section 4.4.2.

4.4.1 Analysis of Low-level Features

Most of the image indexing and annotation frameworks use visual features to obtain high accuracy in image annotation. The visual features widely used are, colour, texture, shape and spatial relationship between objects. In this thesis, the proposed approach uses some visual features for image classification; more specifically colour and texture features. Without losing generality, a brief summary on existing categories is given in this section.

Colour Descriptors

Colour is one of the most important features that can be easily identifiable in visual content. A significant amount of research has been conducted on various aspects of the colour feature which MPEG-7 has standardized a subset of these approaches to form a number of different colour descriptors. In this section of the

⁷ AceMedia (www.acemedia.org) is a collaborative research project from the European Union Sixth Framework Program, in the area of multimedia semantic analysis and processing.

4. PAYOFF CALCULATION AND DECISION MAKING

thesis, most commonly used colour descriptors are reported. Here, before defining colour features, a brief description on the colour space has been described.

Colour Spaces

There are various colour spaces introduced in the literature. Depending on the application, various colour space models are distinguished for different applications. Colour space is a method which creates and visualizes colours. In general, humans define a colour by its attributes of brightness hue and colourfulness. A colour is usually specified using three coordinates, or parameters. These parameters describe the position of the colour within the colour space being used.

RGB is an additive colour space based on tri-chromatic theory often found in electronic devices with CRT display images. Some commonly used colour spaces in literature are HSV, RGB, CMY, HSL, YIQ, YUV, YCbcr etc. The HSV colour space considers human intuition and addresses three of the most important aspects in the perception of the colour hue, saturation and value. Since every space model has its advantage, uniformity is the main requirement for image indexing and retrieval systems.

Colour Histogram

The Colour Histogram is a representation of the distribution of colours in an image. It represents the number of pixels with colour values that fall into given colour ranges; more often a specific colour range is called the colour bin. These bins are defined based on the colour space and quantization levels of the colour. Colour histograms are good representatives of colour distributions across an image, however they lack spatial colour information and to address this issue local colour descriptors, such colour layout or region-based descriptors, have been developed.

Texture Descriptors

Texture is an important factor in visual perception and discrimination of image content. Texture feature has been extensively studied in the research area of image segmentation, image classification, and image retrieval and in other pattern analysis fields. Texture feature characterizes image texture or regions, observing the region homogeneity. Many approaches have been proposed for texture based image retrieval using the multi resolution techniques such as the Wavelet Transform [74]. There are a number of texture descriptors proposed in the literature. The best-established kind relies on comparing values of what are known as second order statistics, calculated from query and stored images. These approaches extract textures by calculating the relative brightness of selected pixel pairs from each image [75]. In [76], a number of texture features are introduced, in particular coarseness, contrast, directionality, regularity, line-likeness and roughness. Among them, the first three are commonly used to extract the texture information. Moreover, the Gabor filter based multi-resolution representation [77] and Grey-Level co-occurrence Matrix (GLCM) [76] are used to extract more texture information in images.

Shape Descriptors

Shape feature provides a powerful clue to object identity and as a consequence it has been used in a similarity search and retrieval of objects. Humans can recognize characteristic objects solely from their shapes. This proves that shape feature is a powerful feature that provides semantic information. This property distinguishes shape from other visual features such as colour or texture.

The image and video world usually deals with 2-D projections of real world objects, where MPEG-7 provides tools to describe 2-D shapes. Generally, shape representations can be classified types: contour-based and region based. The contour-based method expresses shape properties of an object based on its outline. This boundary information may not be available in some cases due to the occlusion,

4. PAYOFF CALCULATION AND DECISION MAKING

noise and vagueness that may occur in digital world. The contour-based shape descriptors are based on the Curvature Scale-space (CSS) representation of the contour and were proposed in [78]. This descriptor is very efficient in applications where high variability in the shape is expected and is robust to noise present in the contour.

The second type, region-based shape descriptors, do not necessarily rely on shape boundary information because they rely on all pixels representing the shape not only on the contour pixels. The Zernike Moment Descriptor [79] is one robust shape descriptor which is invariant to rotation, robustness to noise, expression efficiency and multilevel representation for describing the various shapes of patterns.

MPEG-7 Features Space

MPEG-7 standard is defined by the Moving Picture Expert Group (MPEG) as a standard multimedia content description interface for offering a set of audio-visual descriptions in an effort to provide standardized tools for describing multiple content [80]. MPEG-7 standardizes visual content such as colour descriptions, textual descriptions, shape descriptions, motion descriptions and face descriptions. MPEG-7 defines colour descriptions [80], such as the Dominant Colour Descriptor (DCD) [81] characterize an image or image region of a small number of dominant colour values and some statistical properties related to these. Scalable colour is a colour histogram with efficient encoding based on the Haar Transform [82]. The Colour Structure [83] is an extension of the colour histogram that incorporates some associated structural information. Colour Layout Descriptor [69] describes the spatial layout of colour within an image. Finally, Group of Frames/ Group of Pictures colour is an extension of scalable colour to an image Sequence/collection. MPEG-7 present three descriptors to extract textures featuring of a visual content [80], namely, Homogeneous Texture Descriptor (HTD) [84], Edge Histogram Descriptor (EHD) [71] and Perceptual Browsing Descriptor (PBD) [85]. Here, HTD and EHD describe the statistical distribution of the texture feature of an image and

4. PAYOFF CALCULATION AND DECISION MAKING

are useful for image retrieval applications. PBD is a compact descriptor suitable for quick browsing applications. MPEG 7 describes a number of shape descriptors, Region-based descriptor, contour based descriptor and 2D/3D shape descriptor. The visual descriptors used for experiments in this thesis are presented in detail in this section.

Dominant Colour Descriptor

The Dominant Colour Descriptor (DCD) specifies the representative colours in an image or in an image region. Similarity retrieval in image databases and browsing of image databases based on colour values are the main targets in applications of this descriptor. These colours are computed and quantized for each image or image region. The DCD can be represented with the following vector.

$$\text{DCD} = \{(c_i, p_i, v_i), s\}, \quad i = 1, 2, \dots, N, \quad (4.13)$$

where N is the number of dominant colours, which varies from one image to another and c_i is the i^{th} dominant colour. Each dominant colour c_i represents the colour value vector corresponding to the colour values of the corresponding image. In addition, p_i is the percentage of pixels for the i^{th} dominant colour in the image or image region; v_i is an optional field that expresses the variance describing variation of colour values for pixels in a cluster of a particular colour; and s represents the spatial coherency of the image, i.e. the homogeneity of dominant colours. The spatial coherency is a single number that represents the overall spatial homogeneity of the dominant colours in an image. As it describes the spatial distribution of pixels associated with each representative colour, high values imply that pixels of similar colours are co-located. Consequently, searching for individual colours can be performed efficiently using a 3-D colour space, which thus allows for fast and convenient similarity matching.

Consider two DCD's (F_1 and F_2),

$$F_1 = \{(c_{1i}, p_{1i}, v_{1i}), s_1\}, \quad i = 1, 2, \dots, N_1, \text{ and}$$

4. PAYOFF CALCULATION AND DECISION MAKING

$$F_2 = \{(c_{2j}, p_{2j}, v_{2j}), s_2\}, \quad j = 1, 2, \dots, N_2,$$

The matching function measures the dissimilarity of two descriptors and is given as follows which ignores the optional variance parameter.

$$D^2(F_1, F_2) = \sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2a_{1i,2j} p_{1i} p_{2j} \quad (4.14)$$

where, subscripts 1 and 2 stand for descriptors F_1 and F_2 in all variables, respectively. Moreover, N_1 and N_2 are, respectively the number of dominant colours in descriptor F_1 and F_2 ; $a_{1i,2j}$ is the similarity coefficient between colour clusters c_{1i} and c_{2j} ; p_{1i} is the percentage of pixels for i^{th} dominant colour in the colour cluster c_{1i} ; and p_{2j} is the percentage of pixels for j^{th} dominant colour in the colour cluster c_{2j} . The similarity coefficient $a_{1i,2j}$ between two colours c_{1i} and c_{2j} is defined as follows:

$$a_{1i,2j} = \begin{cases} 1 - d_{1i,2j}/d_{\max}, & d_{1i,2j} \leq \tau_d \\ 0, & d_{1i,2j} > \tau_d \end{cases}$$

where, $d_{1i,2j} = ||c_{1i} - c_{2j}||$ is the Euclidean distance between the colours c_{1i} and c_{2j} . τ_d is the maximal distance for two colours to be considered similar and $d_{\max} = \alpha \tau_d$. This means that any two dominant colours from one single description are at least τ_d distance apart. A recommended value for τ_d is given in [86] as 10 to 20 in the CIE – LUV colour space and for α is between 1.0 and 1.5.

Colour Layout Descriptor

The Colour Layout Descriptor (CLD) is a very compact and resolution-invariant representation of colour for high-speed image retrieval. This descriptor is widely used in variety of similarity based retrieval applications and specially used for spatial structure-based retrieval applications such as sketch based retrieval and video segment identification purposes [86].

4. PAYOFF CALCULATION AND DECISION MAKING

This descriptor divides the impact image into 64 blocks to achieve the resolution or scale invariance and calculates the average colour of the pixels in each block. Then it applies the Discrete Cosine Transform (DCT) on each of the three components in the $YCrCb$ colour space and three sets of 64 DCT coefficients are obtained, which are later zigzaggedly scanned and the first few coefficients are nonlinearly quantized.

$$CLD = \{DY_j, DCr_j, DCb_j\} \quad j = 1, 2, \dots, m \quad (4.15)$$

where DY_j represents the j^{th} DCT coefficient of the Y colour component; DCr_j represents the j^{th} DCT coefficient of the Cr color component; and DCb_j represents the j^{th} DCT coefficient of the Cb color component. Here, m (the maximum number of DCT coefficients) is defined by the user.

For matching two CLDs, e.g. $\{DY, DCr, DCb\}$ and $\{DY', DCr', DCb'\}$, the following distance measure can be used.

$$D = \sqrt{\sum_j w_{yj}(DY_j - DY'_j)^2} + \sqrt{\sum_j w_{rj}(DCr_j - DCr'_j)^2} + \sqrt{\sum_j w_{bj}(DCb_j - DCb'_j)^2} \quad (4.16)$$

where DY_j , DCr_j and DCb_j are the j^{th} coefficients of the Y, Cr and Cb colour components, respectively. Here, w_{yj} , w_{rj} and w_{bj} are the weighting coefficients, which are decreased according to the zigzag scan order. This descriptor is designed to assign greater weights to lower frequency components for the characteristic matching.

Edge Histogram Descriptor

EHD finds the spatial distribution of edges in an image and is used as a strong texture descriptor for similarity search and retrieval. This descriptor divides an

4. PAYOFF CALCULATION AND DECISION MAKING

image into 4×4 subimages and then calculates the local edge distribution for each subimage by a histogram. This histogram contains information about 5 edge categories, vertical, horizontal, diagonal 45 degrees, diagonal 135 degrees and non-directional.

$$\text{EHD} = \{h(l), h^g(m), h^s(n)\} \quad l = 0 \dots 79, m = 0 \dots 4, n = 0 \dots 64, \quad (4.17)$$

where $h(l)$ represents the normalised histogram bin value of the bin count (l) for the local histograms of the given image; $h^g(m)$ represents the normalised bin value of the bin count (m) for the global-edge histograms of the image, which is obtained from the corresponding local histograms $h(l)$. Similarly, $h^s(n)$ represents the histogram bin values for the semi-global edge histograms of the image. For similarity matching, local (80 bins), semi global (65 bins) and global (5 bins) edge histograms are considered in the similarity function [86].

$$\begin{aligned} D(A, B) = & \sum_{k=0}^{79} |h_A(k) - h_B(k)| + 5 \sum_{k=0}^4 |h_A^g(k) - h_B^g(k)| \\ & + \sum_{k=0}^{64} |h_A^s(k) - h_B^s(k)| \end{aligned} \quad (4.18)$$

where $h_A(k)$ and $h_B(k)$ represent the normalised histogram bin values of the bin count (k) of images A and B , respectively. Furthermore, in line with the above, $h_A^g(k)$ and $h_B^g(k)$ represent the normalized bin values of the bin count (k) for the global-edge histograms of the images A and B , respectively, which are obtained from the corresponding local histograms $h_A(k)$ and $h_B(k)$. Finally, $h_A^s(k)$ and $h_B^s(k)$ represent the histogram bin values for the semi-global edge histograms of images A and B , respectively. Since the number of bins of the global histogram is smaller relative to that of local and semi-global histograms, a weighting factor 5 is applied in Equation 4.18.

4.4.2 Fusing MPEG-7 Visual Descriptors for Image Classification

This section introduces the Merging Fusion Method used in the proposed framework. Empirical evidence suggests that, in order to capture the particular properties of each image, it is crucial to select an appropriate set of visual descriptors. This is one of the issues that frequently arise in image classification and thus negatively affect the system performance. In order to address this problem, a technique referred to as Merging Fusion Method is proposed in [73]. This technique applies merge fusion to combine a number of different low-level descriptors, thus improving the image classification performance by reducing the uncertainty and ambiguity of features. In this technique, all the visual descriptors are merged into a unique vector, known as the D_{merged} before the classification process is carried out. Consequently, high image classification performance can be achieved [72]. The merge descriptor is formed based on the following sequence. Let D_1, D_2, \dots, D_M and M be descriptors represented in a vector form. Here, the merged descriptor is formed as follows:

$$D_{merged} = \{D_1, D_2, \dots, D_M\} \quad (4.19)$$

In the framework developed as a part of this study, DCD, CLD and EHD descriptors are used to construct the merged descriptor, as follows:

$$D_{merged} = \{D_{DCD}, D_{CLD}, D_{EHD}\}$$

Given DCD, CLD, and EHD in (4.13), (4.15) and (4.17), respectively, the merged descriptor can be defined as:

$$D_{merged} = [\{(c_i, p_i, v_i), s\}, \{DY_j, DCr_j, DCb_j\}, \{h(l), h^g(m), h^s(n)\}] \quad (4.20)$$

4.5 Summary

In conclusion, the payoff functions described in this chapter are designed to aggregate the player's contribution, previously recorded players' contribution and image classification outcomes in order to obtain useful annotations. In addition, these payoff functions are implemented based on the player's interaction, as well as the use of Game Theories and strategies. The Nash Equilibrium-based decision model forces the agents to behave in a rational manner, thus yielding a decision to the complex problem of image annotation. Nash Equilibrium strategy is simple, yet very successful when it is applied to competitive environments and is proven well suited for multiplayer game models. Although the approach presented here focuses on a single player gaming mode, some simple techniques are being used to adapt this game into a multiplayer model, thus making the Nash Equilibrium based techniques suitable to apply over single player games. In the annotation problem, the NE-based decision model allows decisions to be made based on the player's short- and long-term performance in image annotation. Based on NE's decision, the player is exposed to the most suitable image, i.e. fully annotated, partially annotated or non-annotated. Consequently, the accuracy in image annotation can be significantly improved.

Chapter 5

PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

5.1 Introduction

An important characteristic of a prediction algorithm is the ability to learn from previous experience in order to predict the future outcomes. The need for learning the process has led to vast amounts of research into the construction of prediction algorithms. Typically, prediction of human behaviour is the most difficult task to achieve in practice. The reason for this arises with the human behaviour, which is random and dynamic. This dynamic behaviour has led researchers to predict human outcomes using sequential decision making theories [87] [88]. Sequential decision making involves selecting a sequence of actions to accomplish a goal; that is the prediction of sequential outcomes [87]. In prediction, the objective is to select or predict an action from a finite set of possible actions. When all possible actions correspond to a set of possible outcomes given, the problem that arises is to find the

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

best possible action. Typically, the optimal solution for this case is to choose the action that gives the maximum expected outcome. However, to maximise the expected outcome, the predictive distribution for all the possible outcomes need to be determined. In many cases, the probability distribution is not known explicitly and therefore, it is estimated by sampling the previous obtained data.

In practice, algorithms developed for pattern recognition are widely used in the prediction of new events. One classical application used in sequential prediction is the lossless compression [89]. Application wise, it is widely used in speech and language modelling [90], text-writing recognition [91], and biological sequence analysis [92]. In lossless compression, Hidden Markov Models (HMM) are widely used to predict the state of an outcome [93]. It is flexible in structure and possible to model a complex source of sequential data. Also, HMM's can be adapted to a framework easily. However, an HMM needs a large number of training datasets to produce considerably accurate prediction [94] and that is the main drawback of using HMM. In [95] authors proposed prediction based on Context Tree Weighting (CTW) which is a lossless compression and prediction algorithm that is widely used for prediction. The role of CTW is to combine a number of variable order Markov Models, which can model sequential data of considerable complexity. Another widely used sequential prediction approach is the prediction by partial matching (PPM) [96]. Instead of generating the probability distribution entirely based on the longest sequence match, PMM is designed to blend the predictions of multiple context lengths and assigns a higher weight to longer matches [96].

Although there are a number of prediction algorithms that have been developed, no generic model has yet been developed to predict human outcomes. Since human outcomes are random, dynamic and may not follow a repeated sequence, it makes it even harder to predict. Since, according to [19], human behaviour is mostly governed by current intentions, rather than being based on past performances, predicting human outcomes based on present behaviour may yield promising results. Addressing this dilemma, we focus the literature review on prediction

algorithms based on Markov models, which has been highly used in predicting outcomes based on the present intention. To compare and evaluate the Markov prediction approach, we also focused the literature review on sampling algorithms, in particular Sequential Sampling, where the prediction and decision making is influenced by examining the entire distribution, not only based on the present outcome. Unlike the Markov approach, SS is well known and the involved risk of accepting a defective sample is what makes it admired when compared to the Markov approaches.

5.2 Prediction by Markov Chains

Most of the study of probability has dealt with independent trials of processes. These are the fundamentals of well-known probability theory and statistics. Typically, when a sequence of chance experiments forms an independent trial, the potential outcomes for each experiment may occur the same and with the same probability outcome. Here, the information regarding previous experiments does not influence the prediction of the next experiment. Outcomes of these types of experiment are generally measured by using a single experiment and by constructing a tree that represents the probability distribution. By measuring the tree for a sequence of n experiments, it is possible to answer any probability questions.

In 1907, Markov started the study of a new type of chance process. In this process, the outcome of a given experiment can affect the outcome of the new experiment and is called the Markov property [97], which is the characterization of a system that transits from one state to another. It is concerned with the random process with the Markov property. This process is a Markov model, for a particular type of Markov process in which the process can only be in a finite or countable number of states. Markov decision processes are widely used in many areas. This includes computer science (for predicting memory references) [98], predicting sequential events [99], predicting dynamically changing environments [16] etc.

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

As the theory of Markov chains is well documented, only a short introduction to the topic and some of their basic properties that are used for constructing prediction mechanisms are presented in this thesis. In [100] detailed descriptions about Markov chains are given.

For a discrete time Markov chain (DTMC), the observations of states are done in a discrete set of times. When consider a stochastic process, i.e. a sequence of random variables $\{S_t : t \in 0, 1, 2, \dots\}$ taking discrete values in the state space $\{0, \dots, J-1\}$, are called Markov chains if given the current state of the process S_t , the future S_{t+1} is independent of its past $S_{t-1}, S_{t-2}, \dots, S_0$. For clarification purposes, let's assume that s_0, \dots, s_t, s_{t+1} denote a sequence of observations of a stochastic process $\{S_t, t = 0, 1, \dots\}$. Here, $\{S_t\}$ is a Markov process if it satisfies the Markov property, namely

$$P(S_{t+1} = s_{t+1} | \underbrace{S_t = s_t, S_{t-1} = s_{t-1}, \dots, S_0 = s_0}_{\text{"entire history"}}) = P(S_{t+1} = s_{t+1} | S_t = s_t) \quad (5.1)$$

for all $t \in \{0, 1, \dots\}$.

A Markov chain with stationary transition probabilities $p_{ij} = P(S_{t+1} = j | S_t = i)$ is called homogeneous if the transition probabilities are independent of t . The transition probabilities of a homogeneous J -state Markov chain can be summarized in a $J \times J$ transition probability matrix (TPM).

$$M = \begin{pmatrix} p_{00} & \cdots & p_{0J-1} \\ \vdots & \ddots & \vdots \\ p_{J-10} & \cdots & p_{J-1J-1} \end{pmatrix} \quad (5.2)$$

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

The elements of the transition Matrix M satisfy the following two properties:

$$0 \leq p_{ij} \leq 1 \quad (5.3)$$

for all i ,

$$\sum_{j=0}^{J-1} p_{ij} = 1, \quad i \in \{0, \dots, J-1\} \quad (5.4)$$

The above TPM contains one-step transition probabilities, where it describes the short-term behaviour of the Markov chain. For describing the long-term behaviour of a Markov chain, k -step transition probabilities $p_{ij}(k) := P(S_{t+k} = j \mid S_t = i)$ are defined. It can be shown that the matrix $M(k)$, which contains the k -step transition probabilities can be calculated as the K^{th} power of the transition probability matrix M . That is,

$$M(k) := \begin{pmatrix} p_{00}(k) & \cdots & p_{0J-1}(k) \\ \vdots & \ddots & \vdots \\ p_{J-10}(k) & \cdots & p_{J-1J-1}(k) \end{pmatrix} = M^k \quad (5.5)$$

The proof about K^{th} power transition probabilities is given in [100]. The k -step transition probabilities provide the conditional probabilities to be in state j at time $t+k$, given that the Markov chain is in state i at time t . However, in general, the marginal probability of the Markov chain to be in state i at a given time t is also of interest (this is dependent on the goal of the prediction). Given the probability distribution for the initial state $\pi := (P(S_1 = 1), \dots, P(S_1 = m))$ with $\sum_{i=1}^m \pi_i = 1$, (where π_i is probability distribution of the state i), the distribution of the state at time t can be computed as in (5.6).

$$(P(S_t = 0), \dots, P(S_t = J-1)) = \pi M^{t-1} \quad (5.6)$$

A diagrammatic overview of a typical MM is illustrated in Figure 5.1.

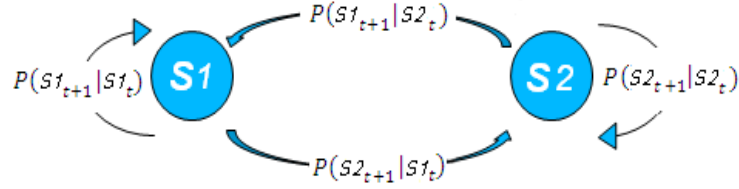


Figure 5.1: A diagrammatic overview of a Markov Model.

5.2.1 Hidden Markov Models

Hidden Markov Models have become an extensively used tool when modelling random process. Moreover, it is widely used in applications, such as speech recognition, radar telecommunications, financial mathematics [101] etc. A typical HMM is characterised by two stochastic processors, an observed process and an unobserved (hidden) process. In a typical MM, the state is directly visible to the observer. However, in HMM, the state is not directly visible but the state dependent output is visible. In HMM, there is a probability distribution in each state, i.e. the probability distributions from each state to possible observations are known. Here, a basic structure of a HMM is illustrated in Figure 5.2.

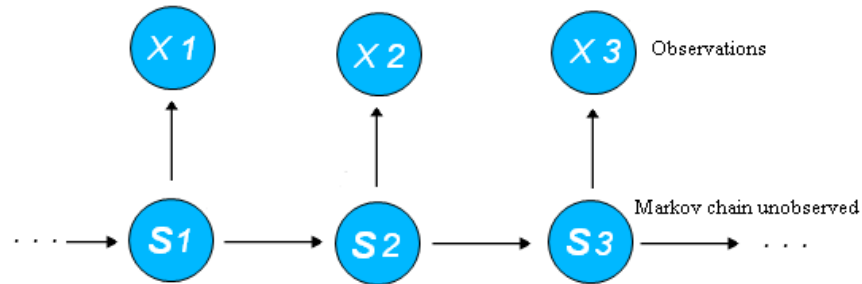


Figure 5.2: A basic structure of a HMM.

A typical HMM characterised with a Markov chain S_t which determines the state at time t , and a state dependent process X_t , which illustrates the observations. The state sequence is governed by a $J \times J$ matrix of transition probabilities of the

form $P(S_{t+1} = j \mid S_t = i)$, that is the conditional probability that the system transit to state j given that the system is in state i at time t . The probability distribution associated with each state describes how the observed data are distributed when the system is in state j .

When the HMM and state transitions are given, the goal is to find the most likely sequence of hidden states. This is normally achieved by taking the advantage of the independent structure of the HMM. Generally, Forward-backward [102] procedure is used to find the most likely sequence of the state using the observations.

5.3 Decision making by Sequential Sampling

Sampling is an important technique in many fields for developing efficient randomized algorithms. A task such as estimating the proportion of instances with a certain property in a given dataset can often be achieved by randomly sampling a relatively small number of instances or so called samples. In general, large industries use sampling plans to measure the quality in product manufacturing due to either ruining the products or the volume of products being too large. The sample size is a very important factor when large sizes of bounds have been used. In practice, the Chernoff bound [103] and Hoeffding bound [104] have been used widely because they derive a theoretical guaranteed sample size sufficient for achieving a given task with given accuracy and confidence [105]. However, there are some cases that bounds can provide us with only over estimated or unrealistic sampling sizes. In practice, Sequential Sampling algorithms are used for some of such cases to design adaptive randomized algorithms with theoretically guaranteed performance [105].

In this thesis, we use Sequential Sampling techniques to predict a player's behaviour. However, without losing the generality, a brief summary on existing

categories of product sampling is given in this section.

5.3.1 Methods used in product sampling

In the literature, a number of sampling methods have been introduced for the purpose of quality checking [106] [107]. Some of them are: no checking, 100% checking, constant percentage sampling, random spot checking, audit sampling (no acceptance and rejection criteria) and acceptance sampling based on probability.

No checking is used in product inspection when the process capability is known and the probability of a defective product is very small. However, most of the manufacturers that use no sampling will check product quality periodically to verify that conditions have not changed.

100% checking is used when it is necessary to check all the products, such as in cases where lives are involved. However, looking at each sample is expensive and time consuming.

Constant percentage sampling is widely used when the number of samples is big. This process will inspect a given percentage of products from a lot. It seems to be more efficient, but the problem with this method is that the sample taken from small lots may be too small and the sample taken from large lots may be too large. As a consequent, accuracy for small lots may not be achieved and too much time and effort may be spent on large lots, therefore the sampling risk involved is not known.

Random spot checking may sometimes be used when the manufacturing process is certified as providing excellent quality products. Therefore, random check is used to verify that the process is in control and to report the product quality level. However, the sampling risk in this process is not known and as a result, this method will not guarantee that the outgoing quality will be at an acceptable level.

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

Audit sampling is a sampling process that is done on a routine basis. Here, the acceptance criterion is not known, however a quality report will be issued to the manufacturing organization to determine what action is to be taken in product manufacturing. Likewise, with random spot-checking, audit sampling is often used when the process of manufacturing is certified as providing excellent quality products.

Acceptance sampling

Acceptance sampling based on probability is the most widely used sampling technique in practice. Most of the acceptance sampling for inspection by attributes are pre-constructed and published and can be easily used by anyone [108] [109]. Typically, acceptance sampling guarantees the performance in theory. When inspection is performed by classifying products good or defective, a number of types of acceptance sampling plans have introduced in the literature. Namely they are, lot by lot - single sampling, lot by lot - double sampling, continuous sampling and Sequential Sampling.

Lot by Lot Single Sampling [110] uses a sample size n selected randomly from a lot size N for quality inspection. Here, a lot will be accepted if the numbers of defects or defectives in the sample do not exceed the acceptance number. Similarly, a lot will be rejected if the numbers of defects or defectives in the sample exceed the acceptance number c . The rejected lots may be re-inspected for the verification purposes of the quality in the inspection process.

Lot by Lot Double Sampling [110] uses two sample sizes n_1 , n_2 and two acceptance numbers c_1, c_2 are specified by the quality inspector. If the number of defects or defectives in the first sample size n_1 exceeds c_2 , the lot will be rejected. If the number of defects in the first sample size n_1 do not exceeds c_1 , the lot will be accepted. When the number of defects in the first sample are greater than c_1 but less than or equal to c_2 , a second sample n_2 will be inspected. If the second sample is inspected and defectives in the combined first and second sample do not exceed c_2 ,

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

the lot is accepted. If defects or defective in the combined samples exceeds c_2 , the lot is rejected.

Continuous Sampling [111] is used when the product flow is continuous and not possible to form into lots. Here, two parameters are specified to form a continuous sampling plan. First is the frequency f and the second is the clearing number i . The frequency is expressed as $1/10$, $1/20$, $1/X$, etc, and i is assigned a number. Here, samples will be 100% checked in the beginning until i parts are found to be defect free. It inspects one out of X samples. This process will continue until a defect sample is found. When a defect sample is found, 100% inspection will resume.

Sequential Sampling [112] is different from single, double or multiple sampling. It classifies a sequence of samples or one sample as good or defective by analysing and checking for specified requirements. When the sequence is one sample at a time, the sampling process is usually called *item-by-item* Sequential Sampling. However, one can also select large sample sizes greater than one, in which case the process is referred to as group Sequential Sampling. Item-by-item is more popular in practice. The advantage of this type of sampling plan is that a decision could be made based on a relatively small sample size.

Sequential Sampling plans make decisions by counting the conforming and nonconforming units. The counted outcomes are compared against the decision criteria to make a decision. Often, counts are graphically represented with accept and reject lines drawn on a graph. In practice the counted result make the decision when the sequential plot crosses one of the lines. If the plotted point falls within parallel lines (acceptance and rejection lines), the process continues by inspecting another sample. As soon as a point falls on or above the upper line (rejection line), the lot is rejected. Also, when a point falls on or below lower line (acceptance line), the lot is accepted. An example of a Sequential Sampling plan is shown by Figure 5.3. It shows the numbers of defectives increases with the observed samples. Here,

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

for each point X-axis represents the total number of items that are selected and Y-axis represents the total number of observed defectives.

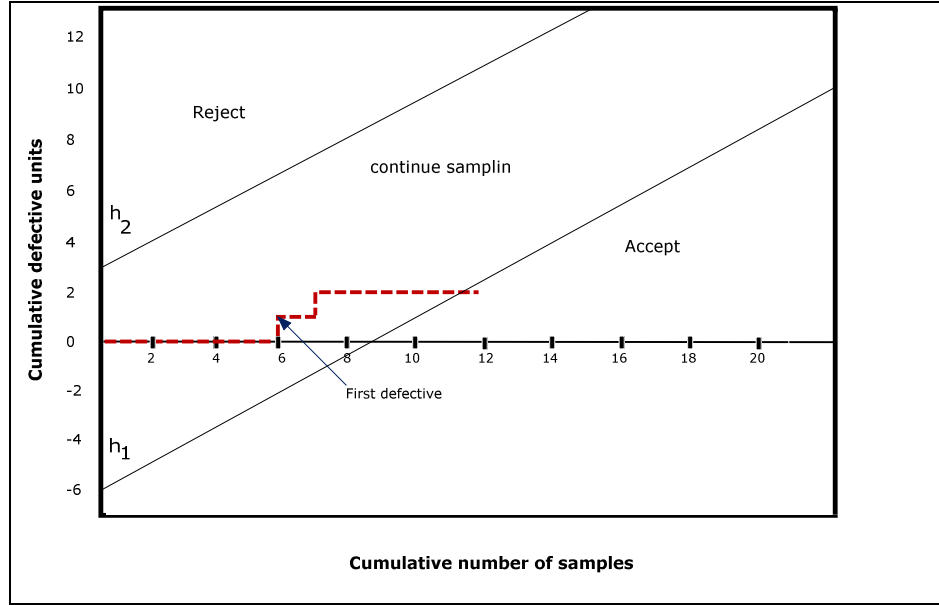


Figure 5.3: An example of Sequential Sampling plan.

5.3.2 Sequential Probability Ratio Test

In [113], Abraham Wald has introduced a new procedure for sequential data analysis known as the Sequential Probability Ratio Test (SPRT) procedure. The SPRT is a specific sequential hypothesis test, which gives a specific rule at any stage of the experiment for making one of the following three decisions: (1) to accept the hypothesis being tested (null hypothesis), (2) to reject the null hypothesis, (3) to continue the experiment by making additional observations. Thus, such a test is carried out sequentially, as described below.

In [114], one of the well-known lemmas proposed by Neyman and Pearson is given. Here, the authors have provided a method of constructing a most powerful test for a simple versus simple hypothesis-testing problem. The process can be explained by assuming that X has a Probability Density Function (PDF) $f(x; P)$ and

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

$H_1: P = P_1$ needs to be tested against $H_2: P = P_2$.

Let $X_1, X_2, X_3, \dots, X_n$ be the sample points (n independent observation points) and

$$\Lambda_n = \frac{\prod_{i=1}^n f(X_i, P_2)}{\prod_{i=1}^n f(X_i, P_1)}$$

Then the most powerful test of H_1 against H_2 is obtained by rejecting H_1 if $\Lambda_n \geq K$, and accepting H_1 if $\Lambda_n < K$, where K is a nonnegative constant.

Based on Neyman and Pearson's lemma given above, Wald proposed the following sequential probability ratio test. In this test, the rule for terminating the experimental procedure is a simple threshold scheme that uses two constants A and B , such that $0 < B < A$, and postulates that H_1 should be accepted if $\Lambda_n \leq B$; or rejected if $\Lambda_n \geq A$. Finally, sampling should be continued if $B < \Lambda_n < A$, when the experiment inspected n samples. Here, the constants A and B are chosen so that sequential test has the desired value α of the probability of a type I error and the desired value β of the probability of a type II error [115]. For Wald's SPRT, A and B were chosen based on the characteristics of the type I error and type II errors as follows (see [113] for a detailed description about SPRT and calculation of A and B).

$$A \cong \frac{1 - \beta}{\alpha} \text{ and } B \cong \frac{\beta}{1 - \alpha}$$

Considering Binomial distribution [116], the SPRT for $H_1: P = P_1$ and $H_2: P = P_2$ is defined using the above mentioned two constants A and B as follows. Here, after n observations, the sampling will continue if

$$B < \frac{P_2^m (1 - P_2)^{n-m}}{P_1^m (1 - P_1)^{n-m}} < A$$

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

that is, if

$$B \left(\frac{1 - P_1}{1 - P_2} \right)^n < \left[\frac{P_2(1 - P_1)}{P_1(1 - P_2)} \right]^m < A \left(\frac{1 - P_1}{1 - P_2} \right)^n$$

or

$$\frac{n \log \left(\frac{1 - P_1}{1 - P_2} \right) + \log B}{\log \left[\frac{P_2(1 - P_1)}{P_1(1 - P_2)} \right]} < m < \frac{n \log \left(\frac{1 - P_1}{1 - P_2} \right) + \log A}{\log \left[\frac{P_2(1 - P_1)}{P_1(1 - P_2)} \right]}$$

In other words, the inequality on which a decision to continue sampling is made of the form

$$Sn + h_1 < m < Sn + h_2$$

where m denotes the number of defective data points and S, h_1 and h_2 are functions of A, B, P_1 and P_2 .

A simplified version of the above decision process determines the system as reliable if m falls below the acceptance line (5.7), and as unreliable if m falls above the rejection line (5.8).

$$\text{Acceptance line: } Y_a = -h_1 + Sn \quad (5.7)$$

$$\text{Rejection line: } Y_r = h_2 + Sn \quad (5.8)$$

In Sequential Sampling, P_1 and P_2 are always given to a system by the designer or the creator of the sampling plan. Typically, P_1 and P_2 symbolise the Acceptance Quality Level (AQL) and the Rejectable Quality Level (RQL) in the Sequential Sampling plan, respectively. Here, AQL represents the quality that is routinely accepted by the sampling plan, and RQL defines the number of defective samples that the sampling plan can tolerate. In Sequential Sampling, the desired value α of

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

the probability of a type I error represents the risk of rejecting a correct unit by the sampling plan, and the desired value β of the probability of a type II error represents the risk of accepting an incorrect unit by the sampling plan. Those risks are measured using an operating characteristic curve (OC) [117], and are discussed in Section 5.3.3. Applying Wald's Sequential Probability Ratio Test, the origin of the acceptance line is computed as:

$$h_1 = \frac{\log\left(\frac{1-\alpha}{\beta}\right)}{k_2} \quad (5.9)$$

Similarly, the origin of rejection line is computed as:

$$h_2 = \frac{\log\left(\frac{1-\beta}{\alpha}\right)}{k_2} \quad (5.10)$$

Finally, the line slope is computed as:

$$S = \frac{\log\left(\frac{1-P_1}{1-P_2}\right)}{k_2} \quad (5.11)$$

where,

$$k_2 = \log\left[\frac{P_2(1-P_1)}{P_1(1-P_2)}\right] \quad (5.12)$$

5.3.3 OC curve and probability distribution

The operating characteristic curve represents the picture of a sampling plan. It describes the probability of acceptance of a lot as a function of its quality [117]. In Sequential Sampling, the OC curve is used to measure the risk of rejecting a correct unit (α) and the risk of accepting an incorrect unit (β). Here, α and β is measured based on the corresponding probability of acceptance of the AQL and RQL,

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

respectively (see Figure 5.4). The OC curve in Figure 5.4 shows that when the percentage defectives in a lot increase, the probability of acceptance decreases. The idea is that the consumer will be accepting a lot of products as long as the process percent defective is below a given level. Each sample plan has a unique OC curve, sample size and acceptance number. It defines the OC curve and determines its shape.

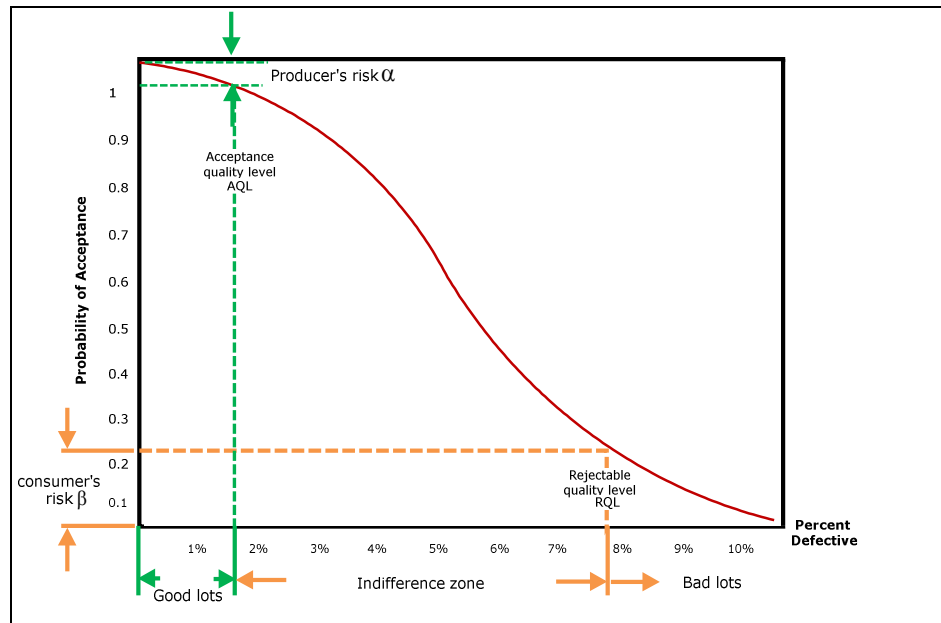


Figure 5.4: Operating characteristic curve.

There are number of probability distribution methods which are introduced in the literature. Some of the widely used distributions are the Hypergeometric distribution, the Binomial distribution and the Poisson distribution.

Hypergeometric Distribution

The Hypergeometric distribution [118] is a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement. This distribution is mostly used when the lot size is very small.

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

The probability of exactly x defective parts in a sample n :

$$P(X) = \frac{\binom{n}{x} \binom{N-n}{n-x}}{\binom{N}{n}} \quad (5.13)$$

where N is the number of total population.

Binomial Distribution

The Binomial distribution [116], also called the Bernoulli experiment or Bernoulli trial, is a discrete probability distribution of the number of successes in a sequence of n independent outcomes. This distribution is sometimes described as sampling with replacement although the parts are not physically replaced. When event $n = 1$, the binomial distribution is a Bernoulli distribution. This distribution is widely used in the literature for the constructing of sampling plans [119] [120].

The probability of exactly x defective parts in a sample n :

$$P(X) = \binom{n}{x} p_{oc}^x (1 - p_{oc})^{n-x} \quad (5.14)$$

where, p_{oc} represents the probability of having defectives of the incoming quality.

Poisson Distribution

The Poisson distribution [121] is so called Poissonian, which is used in sampling plans when the number of defects or defects per units is important, not the number of defective parts. The Poisson distribution can be easily applied to systems with a large number of possible events.

The probability of exactly x defective parts in a sample n :

$$P(X) = \frac{e^{-np} (np)^x}{X!} \quad (5.15)$$

where, e is the constant 2.71828.

5.4 Applying the Markov based and Sequential Sampling based Prediction Techniques

There are players with different attitudes, from the very rational to the very malicious. Rational players mostly produce correct annotations. Thus, their outcomes usually are valid metadata. On the other hand, malicious players try to cheat by entering misleading or incorrect annotations while still trying to achieve high scores in the game. In practice, it is difficult to always correctly distinguish all players. Taking this quandary into consideration, a comparative study on two prediction mechanisms is undertaken. Finding the most responsible material in outcome prediction is what the key-idea of these experiments is.

5.4.1 Player's outcome prediction by Markovian based inference

In [19] authors have suggested that the best way to predict human outcome is by using the present intention. Considering this fact, we used MM based prediction techniques to predict player outcomes. In this approach, initially, the framework feeds players with a number of fully annotated images; it analyzes the player comment in order to measure player confidence, thus, the transition probabilities. This has been done by using a Markovian model [15]. The player outcomes for fully annotated images are sequentially ordered and segmented into set of tags for the purpose of calculating conditional probabilities in the transition matrix (see Section 4.3).

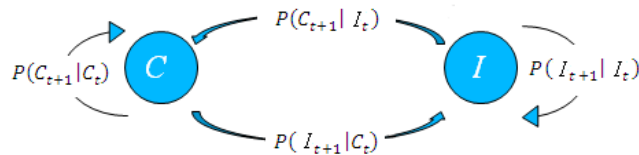


Figure 5.5: Player's probability distribution in gaming.

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

In Figure 5.5, a diagrammatic overview of the proposed MM is given. Since players do not know as to what type of content that they are exposed to, it is sensible to assume that they respond in the same way for any of the three types of content: fully annotated, partially annotated or non-annotated. Based on this assumption, player outcomes are predicted as follows: for example, let's assume that the player gives an 'incorrect' annotation at step t , this can be represented by which the 'correct' entry is 0% and the 'incorrect' entry is 100%, i.e., $x_t = [0 \ 1]$. Since the player's future outcome depends only on the current outcome x_t , the player's future outcome x_{t+1} is predicted as follows:

$$x_{t+1} = x_t M = [0 \ 1] \begin{bmatrix} P(C_{t+1}|C_t) & P(I_{t+1}|C_t) \\ P(C_{t+1}|I_t) & P(I_{t+1}|I_t) \end{bmatrix} = [P(C_{t+1}|I_t) \ P(I_{t+1}|I_t)] \quad (5.16)$$

It is assumed that whenever $(C_{t+1}|I_t) > P(I_{t+1}|I_t)$, the player's potential to provide a correct annotation is high. As a result, the action property a_1 is assigned the value of 1, i.e. $a_1 = 1$. Similarly, whenever $(C_{t+1}|I_t) \leq P(I_{t+1}|I_t)$, it is assumed that the player's potential to provide an incorrect annotation is high, and consequently, the action property a_1 is assigned the value of 0, i.e. $a_1 = 0$.

5.4.2 Player's outcome prediction by Sequential Sampling

To compare and evaluate the Markov prediction approach, we also focused on sampling algorithms, in particular Sequential Sampling, where the prediction and decision making is influenced by examining the entire distribution, not only based on the present behaviour. When considering the acceptance sampling, the risk of accepting an incorrect annotation is well known in advance.

The proposed framework's design has been based on a Sequential Sampling plan, where it uses an Operating Characteristic curve to demonstrate the performances of the player in image annotation. This is the distribution that precisely shows X incorrect annotations in n number of images using a Binomial

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

distribution. The reason for using Binomial distribution is that it describes the sampling with replacement although the units (incorrect annotations) are not physically replaced. This makes the draw of each unit independent, hence it fairly represents the player's performances in gaming. Here, the initial idea is to find the risk that player faces in the game (α), i.e. rejecting of a correct annotation given by the player and the risk that system faces in this game (β), i.e. the accepting of a wrong annotation given by the player, using the OC curve.

$$P(X) = \binom{n}{x} p_{oc}^x (1 - p_{oc})^{n-x} \quad (5.17)$$

where, p_{oc} represents the probability of having incorrect annotations of the incoming quality. In Sequential Sampling, Acceptance Quality Level represents the quality that routinely is accepted by the sampling plan. In our case, it is the measure of incorrect annotations that the system is willing to accept. In the proposed sampling plan, AQL is measured based on the quality of a dictionary mechanism in detecting valid key-words. In practice, there is a risk of rejecting a valid keyword by the dictionary whenever existing keywords are not detected, i.e. a keyword existing in the English language, thus, AQL is a level of product quality that is used by the system.

$$AQL = \left(\frac{N_i}{N_p} \right) \quad (5.18)$$

where N_i is the number of incorrect identifications made by the dictionary mechanism and N_p is the number of key-words inspected by the dictionary mechanism, which exist in the English language.

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

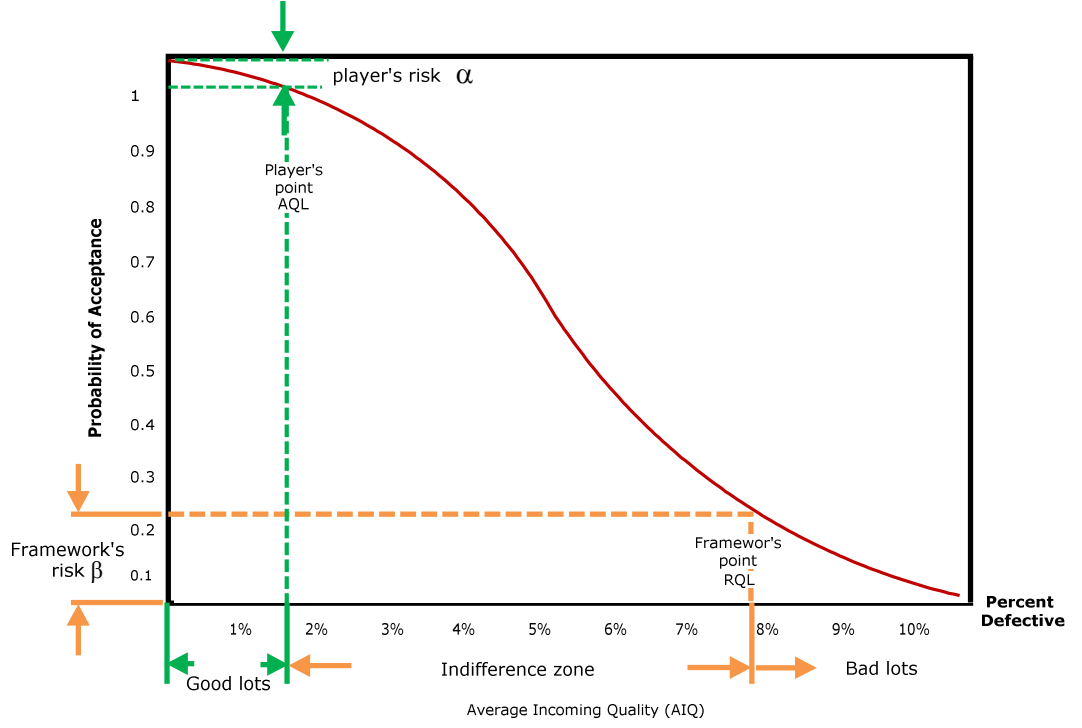


Figure 5.6: Operating characteristic curve.

From the player's point of view, rejecting a valid keyword is the risk that players face in this game and is denoted by α in the proposed sampling plan. This risk is measured by the OC curve and its associated AQL parameter as shown in the Figure 5.6. The Rejectable Quality Level defines the number of defective annotations that are willing to be tolerated by the sampling plan. In the proposed sampling plan, it is the framework's point of rejecting the player's outcome.

$$RQL = \left(\frac{N_w}{N_a} \right) \quad (5.19)$$

where N_w is the number of wrong annotations that are given by the player and N_a is the number of fully annotated contents exposed to the player. The probability of accepting a wrong annotation is the risk that the framework faces here and is denoted by β in the sampling plan. In Figure 5.6, the proposed OC curve is shown. It is often updated using outcomes of the player for fully annotated contents thus

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

this curve represent the latest behaviour of the player in the image annotation.

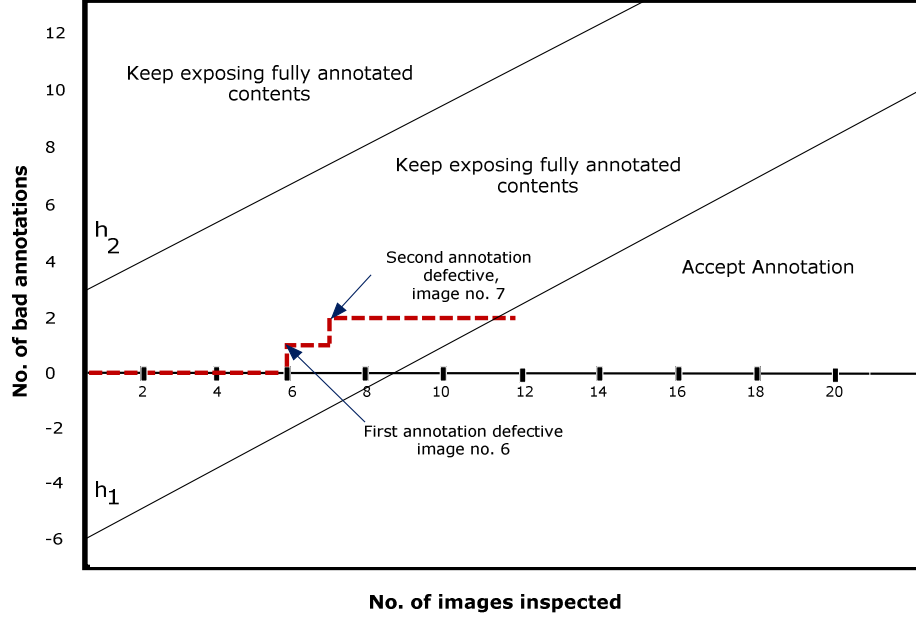


Figure 5.7: Proposed Sequential Sampling plan.

In the proposed approach, item-by-item Sequential Sampling is used. Here, hitting or crossing a line results in making decisions [122]. When given a set of quality levels, AQL, α , RQL, and β , the acceptance (Y_a) and rejection (Y_r) lines are computed as follows:

$$\text{Acceptance line: } Y_a = -h_1 + Sn \quad (5.20)$$

$$\text{Rejection line: } Y_r = h_2 + Sn \quad (5.21)$$

where h_1 is the origin of acceptance line; h_2 is origin of the rejection line; S is the slope of each line and n is the number of inspected samples. Here, we calculated h_1 , h_2 and S by using Equations (5.9) (5.10) and (5.11) respectively.

The increasing numbers of defectives are plotted in Figure 5.7. For each point, X-axis represents the total number of annotations that are selected and Y-axis represents the total number of observed defectives, i.e. wrong annotations. The

proposed prediction mechanism works as follows: Prior to exposing non-annotated content, first the number of defective annotations and exposed fully annotated contents are increased by one. This simulates the worst case in this game, which is of having a wrong annotation. Secondly, the OC curve and related parameters, such as RQL, β , α , Y_a , Y_r and plotted point in the sampling plan is updated. In this instance, if the plotted point falls on or below the lower line, i.e. acceptance line, it is assumed that the player's potential to provide a correct annotation is high. Consequently, the action property a_1 is assigned the value of 1, i.e. $a_1 = 1$. Similarly, the plotted point falls above the lower line it is assumed that the player's potential to provide an incorrect annotation is high, and consequently, the action property a_1 is assigned the value of 0, i.e. $a_1 = 0$.

In practice, action properties a_1 and a_2 are always measured prior to presenting any images to the player. Since empirical evidence suggests that, when the action properties are equal to 1, i.e. $a_1 = a_2 = 1$, the optimal solution in the game is reached, if one of these actions is 0, the player will be exposed to a fully annotated content. Alternatively, a partially annotated or a non-annotated content will be presented to the player, based on the player's overall probability of entering a correct annotation ($P(C_t)$), as described in Section 4.3.1.

5.5 Summary

This chapter introduces the player prediction unit developed as a part of this study. As, in practice, it is difficult to always correctly distinguish all players, a comparative study on two prediction mechanisms was undertaken, one based on Markov chains and the other based on Sequential Sampling. The former predicts player outcome based on the present intention, whereas the latter examines the entire distribution, rather than the present outcome only. However, unlike the Markov approach, SS significantly reduces the risk of accepting a defective sample, making it a superior choice in the context of this study. Since players are not

5. PLAYER'S OUTCOME PREDICTION AND DECISION MAKING

informed in advance what type of content they will be exposed to, it is sensible to assume that they would respond consistently to any of the three types of content fully annotated, partially annotated or non-annotated. This assumption makes using the player's outcome for fully annotated contents a suitable choice of information for predicting the player's outcome. The proposed prediction mechanisms are introduced to obtain more accurate annotations and thus the player's behaviour is predicted prior to exposing non- or partially annotated contents. That makes NE less dependent in decision-making. A comprehensive evaluation on these techniques is reported in the next chapter.

Chapter 6

EXPERIMENTAL EVALUATION

In this section, the proposed game-based annotating framework is comparatively evaluated using three different real image datasets. The first section evaluates the performance of the image classification, while the second section evaluates the usability of the game, in particular excitement, addiction, enjoyment and difficulty in the game play. These factors are compared with the most popular games, ESP and Phetch. Next, the proposed framework efficiency is evaluated, followed by its precision and selected representative results. In the final set of experiments, the precision for different configurations of the framework are evaluated, followed by selected representative results.

6.1 Introduction

In this section, we shall evaluate the proposed algorithm using real world image datasets. For evaluation purposes, two interfaces were introduced, namely INT-1 and INT-2 (see Section 3.3.1) and evaluated with two well known games, ESP and Phetch. In the first set of experiments, we evaluated the proposed SVM classifier for a number of concepts. Secondly we evaluated the excitement factor, addiction,

6. EXPERIMENTAL EVALUATION

enjoyment and the difficulty of playing these games. In the next set of experiments, we evaluated the average number of annotations collected by each game per minute and then, a one-to-one comparison was undertaken to compare their precisions with real world image databases. Here, three different real image datasets were used for evaluation, namely, the ESP image dataset, the Caltech image dataset and the Corel image dataset. On each dataset, we compared our approach of image annotation based on Game Theory with MM prediction (denoted by IA-GTMM), alongside the annotation approach based on Game Theory with Sequential Sampling prediction (denoted by IA-GTSS). The obtained results have been compared with a manually created ground truth of semantic scenes and objects that appeared in the images. In the final set of experiments, the precision of the proposed framework for different configurations is evaluated, i.e. IA-GTMM, IA-GTSS and a framework that uses only the two-player game model (a game that use no prediction mechanisms) are compared.

To evaluate the proposed framework, we first exposed the players to a number of fully annotated images via the visual interface (players are asked to annotate the main object or the character). Outcomes from these images were used for predicting the player's outcome. If the outcome indicates that the player is cheating, he/she will be exposed to fully annotated content. In the other hand, if the outcome indicates that the player is honest, Nash Equilibrium based equilibrium analysis is used. Whenever Nash's equilibrium is formed, players will be exposed to a non-annotated or a partially annotated content and that is based on the outcome of T_3 threshold, or else the player will be exposed to a fully annotated content. No matter what image the player is exposed to, they are encouraged to comment on it using a string of characters (the game accepts only one string at a time). This string will be then passed onto the text analysis unit, where it ensures that the player has entered a valid keyword by using its inbuilt spellchecker software. In the case of the player entering an invalid keyword, he/she will be asked to re-enter a valid keyword.

6. EXPERIMENTAL EVALUATION

The text analysis module uses a huge database of words that consists of more than 100,000 words from the English language. It measures the Hamming distance between the player's input string and database words to make clear that the player is giving a valid keyword. The Hamming distance is the number of positions for which the corresponding characters are different in two strings of equal length. For example, if p and q are two strings of the same length, the Hamming distance $H(p, q)$ is the number of places in which the two character strings differ, i.e. the number of substitutions required to make them equal. More formally, the distance between strings p and q with equal length of l characters is $\sum_{i=1}^l |p_i - q_i|$, here, i represents a character position in the given strings. In the proposed framework, the number of characters in the input string is measured and compared with the length of each word in the dictionary word database. Thus, whenever the framework finds a word of the same length with Hamming distance 0, the player is assumed to have entered a valid keyword, otherwise the WordNet lexical database [123] and its associated software tools are used to examine the input string further. Here, two databases (database containing 100,000 words and the WordNet lexical database) are used in order to minimize the risk of rejecting a valid keyword by the framework. In addition, the WordNet lexical database is also used for finding similarities or synonyms among the player input keywords and given annotations of the other players.

When the string is a valid word, WordNet usually produces a stemmed word and its associated synonyms, and the system assumes that the player has entered a valid keyword. Similarly, if the string does not represent a valid word, no outcomes will be produced by the WordNet, implying that the player has entered an invalid keyword. Whenever an invalid string is detected, the player will be asked to re-enter a valid keyword. Following the word search, the proposed payoff calculation unit calculates the player's payoff and finally, the score computation module calculates the player's score based on the payoff function and its outcome. This loop will be repeated until the end of the game session.

6.2 Experimental Evaluation

For testing purposes, all experiments were conducted with a set of threshold parameters, which were chosen using a validation set of images (not a part of the test set that was used to measure performances of the proposed model). The experiment parameters were chosen based on a testing experiment, which was conducted with a group of 30 game players. The experiment was conducted with 50 images chosen from the Caltech database (all of which were non-annotated). During the experiment using the interface INT-2, action properties a_1 and a_2 were assigned the value of 1 in order to simulate the worst case scenario, whereby all the annotations given by the players are accepted.

Given that test results indicate that a minimum of 1 and a maximum of 5 annotations are needed to find a single correct annotation, the threshold T_1 (allocated cost per annotation or limiting the maximum number of annotations per image) was set to 0.2, which would allow the players to label images using 5 different annotations. Based on the experiment outcome, T_2 (used for a_2 calculation) is assigned the value of 301. This is perceived as an acceptable value for the game, as previous experiments suggested that rationally minded players would not complete three incorrect annotations in a single row. Similarly, threshold T_3 (exposing partially or non-annotated content) is offset to 0.63, as it corresponds to the average value of valid contributions of trustworthy game players. By assigning this value, most of the partially annotated contents will be exposed to true game players, i.e. players who mostly enter correct annotations. As a consequence more accurate annotations are extracted. The AQL is assigned a value of 0.03. It has been found that the dictionary mechanism fails to identify 3% (0.03) of the valid keywords in practice. In practice, the WordNet lexical database failed in recognising some valid keywords, such as ‘Binocular’ and ‘Scissor’. However, as whenever the WordNet rejects a valid keyword, synonyms associated to the word will not be illustrated, the risk of rejecting a valid keyword by the framework

increases. This risk is measured by the AQL, which represents the incorrect annotations that the system is willing to accept in practice.

6.2.1 Performance measure in image classification

In practice, different types of players can be recognized, one of which is ‘random cheater’. Such players tend to cheat at random time intervals and will, thus, enter both correct and incorrect annotations during the game play. Since Player 2 in this game is not a fully independent player of Player 1, the payoff of Player 2 will be low when based on the scores of random cheaters. This is one of the reasons why Player 2’s payoff is designed to be dependent on a number of factors, including classification outcomes and the probability of entering a given annotation. More formally, there is a risk that a correct annotation entered by a random cheater could be rejected when the Player 2’s good contribution is lower than the associated cost, i.e. $P(G_2) \leq P(B_2)$. This problem has been partially mitigated by introducing the classifier. As a result, whenever Player 1’s payoff is low, classification outcomes are used to weight the Player 2’s good contribution ($P(G_2)$) and could thus make it higher than the corresponding cost, i.e. $P(G_2) > P(B_2)$, for a good annotation. Since the trained concept based on the players input keyword is selected, it is fair to use classification outcomes for weighting the Player 2’s good contribution. In practice, this process increases the probability of accepting a correct keyword given by a player inclined to cheat, whilst rarely accepting an incorrect keyword.

Two SVM classifiers were used for testing, trained with 50 and 500 positive and negative images and referred to as classifier-1 and classifier-2, respectively. Table 6.1 shows the results (precision) obtained for classifier-1. The performance is tested for the following concepts: butterfly, cougar, tree, building, cloud and tiger. Table 6.2 shows the Correct Rejection Rates (CRR) for classifier-1, calculated by dividing the number of correct rejections by the total rejections made by the classifier. Here, descriptors CLD, DCD and EHD were merged to form a new descriptor, referred to as the ‘Merged descriptor’, which is constructed based on ‘Merging Fusion

6. EXPERIMENTAL EVALUATION

Method’ technique (see Section 4.4.2 for more details).

Table 6.1: Performances of the SVM classifier (Precision)

Precision	Butterfly	Cougar	Tree	Building	Cloud	Tiger
CLD	45%	12%	65%	45%	62%	50%
DCD	30%	5%	40%	20%	54%	45%
EHD	45%	12%	40%	65%	73%	53%
Merged descriptor	53%	16%	75%	75%	76%	58%

Table 6.2: Performances of the SVM classifier (CRR)

CRR	Butterfly	Cougar	Tree	Building	Cloud	Tiger
CLD	68%	23%	59%	41%	63%	52%
DCD	54%	17%	37%	43%	61%	64%
EHD	41%	34%	63%	58%	71%	44%
Merged descriptor	61%	34%	77%	67%	71%	66%

Table 6.1 and Table 6.2 clearly show that, even when using default parameters, the ‘Merging Fusion Method’ provides better performance in image classification for both precision and CRR.

Table 6.3 and Table 6.4 show the precision and CRR of the classifier-2, respectively, obtained under experimental settings identical to those used for classifier-1.

Table 6.3: Precision when trained with 500 images

Precision	Butterfly	Cougar	Tree	Building	Cloud	Tiger
CLD	66%	54%	77%	61%	79%	71%
DCD	71%	53%	71%	66%	66%	63%
EHD	64%	67%	65%	81%	77%	77%
Merged descriptor	73%	66%	82%	84%	77%	81%

6. EXPERIMENTAL EVALUATION

Table 6.4: CRR when trained with 500 images

CRR	Butterfly	Cougar	Tree	Building	Cloud	Tiger
CLD	71%	39%	64%	46%	77%	69%
DCD	59%	48%	40%	67%	64%	66%
EHD	61%	52%	71%	73%	78%	82%
Merged descriptor	71%	59%	73%	77%	75%	88%

Compared to classifier-1, classifier-2 shows better performance with the merged descriptor (for both precision and CRR). The experiment shows that image classification performance significantly increases whenever larger training sets are used. However, it must be noted that, in practice, classification outcomes can be degraded when a classifier is over-trained.

The advantage of using a classifier in the annotation framework developed as a part of this study is tested for 50 game players that were asked randomly enter correct and incorrect annotations (thus simulating random cheater behaviour). A set of 40 images from the Caltech database (20 fully annotated and 20 non-annotated), as well as interface INT-2 and IA-GTSS, were used in the experiment. In addition, the final experiment setup included a phase of training (5 minutes), conducted with 40 images (20 fully annotated and 20 non-annotated) from the Caltech database, designed to acquaint the players with the game. The test results indicate that the precision of image annotation improved by 11% and 19% with classifier-1 and classifier-2, respectively. Although classifier-2 gives better performance in image classification, classifier-1 was used in all the experiments described in this thesis, as classifier-2 was not available at the time the proposed game framework was tested. In practice, it is difficult to make a large database of trained concepts. Therefore, the proposed framework uses a limited number of trained concepts and as a consequence most of the time, Player 2's payoff is calculated based on Player 1's payoff and $P(K)$.

6.2.2 Measure of Usability

The games developed as a part of this study were evaluated with two popular games—ESP and Phetch. Although player experience is essential to performance in computer games, no universal model that can measure the player experience currently exists [124] [125]. Although several heuristic works are available in the literature, based on elements such as the game interface, mechanics and game play [126] [127], no general model has been developed yet. Thus, in this work a one-to-one comparison was used for evaluating players' excitement level, enjoyment, addiction and the difficulty of playing games. The test was conducted with IA-GTMM. Since both ESP and Phetch are online games, they could not be linked to the database used by INT-1 and INT-2. However, in practice images use by ESP and Phetch were very similar to those given in the ESP dataset⁸ (ESP dataset is created from images that are been annotated by ESP game). Therefore, in order to enable a fair evaluation, INT-1 and INT-2 were given a set of 50 images (20 fully annotated and 30 non-annotated) from the ESP dataset, which were similar to those used for testing ESP and Phetch. A representative sample of images is depicted in Appendix B.

As before, the final experimental setup included a training phase (5 minutes per game) using 40 images (20 fully annotated and 20 non-annotated) from the ESP database. Since the objective is to measure the usability, players were instructed to play games as they wish to do in both training and testing sessions. However, all the players were fully informed of the purpose of the test and the objectives of the experiment. Once the training phase was complete, each player was tested by playing the game for 5 minutes, after which he or she was asked to enter their gender (male, female), age and occupation, as well as to provide 1 (low value) to 10 (high value) ratings on excitement factor, enjoyment, addiction and the perceived

⁸ <http://www.cs.cmu.edu/~biglou/resources/>

6. EXPERIMENTAL EVALUATION

level of game difficulty. As this test aimed to measure usability and the test time was restricted to 5 minutes, the numbers of images that have been used by any of the four games, i.e. INT-1, INT-2, ESP and Phetch were not taken into account. Because ESP and Phetch are played online, when playing those games, the players could not be exposed to a particular image or image database. Consequently, for consistency, the effects of using different images were ignored in this test.

As empirical evidence has shown that people usually do not like filling long questionnaires, the one used in this study was made as simple as possible and asked only the necessary questions. A template of the questionnaire used for testing is given in Appendix C. Thus, the data yielded by the survey was used to measure the mean percentage of the each usability question, i.e. excitement factor, enjoyment, addiction, game difficulty level. Figures 6.2 to 6.5 show the resulting mean percentage for each usability question as a histogram.

In order to establish the significance of the results, two different tests of analysis of variance were conducted [128] (both from the ANOVA family, used for determining the existence of a statistically significant difference among several group means). Here, the statistical analysis software Analyse-it⁹ was used to perform the ANOVA testing. A one-way ANOVA was conducted for each usability question (excitement factor, enjoyment, addiction and game difficulty level) across the different age categories. The second one-way ANOVA was conducted for each usability question across the all four games, whereby the results were considered significant if $p \leq 0.05$, where p is the probability statement which represents the p -value or significance among the data. In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis. One often rejects the null hypothesis when the p -value is less than 0.05, which is the value generally accepted for statistical significance testing. For each ANOVA test, the p and F values were

⁹ Which is one of the widely used software tools in ANOVA testing (<http://www.analyse-it.com/>)

6. EXPERIMENTAL EVALUATION

given, where F indicates how different the means are relative to the variability within each sample [128]. The larger this value, the greater the likelihood that the differences between the means are due to other factors, rather than chance alone, which would indicate a real effect. Appendix D shows all outcomes of the ANOVA tests.

The usability test was carried out with 440 players, which yielded a set of 440 datasets. The sets showed a reasonable distribution of gender; 182 (41%) test players were female and 258 (59%) were male. 231 players were students (this including higher-educational and college students), 62 were job seekers, 121 were employed and 26 were retired people. The age distribution is depicted in Figure 6.1.

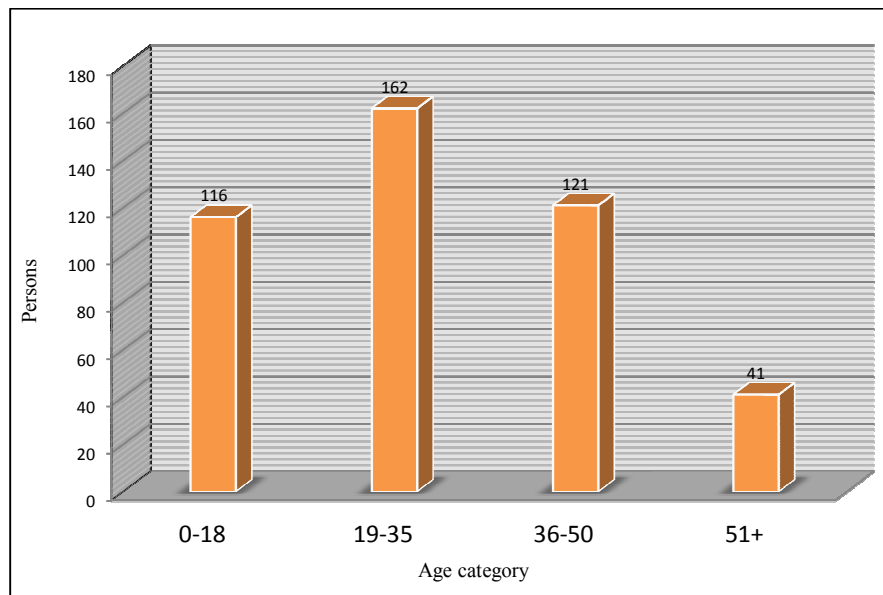


Figure 6.1: Age distribution of the players.

Measure of Excitement

The intention of this experiment was to find out the player's excitement level, which is potentially an important consequence of gaming. One can argue that the more excitement obtained, a larger audience will be attracted to a game [129].

6. EXPERIMENTAL EVALUATION

Figure 6.2 shows the obtained results of excitement. A one-way ANOVA shows that there was a significant difference in excitement levels across different age categories ($F = 32.71$, $p < 0.0001$), shown in Figure 6.2. More formally, the excitement level increased with players' age. Hence, the excitement level of the younger generation, (0-18), exhibits a low figure when compared to the other age groups. The reason is that the younger generation is attracted by more challenging games. However, overall results show that the excitement level of Phetch increased largely with age and therefore Phetch was able to outperform all the other games. The ANOVA conducted for excitement levels between data sets collected for the four games shows that there was a significant difference in excitement among all the 4 games ($F = 27.58$, $p < 0.0001$).

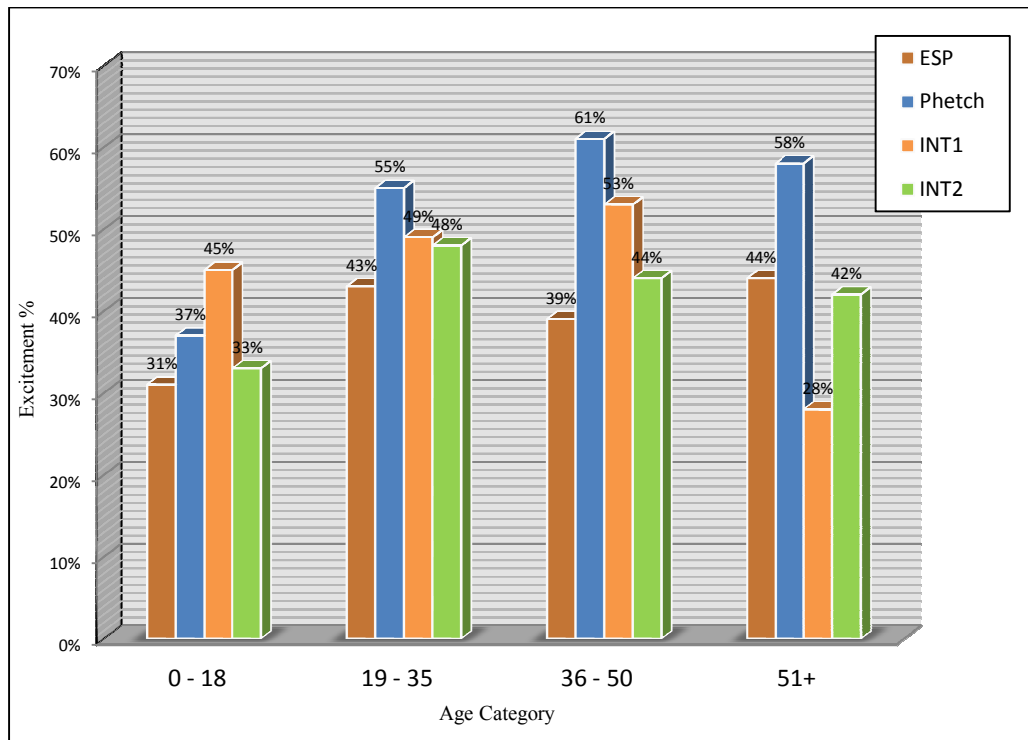


Figure 6.2: Excitement level of games.

6. EXPERIMENTAL EVALUATION

Measure of Addiction

Addiction is an important criterion that shows the attraction of players in gaming. That indicates the player's potential of repeatedly playing games. In Figure 6.3, the overall player's addiction is depicted. A one-way ANOVA shows that there were no significant differences in addiction levels across different age categories ($F = 1.77$, $p < 0.1513$). However, Figure 6.3 shows that reported addiction levels related to INT-2 remain stable among all the age groups and is higher than that associated with other games. The ANOVA tests for addiction levels across the four games show that there is a significant difference in addiction levels reported for individual games ($F = 16.52$, $p < 0.0001$).

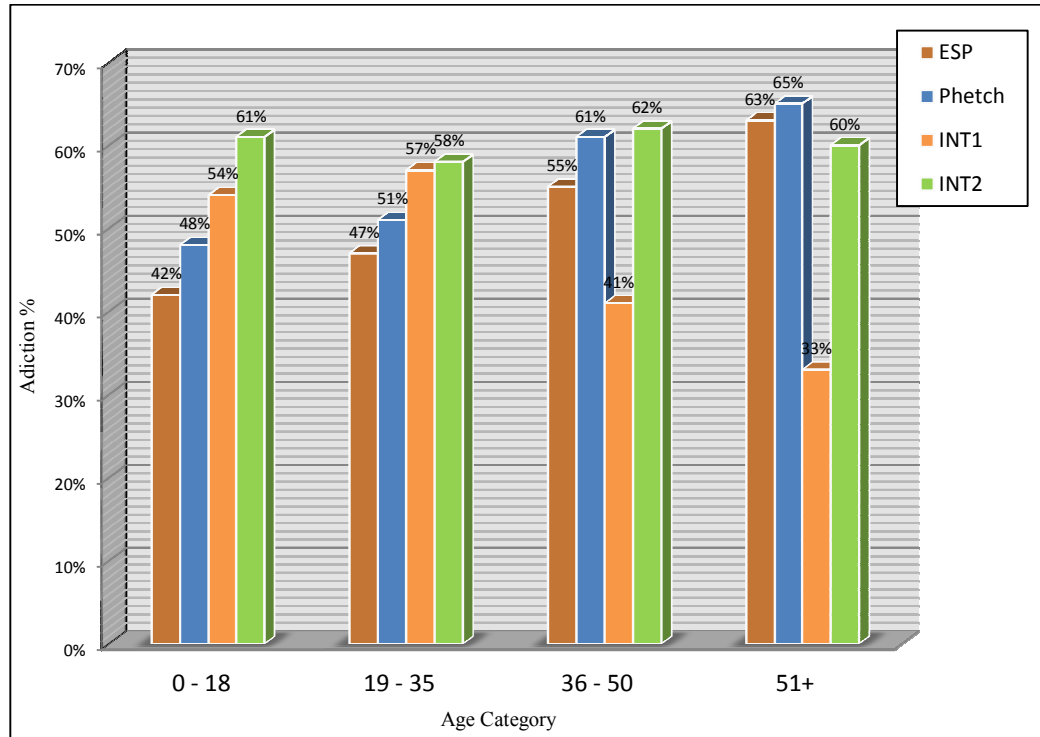


Figure 6.3: Addiction of games.

Measure of Enjoyment

It is assumed that a game is enjoyable when a number of conditions have been met. Overall, a game should offer appropriate challenge to the player, i.e. the task difficulty should match the user's current skill level, or rather be just a little bit more

6. EXPERIMENTAL EVALUATION

demanding so the user is required to put extra effort into the game¹⁰. Figure 6.4 shows the overall distribution of enjoyment. A one-way ANOVA shows that there was no significant difference in enjoyment level reported by players in different age categories ($F = 0.24$, $p < 0.8675$). Nonetheless, Figure 6.4 shows that enjoyment increases slightly for ESP and remains stable (or slightly decreases) with Phetch and INT-2 with an increase in players' age. In contrast, for INT-1, the enjoyment decreases significantly with an increase in age. The ANOVA conducted for enjoyment levels across the four games shows a significant difference ($F = 23.56$, $p < 0.0001$).

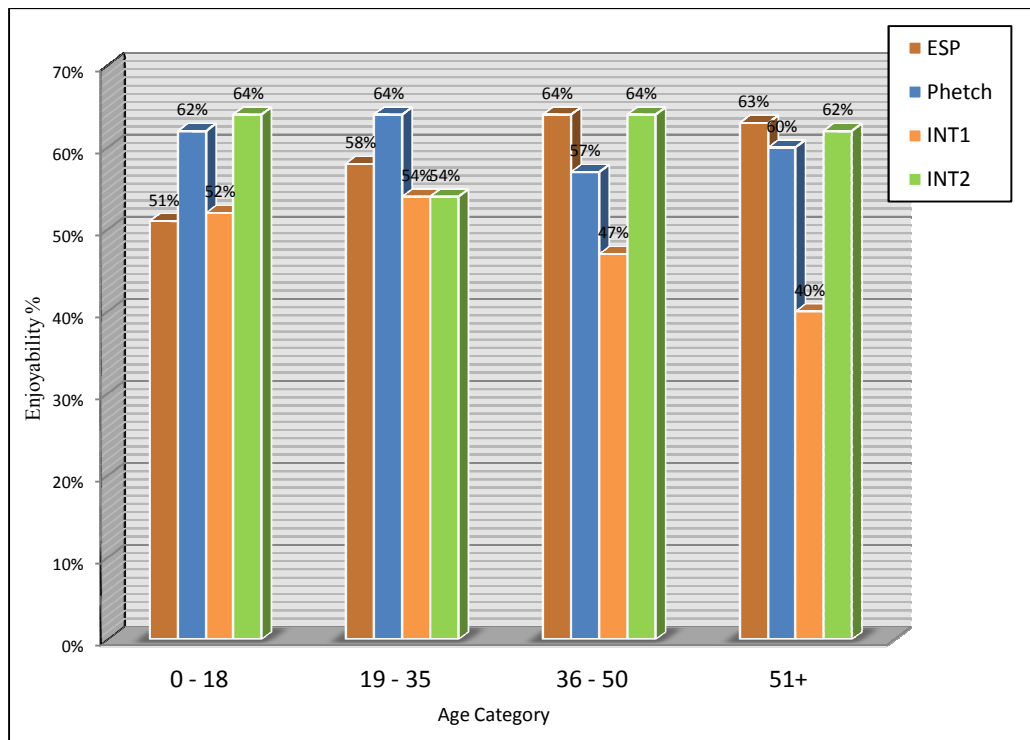


Figure 6.4: Enjoyability.

¹⁰www.upassoc.org/conferences_and_events/upa_conference/2004/program/Workshops/DesigningFun.html

6. EXPERIMENTAL EVALUATION

Measure of Difficulty

In practice, a game has to be of medium difficulty, so that the player is not bored by winning too easily. Also, we expect a player has to be mildly challenged to consider a game to be entertaining[126]. How much a game is perceived to be challenging is dependent on the player's skills in playing computer games in general. The intention of this experiment is to find out how difficult these games are to play in practice. As already noted, a game is more entertaining if it has a medium difficulty. Figure 6.5 shows the difficulty of playing all of the four games. A one-way ANOVA shows that there were significant differences between game difficulty levels reported by players in different age categories ($F = 12.27, p < 0.0001$). More formally, Figure 6.5 shows that INT-1 is the hardest game to play and ESP is the easiest. Overall, Phetch and INT-2 are the medium-hard games in practice. The ANOVA conducted across the four games with respect to game difficulty level indicates a significant difference in reported difficulty associated with playing the individual games ($F = 344.92, p < 0.0001$).

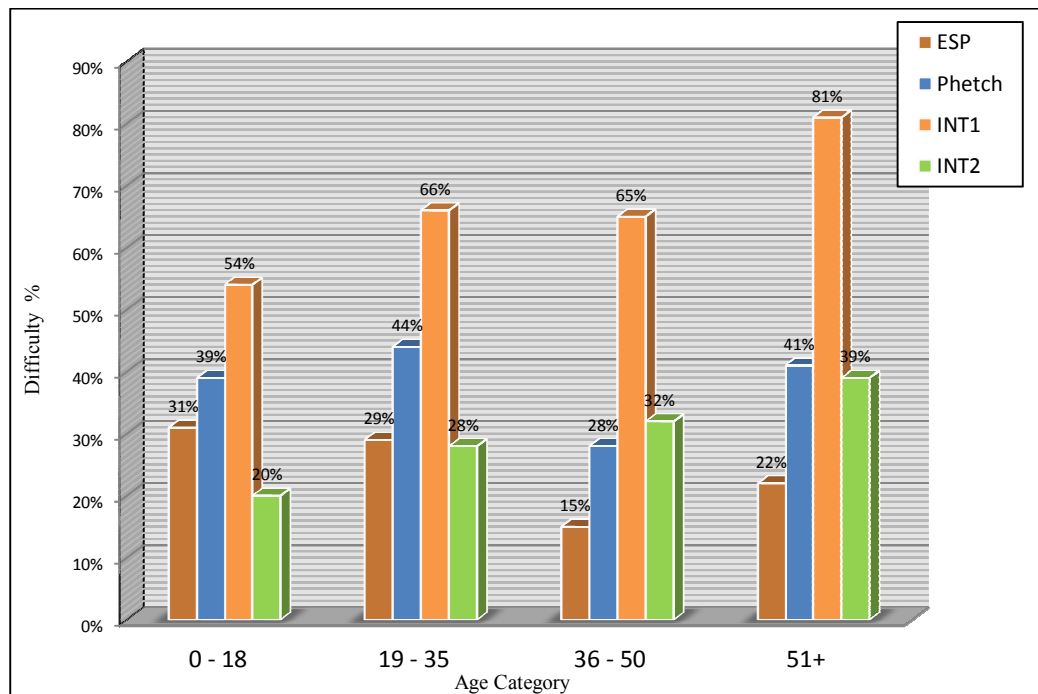


Figure 6.5: Difficulty in game play.

6. EXPERIMENTAL EVALUATION

From the above experiments, it is shown that the younger generation is more attracted by the complex games such as INT-1 rather than playing simple games like ESP, Phetch and INT-2. INT-1 is relatively hard to play when compared to the other games. However, with increased age, more players like playing less complex games such as ESP, Phetch and INT-2. Although Phetch outperforms INT-2 slightly in excitement, overall results show that more players like playing INT-2. It is deemed that introducing some new features, such as time limits and more destruction, may help to improve INT-2's excitement factor. However, it must be noted that this might further increase the difficulty level, which, as Figure 6.5 shows, increases with age for INT-2.

In practice, the order in which the individual games are played influences the usability test. In particular, significant changes in the perceived level of difficulty are noted. More formally, whenever INT-1 is played first, the reported difficulty level is moderate; whereas, if INT-1 is played as the last game, the difficulty is reported as high. Clearly, experience gained in playing other games affects the perception of difficulty level of any particular game. Consequently, the INT-1 is perceived as difficult, when no prior gaming experience can be used to help progress through the game, or serve as a reference point for comparison. In the tests conducted here, this problem is mitigated to some extent by randomly assigning the playing order to each player. This process was assumed to sufficiently reduce the influence of the order in which games were played, thus more accurate results were obtained from the usability test.

6.2.3 Measure of Efficiency

The purpose of this test was to measure the efficiency, i.e. the average number of annotations collected per minute by each game. Since information regarding efficiency is available for ESP, Phetch and KissKissBan (obtained from [3] [8] [39] respectively), these games have been evaluated with INT-1 and INT-2. Given that all three games use the ESP dataset, the efficiency of INT-1 and INT-2 were tested

6. EXPERIMENTAL EVALUATION

with the same dataset in order to obtain a fair evaluation. As before, the final experimental setup commenced with a training phase (5 minutes per game), which served to acquaint the players with INT-1 and INT-2. The training was conducted with 40 images (20 fully annotated and 20 non-annotated) from the ESP database.

A group of 30 game players participated in the tests. They were instructed to play games as they wished, without any time limitations. Due to the possibility that some players would be inclined to cheat, no information regarding the objective of the test was shared with the participants. It was assumed that, if players know the objective, there is a possibility that malicious players would cheat and rational players would enter correct annotations, which may influence the precision results. The test was conducted with 40 images (20 fully annotated and 20 non-annotated) using the IA-GTMM framework. Representative sample of images are depicted in Appendix B.

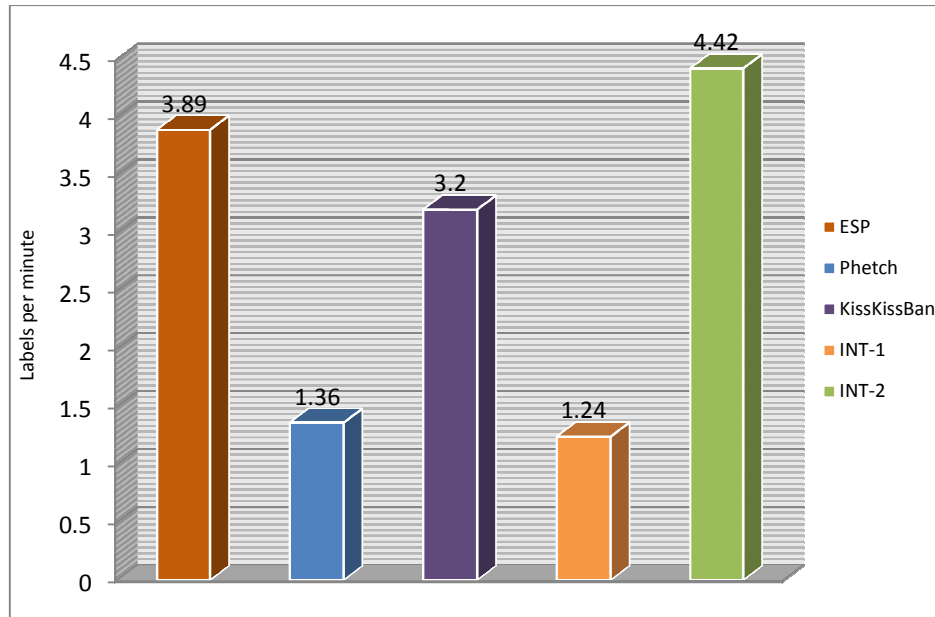


Figure 6.6: Average labels per minute.

Figure 6.6 shows the average number of labels collected per minute by each game. It clearly shows INT-2 outperformed other games by collecting more

annotations and that the efficiency is high in INT-2, i.e. the potential of obtaining a large number of annotation is high within a short period. Here, the overall number of annotations for ESP is slightly lower than that associated with INT-2, most likely due to the fact that ESP uses taboo words, which the players are not permitted to use. Given that these keywords have previously been used for annotating an image, ESP had already listed those keywords as the taboo words and would not allow players to use them in subsequent annotations. However, given the rather large number of taboo words, the players struggled to find matching keywords, thus wasting the gaming time.

6.2.4 Measure of Precision

Figures 6.7 to 6.9 show the precision for obtaining useful labels. Precisions of IA-GTMM and IA-GTSS were measured for three image datasets, namely ESP, Caltech and Corel. Each dataset was tested with 400 players (200 with IA-GTMM and 200 with IA-GTSS), so 1200 tests altogether with all 3 datasets. The precision of ESP and KissKissBan was evaluated for only the ESP dataset. This information is obtained from [3] [8] respectively and we did not perform that test by ourselves. Here, we did not concern ourselves with the precision of Phetch as the information was not available to use. To measure the precision, an independent rater was asked to give an opinion on whether the labels generated using the game were appropriate with respect to the images. Based on this outcome, the precision for each image was measured by dividing the obtained number of correct annotations by the total number of obtained annotations.

Since INT-2 outperformed INT-1 with respect to perceived usability, INT-2 was used in all the testing. As before, the final experimental setup included a training phase (5 minute) to acquaint the players with INT-2. For training purposes, 40 images (20 fully annotated and 20 non-annotated) were used from each dataset. Players were instructed to play games as they wished, with no limitations imposed on the duration of the testing session. As before, to avoid potential for cheating, the

6. EXPERIMENTAL EVALUATION

players were not informed of the objective of this test.

Since a large number of game players participated in this test, each image database was divided into 5 different groups of images, with 40 images (20 fully annotated and 20 non-annotated) in each group. In addition, each group was played by a set of 40 game players. Therefore, each IA-GTMM and IA-GTSS was tested with 200 game players (5×40). Furthermore, as each image database was divided into 5 different groups, this allowed for the average precision rate for each database to be measured. Figure 6.7 to 6.9 show the average precision rate for ESP, Caltech and Corel databases, respectively.

ESP Image Dataset

This dataset contained images from the World Wide Web. Dataset, which consisted of 200 images: 100 fully annotated and 100 non-annotated images. These images contain complex scenes and scenarios with large numbers of objects present, such as busy streets, seaside, landscape, office environments etc. Therefore, they cannot be categorised into a particular semantic category. Representative samples of images for each category are depicted in Appendix B. Average precision values are shown in Figure 6.7. Here, precision of ESP and KissKissBan is measured only for the ESP dataset (the information is not available for the other datasets).

6. EXPERIMENTAL EVALUATION

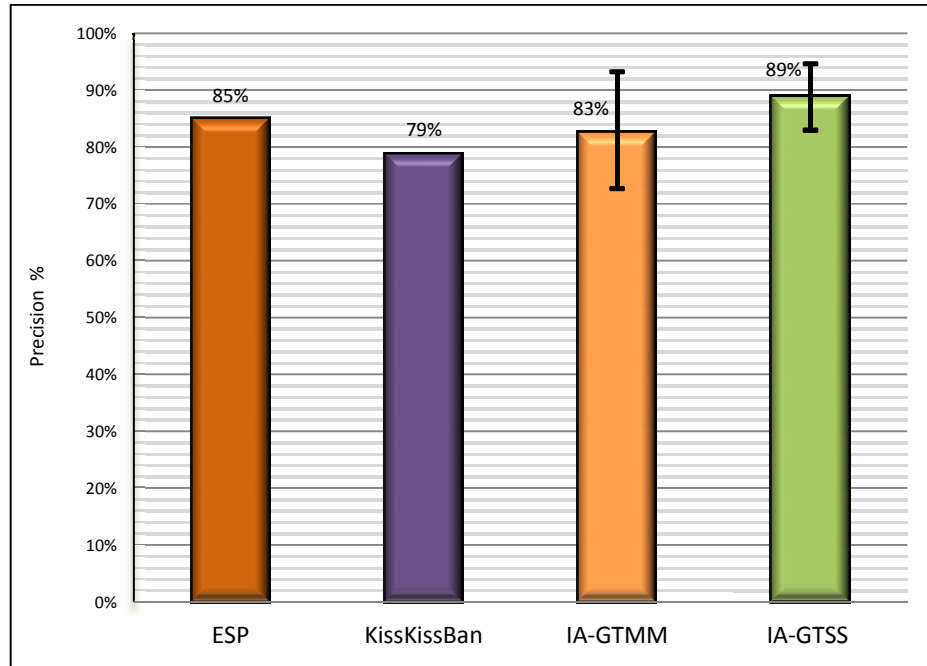


Figure 6.7: Average precision for ESP dataset.

Figure 6.7 clearly shows that IA-GTSS outperforms ESP, KissKissBan and IA-GTMM. Standard deviation (SD) for IA-GTMM and IA-GTSS was measured as 10.27 and 5.69, respectively, indicating that the spread of results associated with IA-GTSS is relatively low, compared to IA-GTMM. Since SD for ESP and KissKissBan was unavailable, only SD of IA-GTMM and IA-GTSS is shown in Figure 6.7. ESP is concerned with mainly obtaining different labels, rather than the major objects or characters that are more related to an image. ESP uses taboo words that are not permitted to be used by players. As a result, players are forced to describe an image using different labels which is not related to the image. This is a failure of ESP's precision when compared to IA-GTSS in practice.

Caltech Image Dataset

The second dataset was obtained from the Caltech image database and included 200 images (100 fully annotated and 100 non-annotated images) divided into 20 categories, each consisting of 10 images. The representative samples of images for each category are depicted in Appendix B.

6. EXPERIMENTAL EVALUATION

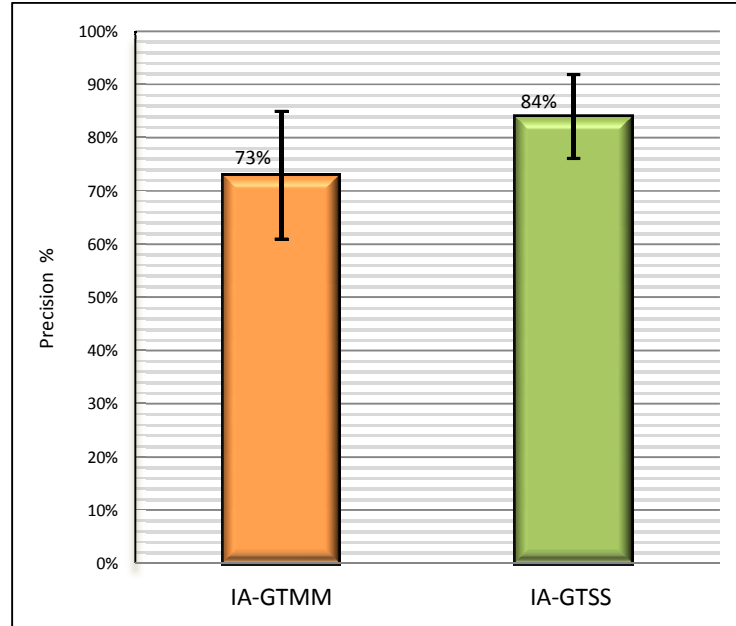


Figure 6.8: Average precision rates for Caltech dataset.

Figure 6.8 shows the performances for IA-GTMM and IA-GTSS. It shows that IA-GTSS gives the best results in image annotation. However, the number of annotations, i.e. the different labels obtained, is relatively low when compared to GTMM. SD associated with IA-GTMM and IA-GTSS was 12.01 and 7.89, respectively, suggesting that the spread of results reported for IA-GTSS is relatively low, compared to IA-GTMM.

Corel Image Dataset

A third set of images were obtained from the Corel image database. The Dataset consisted of 200 images; 100 fully annotated and 100 non-annotated images, divided to 7 semantic categories. Representative samples of images for each category are depicted in Appendix B. Average annotation precision rates for IA-GTMM and IA-GTSS are shown in Figure 6.9.

6. EXPERIMENTAL EVALUATION

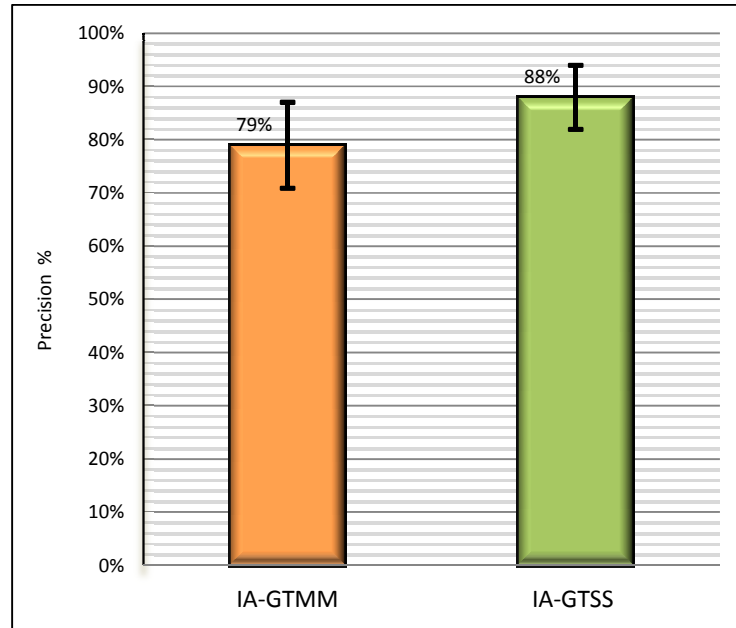


Figure 6.9: Average precision rates for Corel dataset.

Figure 6.9 shows the performances for IA-GTMM and IA-GTSS. Similar to Caltech image dataset, the figure shows that IA-GTSS yields the best results in image annotation. However, the number of annotations, i.e. the different labels obtained, is relatively low when compared to IA-GTMM. SD for IA-GTMM and IA-GTSS was 8.07 and 6.03, respectively.

Overall, results show that, for both IA-GTMM and IA-GTSS, the precision of ESP dataset is considerably higher, compared to the other datasets. More formally, whenever players are exposed to a set of similar images, i.e. images that are related to the same concept, such as images in Corel and Caltech databases, the obtained precision is relatively low, compared to that associated with the ESP database. Since ESP dataset contains different and complex images, it is likely that this influenced the players' behaviour, i.e. motivated them to perform good quality annotations.

Overall, the precision is the highest for IA-GTSS although number of annotations, i.e. different labels obtained by this framework is relatively low when

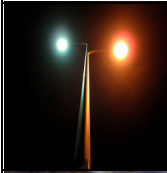


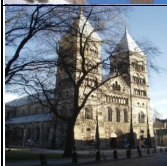
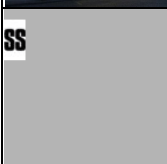
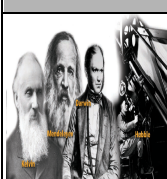
6. EXPERIMENTAL EVALUATION

compared to IA-GTMM. This issue is also found in practice, whereby IA-GTSS required a large number of samples to make decisions, i.e. for predicting the player's outcome. In other words, players need to be exposed to a large number of fully annotated contents before the prediction can be made. As a result, the efficiency of the system is low compared to IA-GTMM. This is one of the disadvantages that should be investigated in the future. However, for all three databases, IA-GTSS shows the best precision. Thus, the quality of most of the annotations obtained by IA-GTSS is high, and can thus be used to represent an image.

Table 6.5 and Table 6.6 show the keywords obtained by the proposed IA-GTMM and IA-GTSS frameworks for a part of the ESP dataset. The test was conducted with 20 game players, yielding 40 tests in total. Both IA-GTMM and IA-GTSS were tested with 40 images (20 fully annotated and 20 non-annotated images). In both tables, the column labelled as 'Keywords for IA-GTMM' shows the annotation obtained for the IA-GTMM and that labelled as 'Keywords for IA-GTSS' shows the annotations obtained for IA-GTSS. The column marked as 'Votes' represents the total number of times the particular annotation has been described by the players. Here, INT-2 was used for the testing and, as before, the final experiment setup included a 5-minute training phase. For training purposes, 40 images (20 fully annotated and 20 non-annotated) from ESP database were used. Players were instructed to play games as they wished and no time limitations were applied during the testing session. To avoid cheating, players were not educated about the objective of this test, as this knowledge might lead them to behave in a different manner in order to affect results in a different way.






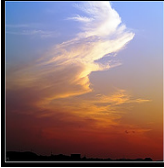
6. EXPERIMENTAL EVALUATION

Table 6.5: Performances of the proposed frameworks for ESP Dataset (part 1).

Image	Keywords for IA-GTMM	Votes	Keywords for IA-GTSS	Votes
	Light Dark Black Sun Animal	6 3 5 2 1	Light	4
	Woman Girl Black Animal	4 2 3 1	Girl Woman sexy	2 4 1
	Plane Cartoon Sky Animal Shit	5 3 5 2 1	Plane Sky	5 4
	Building Church Library Sky	5 2 1 2	Building Church Car	6 2 1
	Gray Characters Web Shit	7 2 1 1	Gray	3
	Men Kelvin Darwin Cartoon	7 2 3 1	Kelvin Men	1 3

6. EXPERIMENTAL EVALUATION

Table 6.6: Performances of the proposed frameworks for ESP Dataset (part 2).

Image	Keywords for IA-GTMM	Votes	Keywords for IA-GTSS	Votes
	Helmet Blood Killed Shit	5 6 2 1	Helmet Blood	5 3
	Sunglass Sunglasses Sunspeccs Black Nice	3 5 2 1 1	Sunglasses Sunglass	3 2
	Woman Girl Sexy Wow	7 3 2 1	Woman Girl	5 3
	Road Building House	6 6 2	Building Road	3 4
	Man Sunglasses Thug Bounty Police	5 4 2 2 1	Sunglasses Man Bounty	3 3 1
	Sky Sunset Nice Dark	6 3 2 1	Sunset Sky	2 3

6. EXPERIMENTAL EVALUATION

Table 6.5 and Table 6.6 show that incorrect annotations were associated with very few votes. Therefore, performance in image annotation could be further improved by eliminating the annotations associated with very few votes.

Payoff Representation

Figure 6.10 and Figure 6.11 show the payoff outcomes for IA-GTMM and IA-GTSS frameworks. The test used to collect this data was conducted with INT-2, whereby 10 game players were given a set of 20 fully annotated and 20 non-annotated images from the ESP database. As before, the final experiment setup included a 5-minute training phase. For training purposes, 40 images (20 fully annotated and 20 non-annotated) were used from the ESP database. Players were instructed to play games as they wished and no time limitations were applied during the testing session. Furthermore, in order to avoid cheating, players were not informed of the test objective.

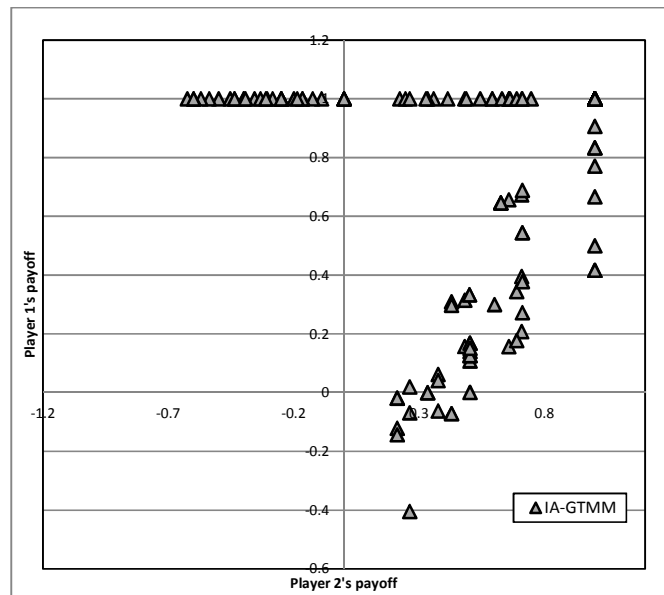


Figure 6.10: Payoff outcome for IA-GTMM

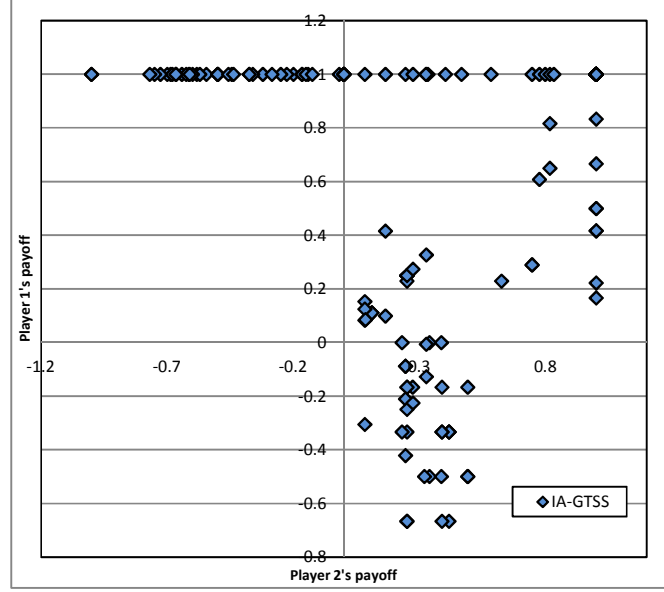


Figure 6.11: Payoff outcome for IA-GTSS

Figure 6.10 and Figure 6.11 clearly show that the feasible region is inside a convex hull of: 1. $(0,0)$, 2. $(-P(B_1), P(G_2))$, 3. $(P(G_1) - P(B_1), P(G_2) - P(B_2))$, 4. $(P(G_1), -P(B_2))$, which confirms the theoretical explanation given in Section 4.3.1 regarding the payoff outcome representation.

6.3 Result Obtained for Different Configurations of the System

As game players can be classified into several categories, here, the players were categorised into three major groups for experimental purposes-classical, random and true players. The classification is based on the player behaviour and propensity to cheat, whereby players that entered correct annotations in the beginning and later started to cheat are referred to as 'classical players'. Similarly, players that often changed their behaviour, i.e. entered both correct and incorrect annotations throughout the game, are defined as 'random players'. Finally, players that entered correct annotations most of the time are called 'true game players'. This experiment was conducted with 90 game players, using INT-2 interface, as it outperformed

6. EXPERIMENTAL EVALUATION

INT-1 with respect to usability rating. As before, the final experiment setup included a 5-minute training phase. For training purposes, 40 images (20 fully annotated and 20 non-annotated) from the Caltech image database were used. Furthermore, the players were not informed of the objective of the experiment, and were instructed to play as they wished, without any time restrictions to the test sessions.

The algorithms developed as a part of this study (IA-GTMM and IA-GTSS) were applied and their outcomes compared with those obtained for a framework with no prediction mechanisms installed. The chosen framework was based on a two-player game model with no prediction mechanisms installed and, thus, the action property a_1 was assigned the value of 1. During the test, a fully annotated content the player is exposed to is indicated by a blue square. Similarly, a correct annotation detected by the framework, i.e. true positive, is represented by a green square, and an incorrect annotation completed by the framework, i.e., false positive, is marked by a red triangle on the player confidence line $P(C_t)$. The confidence line shows the player's overall probability of entering a correct annotation in the game. Since players' behaviours are dynamic, they cannot be expressed in a generic way. To address this problem, a selected set of results from different players is illustrated. Thus, the results show individual player's output distribution for a set of images. Whenever player plays IA-GTMM, the player output distribution is represented by the probability distribution of the player's outcome, i.e, $P(C_{t+1}|C_t)$, $P(C_{t+1}|I_t)$, $P(I_{t+1}|C_t)$, $P(I_{t+1}|I_t)$ and $P(C_t)$. On the other hand, whenever player plays IA-GTSS or a two-player game model with no prediction mechanisms installed, the player's output distribution is represented in the figures by the probability distributions of $P(C_t)$ and $P(I_t)$. In all the cases, $P(C_{t+1}|C_t)$, $P(C_{t+1}|I_t)$, $P(I_{t+1}|C_t)$, $P(I_{t+1}|I_t)$, $P(C_t)$ and $P(I_t)$ are shown as $P(C|C)$, $P(C|I)$, $P(I|C)$, $P(I|I)$, $P(C)$ and $P(I)$, respectively.

The test was conducted with the Caltech image database and includes 240 images (110 fully annotated, 50 partially annotated and 80 non-annotated images),

divided to 40 categories, each consisting of 6 images. The representative samples of images for each category are depicted in Appendix B.

6.3.1 Two-player game model with no prediction mechanisms installed

Classical players

Figure 6.12 and Figure 6.13 shows the performances of the framework for two classical players. In Figure 6.12, a player cheated from point 9 to 14 and the framework was not able to detect them because Nash Equilibrium policy accepted those annotations as correct. This results in exposing more non-annotated contents to the player. In Figure 6.13, the framework located most of the incorrect annotations, and therefore fully annotated contents were exposed to the player. The experimental results indicate that this configuration can detect classical cheaters with 60% accuracy in image annotation.

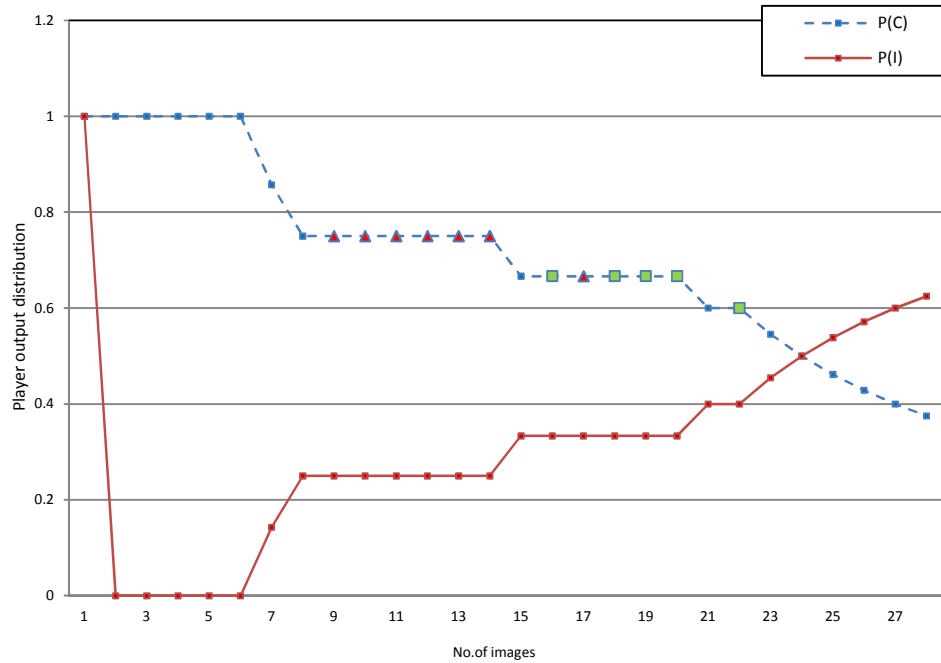


Figure 6.12: Performance measure for classical players, example - 1.

6. EXPERIMENTAL EVALUATION

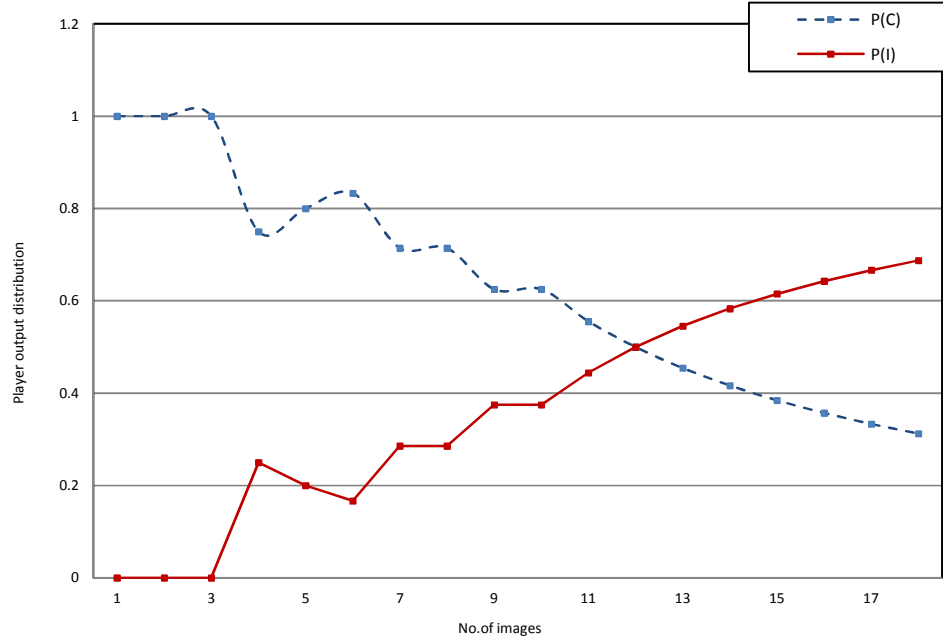


Figure 6.13: Performance measure for classical players, example - 2.

Random cheaters

It is difficult to measure performances of random cheaters. Figure 6.14 and Figure 6.15 shows the performances of the framework for two random cheaters. In Figure 6.14, two wrong annotations made by the player were not detected. In practice, most of the random cheaters acquired fewer contributions, i.e. low $P(C_t)$ in gaming. As a result, they are mostly exposed to fully annotated contents. The experimental results indicate that the two-player game model obtains a precision of 57% in image annotation for random players.

6. EXPERIMENTAL EVALUATION

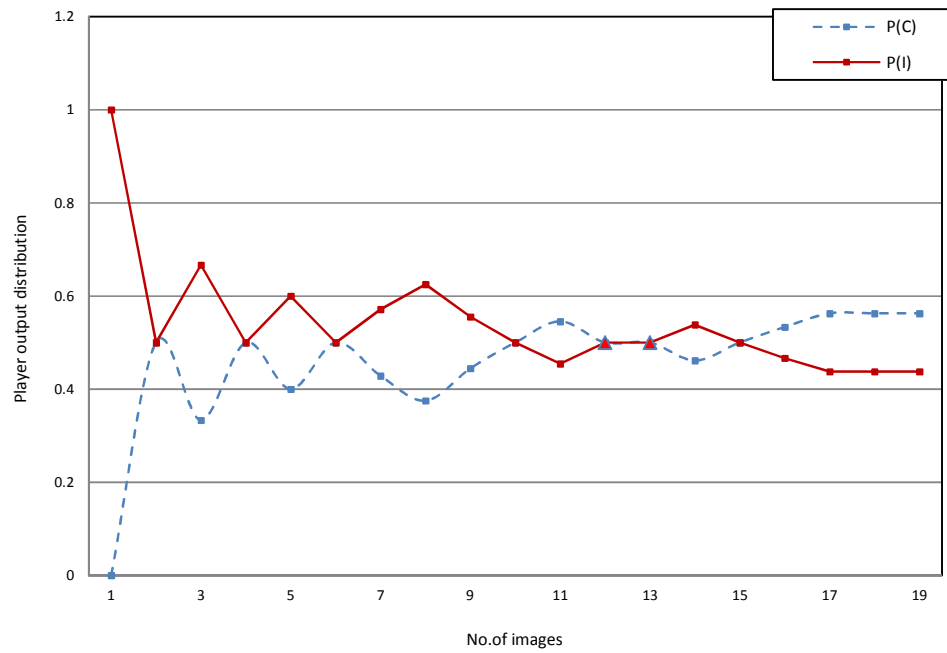


Figure 6.14: Performance measure for random players, example – 1.

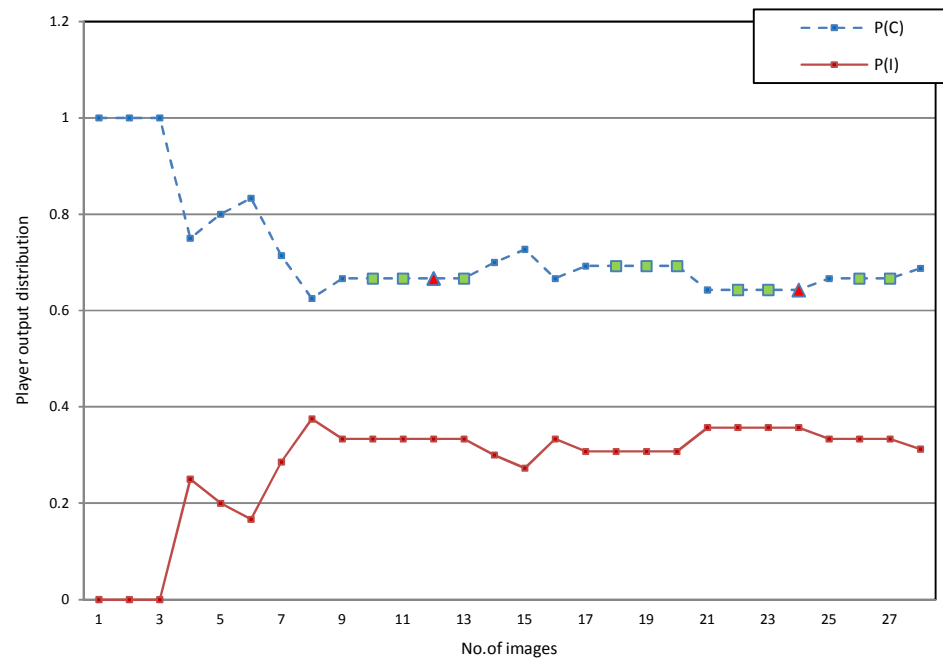


Figure 6.15: Performance measure for random players, example – 2.

Genuine players

Figure 6.16 and Figure 6.17 show the behaviour of two genuine game players. Here, players are performing well in image annotation by entering good annotations. In Figure 6.16, a player mistakenly made one wrong annotation which eventually degrades the players overall good contribution. The experimental results indicate that two-player game model is capable of detecting genuine players with 78% accuracy.

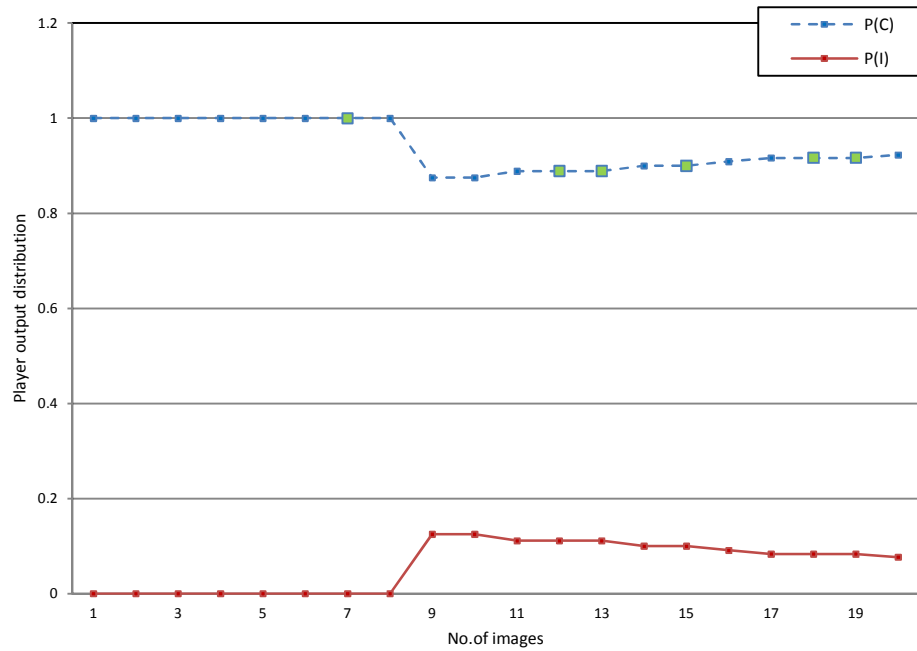


Figure 6.16: Performance measure for true players, example -1.

6. EXPERIMENTAL EVALUATION

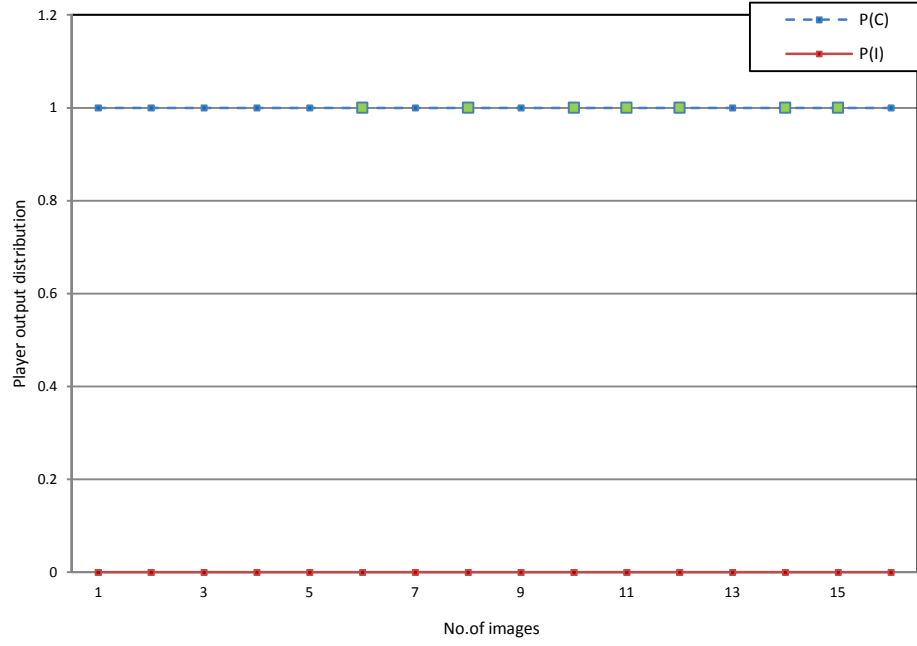


Figure 6.17: Performance measure for true players, example -2.

6.3.2 Two-player game model followed by the Markovian prediction

Figure 6.18 to Figure 6.23 show the outcome of this framework for classical, random and genuine players.

Classical cheaters

Some outcomes obtained from classical cheaters are shown in Figure 6.18 and Figure 6.19. In Figure 6.18, the framework correctly annotated 4 out of 5 images. Here, the player gave an incorrect annotation at 14 and, as a result, the framework measured $P(I_{t+1}|I_t) < P(C_{t+1}|I_t)$ to predict the players future outcome. In this instance, $P(I_{t+1}|I_t) = P(C_{t+1}|I_t) = 0$, and as a result, the framework exposes a fully annotated content to the player. However, when the player gave an incorrect annotation at 15, the framework measured $P(I_{t+1}|I_t)$ and $P(C_{t+1}|I_t)$ to predict the player's future outcome. Given that $P(I_{t+1}|I_t)$ is the greater of the two probabilities, this indicates that the player would enter an incorrect annotation with

6. EXPERIMENTAL EVALUATION

high probability. This results in exposing a fully annotated content to the player at 16. The experimental results show that this approach is capable of detecting classical cheaters in 81% of image annotation cases.

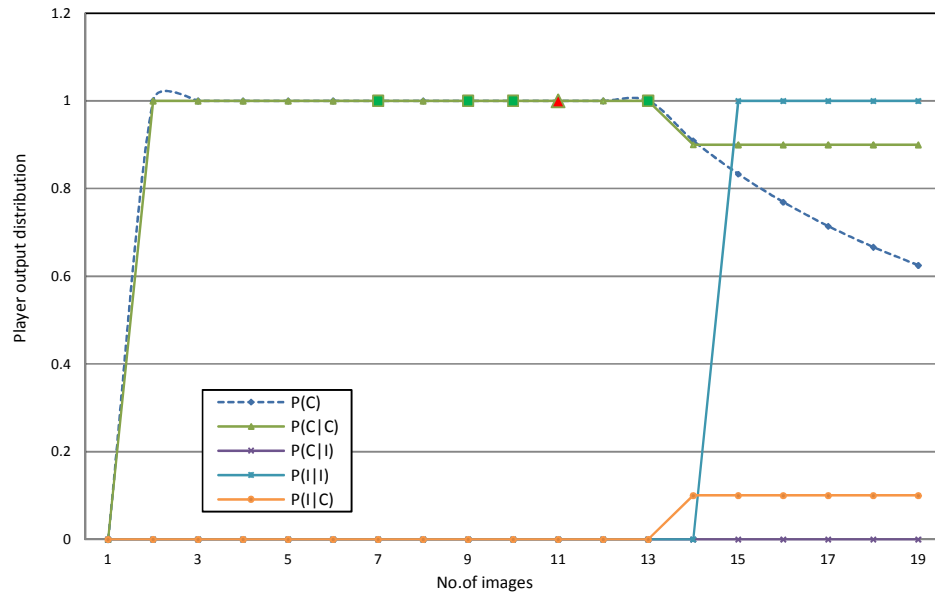


Figure 6.18: Performance measure for classical players, example -1.

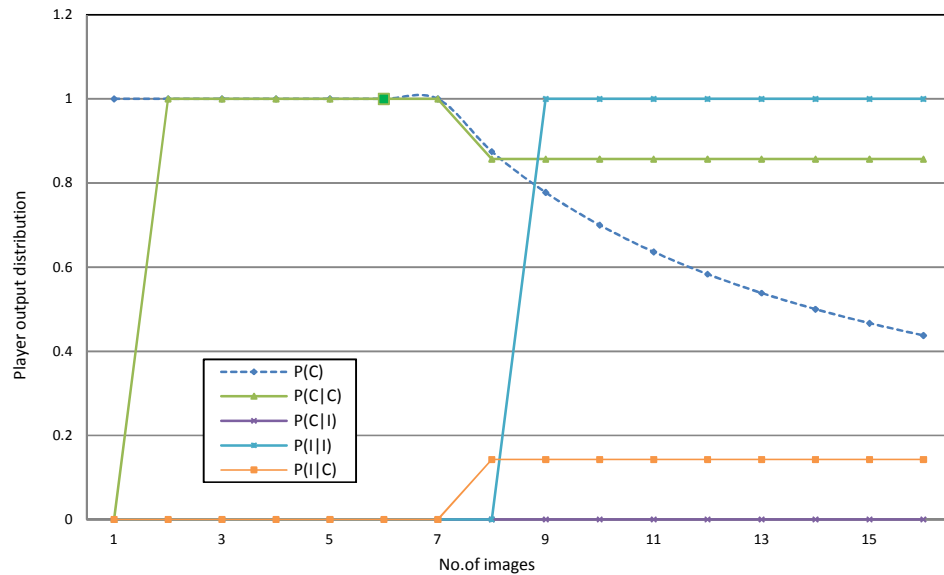


Figure 6.19: Performance measure for classical players, example -2.

Random cheaters

Random cheaters are the most difficult factor to detect in practice. Figure 6.20 and Figure 6.21 show the behaviour of 2 random cheaters. In Figure 6.20, a player annotated 9 out of 14 images correctly. Here, the player gave a wrong annotation at 11, resulting in the next outcome being predicted by $P(I_{t+1}|I_t) < P(C_{t+1}|I_t)$. It should be noted that, although $P(C_{t+1}|I_t)$ is greater than $P(I_{t+1}|I_t)$, a fully annotated content is presented to the player, based on the decision made by the Random Content Selection module (see Section 3.2). Here the player gave a wrong annotation at 17 and the next outcome is predicted by $P(I_{t+1}|I_t) < P(C_{t+1}|I_t)$. Outcome from 17 shows $P(C_{t+1}|I_t)$ is the largest probability that indicates the player's next outcome is good. Therefore, the player is exposed to a non-annotated content. The experimental results indicate that, for random cheaters, this approach obtains correct results in 65% cases.

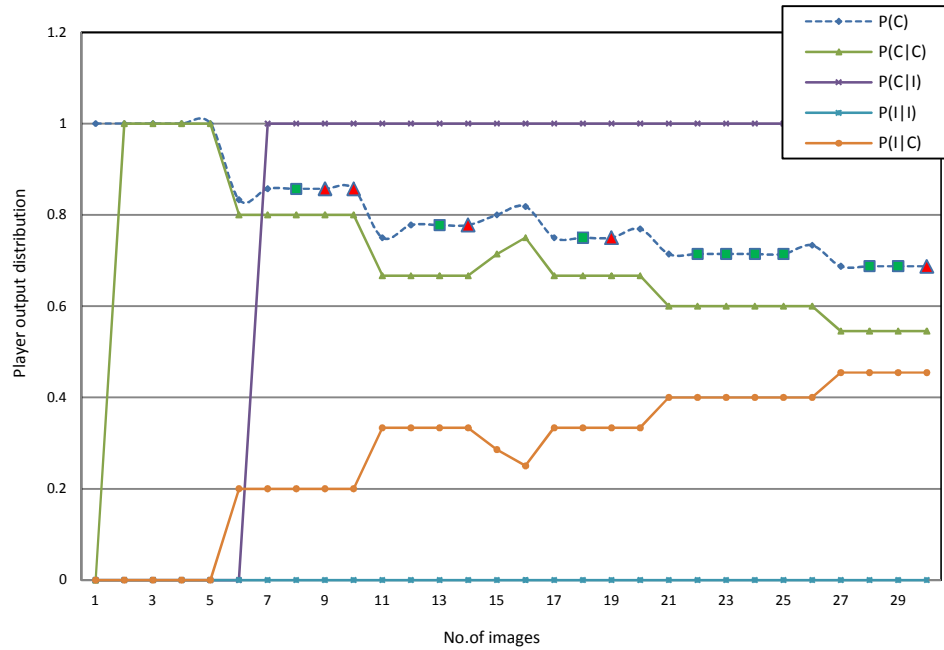


Figure 6.20: Performances measure for random players, example -1.

6. EXPERIMENTAL EVALUATION

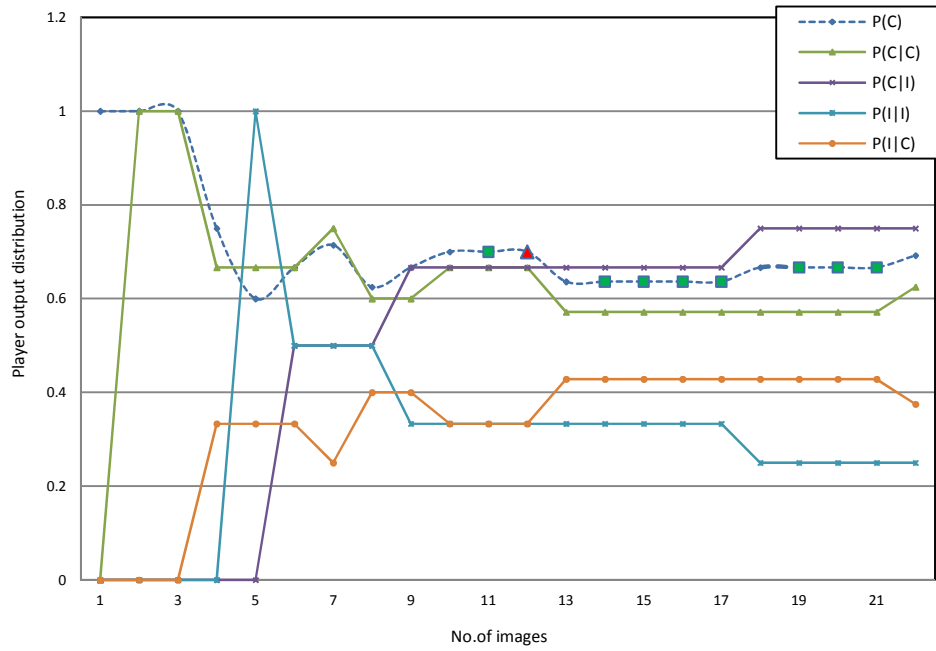


Figure 6.21: Performances measure for random players, example -2.

Genuine game players

Figure 6.22 and Figure 6.23 show performance of the framework for 2 genuine players. In general, the proposed framework performed well in detecting genuine players (players are also performing well in annotating images as they were more interested in collecting game points). In Figure 6.22, player annotated 19 out of 21 images correctly. Here, the player gave wrong annotations for three fully annotated contents and, as a result, the player's overall good contribution level was considerably reduced. The experimental results indicate that this approach is capable of detecting genuine players 84% of the time.

6. EXPERIMENTAL EVALUATION

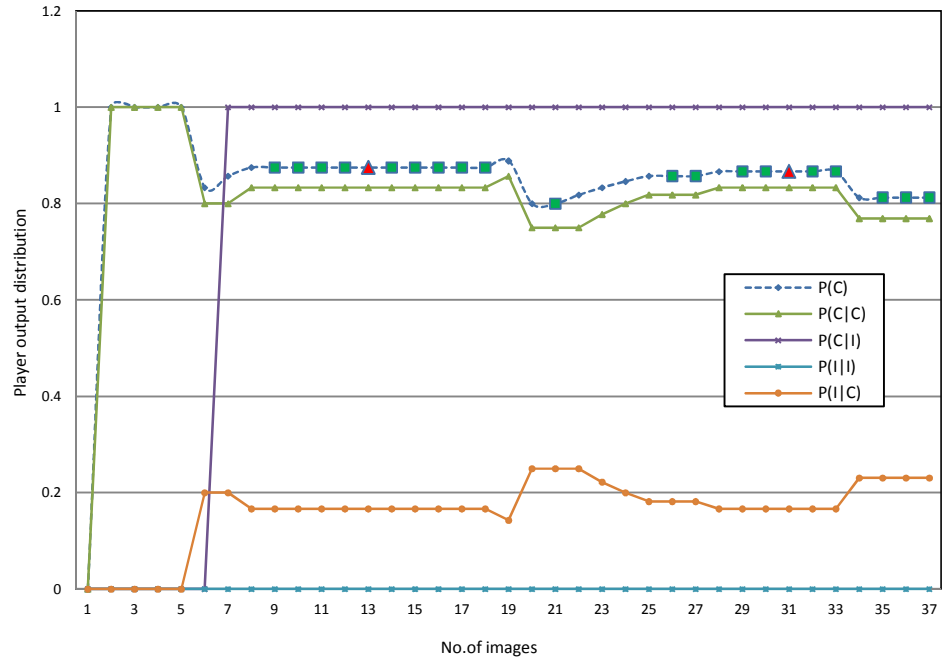


Figure 6.22: Performances measure for genuine players, example - 1.

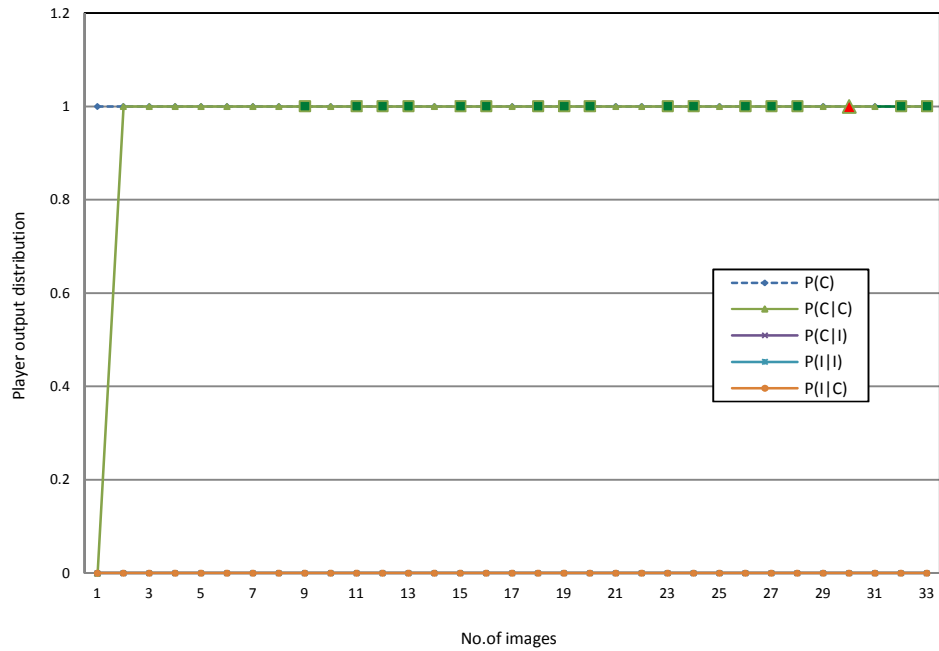


Figure 6.23: Performances measure for genuine players, example - 2.

6.3.3 Two-player game model followed by the proposed sampling prediction mechanism

In Figure 6.24 to Figure 6.26, outcomes for classical, random and genuine players are shown respectively. This approach appeared to have problems with exposing non-annotated images to the player, thus it exposes a large number of fully annotated contents and that the efficiency of this system is low, i.e. was able to collect very few annotations. However, for classical cheaters, the overall precision of this configuration was about 84%, for random cheaters it was about 79% and for true game players it was about 89%. This makes an overall precision of the system to 84% in image annotation.

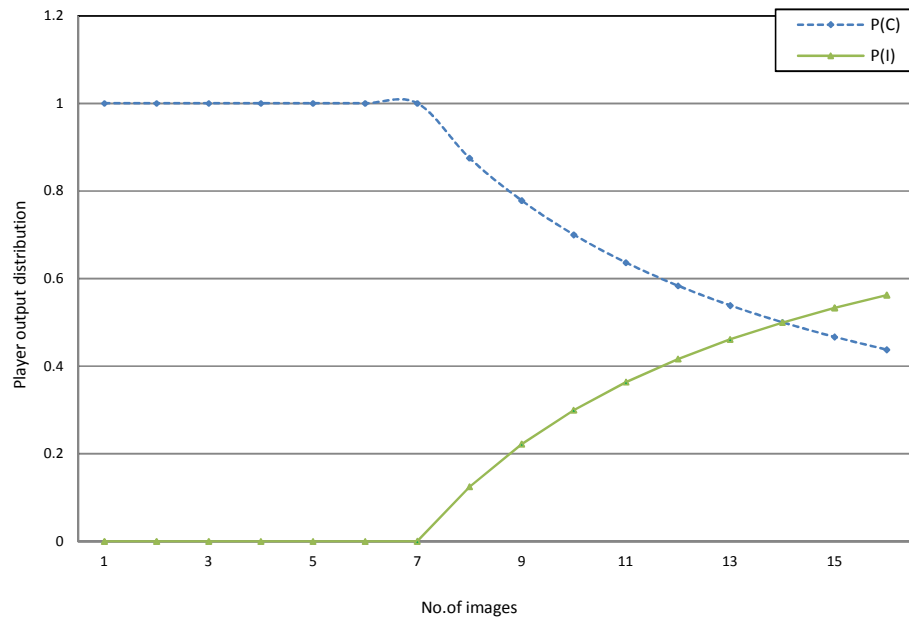


Figure 6.24: Outcome measure for classical players (Prediction by sampling).

6. EXPERIMENTAL EVALUATION

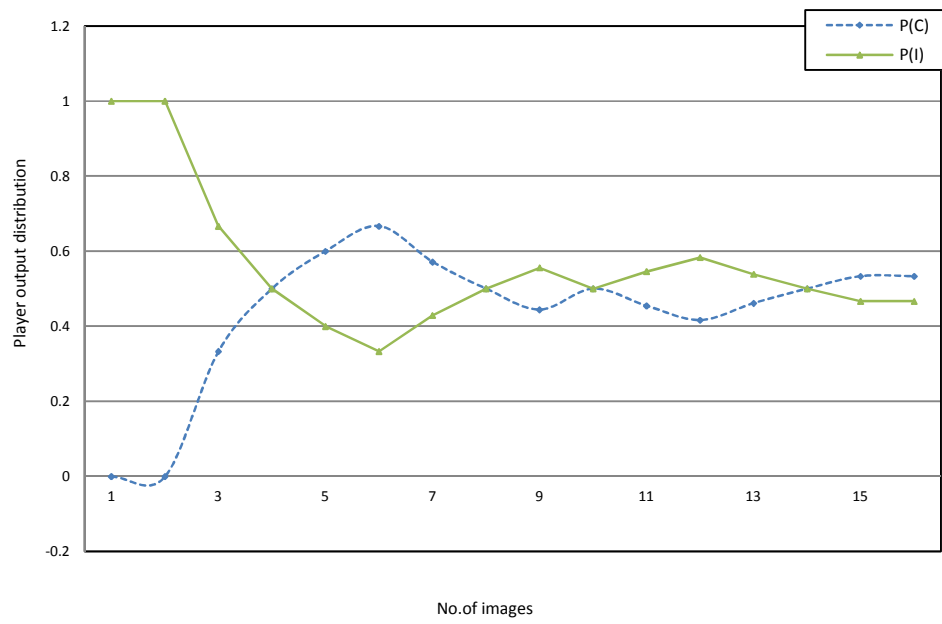


Figure 6.25: Outcome measure for random players (Prediction by sampling).

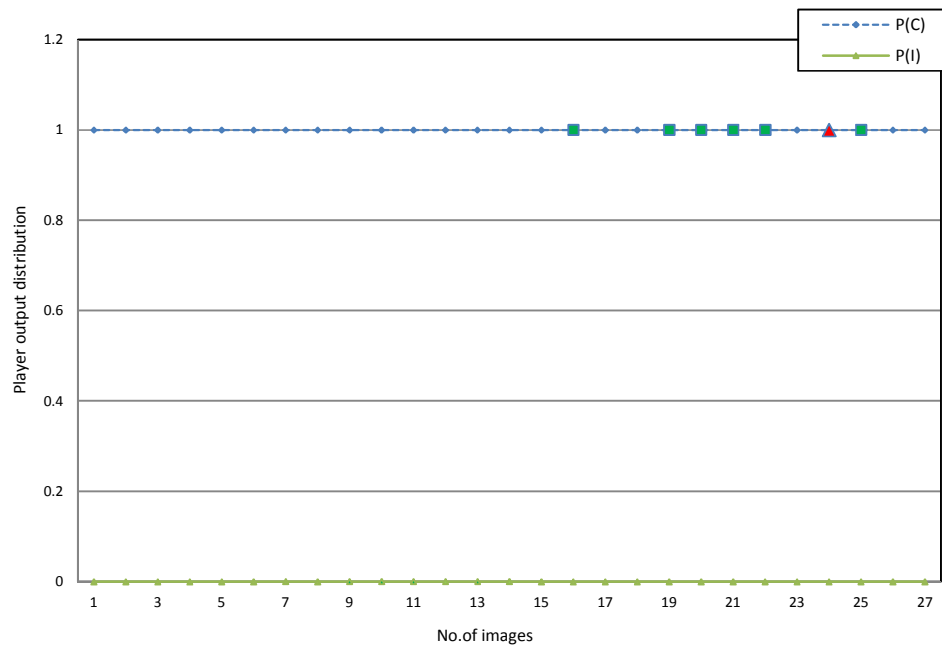


Figure 6.26: Outcome measure for genuine players (Prediction by sampling).

6.4 Summary

A comprehensive evaluation of the proposed framework was given in this chapter. The findings suggest that, with respect to image classification, SVM shows better results with the merged descriptor, compared to using any other single descriptor. With respect to perceived usability, with the exception of excitement, INT-2 outperformed all other games in player-reported levels of addiction, enjoyment and game difficulty level. Moreover, INT-2 has managed to capture more annotations in a given timeframe, indicating higher efficiency compared to other tested games. Regarding precision, IA-GTSS outperformed all other games, as it obtained high precision for all datasets. The results of the experiment conducted for different system configurations indicate higher precision associated with IA-GTSS, compared to IA-GTMM and the framework with no prediction mechanism installed. Here, IA-GTSS managed to obtain a large number of high-quality annotations that can be used to describe an image. Overall, it can be concluded that IA-GTSS predicts the player outcome more reliably than IA-GTMM and framework with no prediction mechanism installed. Based on all the experiments conducted as a part of this work, it can be concluded that IA-GTSS obtained higher precision and, thus, outperformed all other games with respect to image annotation. Although IA-GTSS obtained high precision, the main drawback of this algorithm is that it exposes players to only a few non-annotated and partially annotated contents. In fact, only a few different annotations were obtained with IA-GTSS, compared to IA-GTMM. This is one of the disadvantages that will be investigated in the future.

Chapter 7

CONCLUSION AND FUTURE WORK

The major issue in visual information indexing and retrieving is finding a way of connecting low-level information and high-level semantic information to represent the way humans perceive the world. Over the last decade, research has moved into a number of different directions to address this problem. One such approach is to use humans in a loop to solve complex problems by harvesting their brainpower. Game-based annotation of visual information is a computer vision application to the problem of data indexing and retrieval. This is based on actual content manually extracted by the players. The critical issue in game-based annotation is how to filter out bad annotations given by malicious players. Traditional game-based approaches use online multiplayer game strategies to tackle this problem. However, this technique faces some problems, such as that it cannot be installed in applications where only single isolated players are available: i.e. for the gadgets with no Internet connectivity. Furthermore, recent research shows that only 27% of teens were interested in online games and this is not what is concerned in the existing work of

7. CONCLUSION AND FUTURE WORK

GWAP. Moreover, traditional frameworks are not designed to tackle the major problem in online approaches, which is following a given strategy that leads to cheating. Addressing this and other drawbacks, it is worthwhile to design a system for tackling these issues. In this thesis, the appropriateness of standalone game inspired models to tackle these problems is derived. The proposed technique is inspired by Game Theories and their associated techniques.

In our proposal, the problem of making decisions; i.e. accepting or rejecting annotations, is tackled by Game Theory and its driven techniques. Player outcomes are always predicted prior to exposing non-annotated contents. For comparative purposes, two prediction techniques are proposed, one based on Markov models and the other based on Sequential Sampling algorithms. In the first proposal, Game Theory based decision process enhanced by prediction based on Markovian inference is derived. The evaluation experiments show the potential for using Game Theories and Markov models in image annotation. Here, the cheating oriented players are well recognised and thus the framework was able to capture correct annotations. The Markovian approach makes the Game Theory based decision model less dependent in decision making by predicting malicious players prior to exposing non-annotated contents. And it is the same with prediction by Sequential Sampling. However, Sequential Sampling technique exposes a large number of fully annotated contents to the players thus this makes the low efficiency in this approach. The experimental results for all three databases show that the Sequential Sampling approach yields high precision in image annotation, thus providing more accurate annotations than all the other approaches did.

Although the main application of the proposed approach is to label images, it can simply be used for solving large-scale problems such as labelling videos, sounds or even giving a meaningful sense to words etc. In addition, this work has turned a tedious work into something that people wanted to do in their spare time.

7. CONCLUSION AND FUTURE WORK

Potential enhancements and extensions for the proposed research are conceivable. Additionally, the proposed framework is modular and easily expandable to allow additional functionalities in the future. In summary, future directions for the development of the presented research may include the following actions:

- to investigate Markov model prediction performances by introducing higher-order Markov models.
- to investigate the issue of exposing small numbers of non-annotated images by the SS algorithm.
- to improve the performance of Nash Equilibrium based decision model by introducing more strategic actions, such as actions based on short-term and long-term historical performances of the player.
- to investigate the performance of the framework in places where the game is available, such as in mobile environments where large number of gamers exist every day.
- to investigate the framework's performance for different multimedia contents, such as for audio and video contents.
- to investigate the framework performances in multiplayer model games.

As can be seen, there are still many directions that exist to be covered for enhancing the performance of image annotation. However, the research presented in this thesis has provided suitable strategies for future research towards an enhanced game-based system for image annotation and shown the importance of using Game Theory driven mechanisms in decision making.

Bibliography

- [1] R. C. Veltkamp and M. Tanase, "Content-Based Image Retrieval Systems: A Survey," Utrecht University, Technical Report UU-CS-2000-34, 2002.
- [2] L. Kristina and A. Laurie, "Social Browsing on Flickr," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2007.
- [3] L. von Ahn and L. Dabbish, "Labelling images with a computer game," in *proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2004.
- [4] E. L. M. Law, L. von Ahn, R. B. Dannenberg, and M. Crawford, "TAGATUNE: A Game for Music and Sound Annotation," in *proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.
- [5] L. von Ahn, M. Kedia, and M. Blum, "Verbosity: A Game for Collecting Common-Sense Facts," in *Conference on Human Factors in Computing Systems*, 2006.
- [6] R. Jesus, D. Goncalves, A. Abrantes, and N. Correia, "Playing Games as a Way to Improve Automatic Image Annotation," in *Computer Vision and Pattern Recognition*, 2008.
- [7] V. Tuulos, J. Scheible, and H. Nyholm, "Combining Web, Mobile Phones and Public Displays in Large-Scale: Manhattan Story Mashup," in *proceedings of the 5th international conference on Pervasive computing*, 2007, pp. 37-54.

- [8] J. Chien, H. Tao, Y. Jane, and C. Kuan, "KissKissBan: A Competitive Human Computation Game for Image Annotation," in *proceedings of the ACM SIGKDD Workshop on Human Computation*, 2009, pp. 11-14.
- [9] J. Montgomery. (2010, January) Mobile Marketing Watch. [Online]. <http://www.mobilemarketingwatch.com/latest-study-76-of-users-dont-use-the-mobile-web-5040/>
- [10] A. Lenhart, J. Kahne, and E. Middaugh, "Teens, Video Games, and Civics," PEW Internet & American Life Project, survey 2008.
- [11] L. von Ahn, "Invited Talk: Human Computation," in *Knowledge Capture*, 2007, pp. 5-6.
- [12] C. Ho, T. Chang, and Y. Hsu, "PhotoSlap: A Multi-player Online Game for Semantic Annotation," in *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, 2007, pp. 1359-1364.
- [13] E. Rasmusen, *Games and information: an introduction to game theory*, 4th ed.: Wiley-Blackwell Publishing, 2007.
- [14] Yu. Xiaohan, Xu. Zeshui, and Qi. Chen, "A game model based on multi-attribute aggregation," *International Journal of Intelligent Systems*, vol. 26, no. 4, pp. 323-339, January 2011.
- [15] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.*: John Wiley & Sons, Inc, 1998.
- [16] Y. Xue, A. Leetmaa, and Ji. Ming, "ENSO Prediction with Markov Models: The Impact of Sea Level," *Journal of Climate*, vol. 13, no. 4, pp. 849-871, 2000.
- [17] K. P. Engelbrecht, F. Godde, F. Hartard, H. Ketabdar, and S. Moller, "Modeling User Satisfaction with Hidden Markov Models," in *10th Annual*

- Meeting of the Special Interest Group in Discourse and Dialogue*, 2009, pp. 170-177.
- [18] D. Chaudhuri, S. K. Ghosh, and A. R. Mukhopadhyay, "A Discursion on the Issues of Questionnaire Design for Sample Survey," *International Referred Research Journal*, vol. 8, no. 8, pp. 60-62, 2009.
- [19] D. J. Koehler and C. S. K. Poon, "Self-Predictions Overweight Strength of Current Intentions," *Journal of Experimental Social Psychology*, vol. 42, no. 4, pp. 517–524, 2006.
- [20] A. N. Hampton, P. Bossaerts, and J. P. O'Doherty, "The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans," *The Journal of Neuroscience*, pp. 8360-8367, 2006.
- [21] Z. Ghahramani and M. I. Jordan, "Factorial Hidden Markov Models," *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, 1997.
- [22] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," in *European Conference on Machine Learning*, 2010.
- [23] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image Annotation with TagProp on the MIRFLICKR Set," in *11th ACM SIGMM International Conference on Multimedia Information Retrieval*, 2010.
- [24] P. Koletsis and E. G. M. Petrakis, "SIA: Semantic Image Annotation using Ontologies and Image Content Analysis," Technical University of Crete (TUC), Lecture Notes in Computer Science 2010.
- [25] Li-Jia Li and Li Fei-Fei, "OPTIMOL: Automatic Online Picture Collection via Incremental Model Learning," *International journal of computer vision*, vol. 88, no. 2, pp. 147-168, 2010.

- [26] B Ommer and J. M. Buhmann, "Learning the Compositional Nature of Visual Object Categories for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, 2010.
- [27] F. D. Rivera, "Visual templates in pattern generalization activity," *Educational Studies in Mathematics*, vol. 73, no. 3, pp. 297-328, 2009.
- [28] K. Wang, X. Wang, and Y. Zhong, "A Weighted Feature Support Vector Machines Method for Semantic Image Classification," in *International Conference on Measuring Technology and Mechatronics*, 2010.
- [29] C. Galleguillosa and B. S. Belongiea, "Context based object categorization: A critical survey Computer Vision and Image Understanding," *Special Issue on Multi-Camera and Multi-Modal Sensor Fusion*, vol. 114, no. 6, pp. 712-722, 2010.
- [30] L. Zhou, C. Zhang, W. Wan, J. Birch, and Wei-Bang Chen, "An Image Clustering and Retrieval Framework Using Feedback-based Integrated Region Matching," in *International Conference on Machine Learning and Applications*, 2009.
- [31] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support Vector Machines for Histogram-Based Image Classification," *IEEE Transactions On Neural Networks*, vol. 10, no. 5, pp. 1055-1064, 1999.
- [32] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods.*: Cambridge University Press, 2000.
- [33] J. Huang, S. R. Kumar, and R. Zabih, "An automatic hierarchical image classification scheme," in *Proceedings of the sixth ACM international conference on Multimedia*, 1998.
- [34] A. A. Goodrum, "Image Information Retrieval: An Overview of Current

- Research," *Special Issue on Information Science Research*, vol. 3, no. 2, pp. 63-67, 2000.
- [35] S. Moran, "Automatic Image Tagging," School of Informatics University of Edinburgh, Master of Science 2009.
- [36] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, 2008.
- [37] Amazon Web Services, "Amazon Mechanical Turk: Developer Guide," Amazon Web Services, Developer Guide 2008.
- [38] L. von Ahn, M. Kedia, and M. Blum, "A Game for Collecting Common-Sense Facts," in *proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 75-78.
- [39] L. von Ahn, S. Ginosar, M. Kedia, and M. Blum, "Improving Image Search with PHETCH," in *International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 15-20.
- [40] S. Hacker and L. von Ahn, "Matchin: Eliciting User Preferences with an Online Game," in *proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, 2009, pp. 1207-1216.
- [41] L. von Ahn, R. Liu, and M. Blum, "Peekaboom: A Game for Locating Objects in Images," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2006.
- [42] I. King, Li. Jiexing, and K. T. Chan, "A Brief Survey of Computational Approaches in Social Computing," in *Proceedings of International Joint Conference on Neural Networks*, 2009.
- [43] E. Law and L. von Ahn, "Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games," in *Conference on*

- Human Factors in Computing Systems*, 2009.
- [44] L. Lee, G. Loewenstein, D. Ariely, J. Hong, and J. Young, "If I'm Not Hot, Are You Hot or Not? Physical-Attractiveness Evaluations and Dating Preferences as a Function of One's Own Attractiveness," *Psychological Science*, vol. 19, no. 7, pp. 669-677, 2008.
- [45] A. Nasr, F. Bechet, and A. Volanschi, "Tagging with hidden Markov models using ambiguous tags," in *proceedings of the 20th International Conference on Computational*, 2004.
- [46] G. Scutari, D. P. Palomar, and S. Barbarossa, "Optimal Linear Precoding Strategies for Wideband Noncooperative Systems Based on Game Theory- Part I: Nash Equilibria," in *IEEE Transactions on Signal Processing*, vol. 56, no. 3, 2008, pp. 1230-1249.
- [47] Khronos-Group. (2011) The Industry's Foundation for High Performance Graphics from games to virtual reality, mobile phones to supercomputers. [Online]. <http://www.opengl.org/about/overview/#1>
- [48] J. N. Webb, *Game Theory- Decision, Interaction and Evolution*, Mathematics Subject Classification ed. USA: Springer - Verlag London, 2007.
- [49] R. A. McCain, *GAME THEORY: A Nontechnical Introduction to the Analysis of Strategy*, Revised Edition ed.: World Scientific Publishing, 2010.
- [50] A. Kelly, *Decision Making using Game Theory: An Introduction for Managers*, 1st ed.: Cambridge University Press, 2003.
- [51] J. V. Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour*, 1st ed.: Princeton University Press, 1944.
- [52] J. Jr. Nash, "The Bargaining Problem," *Econometrica*, vol. 18, no. 2, pp. 155-162, 1950.

- [53] L. S. Shapley, *A value for n-person games*, 1st ed.: Defense Technical Information Center, 1952.
- [54] A. Perea, *Rationality in extensive form games*, 29th ed.: Springer, 2001.
- [55] T. L. Turocy and B. von Stengel, "Game Theory," Texas A&M University, Research Report 2001.
- [56] E. Rasmusen, *Games and information: an introduction to game theory*, 4th ed.: Blackwell Publishing, 2007.
- [57] B. C. Schipper, "Pure vs. Mixed Motive Games: On the Perception of Payoff-Orders," University of Bonn, JEL-Classifications 2001.
- [58] S. Morris and H. S. Shin, *Global Games: Theory and Applications.*: Cambridge University Press, 2006.
- [59] J. F. Nash, "Equilibrium Points in n-Person Games," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48-49, 1950.
- [60] E. Kalai and E. Lehrer, "Rational Learning Leads to Nash Equilibrium," *Econometrica*, vol. 61, no. 5, pp. 1019-1045, 1993.
- [61] A. J. Jones, *Game theory: Mathematical model of conflict.*: Horwood publishing, 2002.
- [62] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval," in *International Workshop on Content-Based Access of Image and Video Databases*, 1998.
- [63] J. Jantzen, "Introduction To Perceptron Networks," Technical University of Denmark, Technical report 1998.

- [64] T. Liu, "Fast Nonparametric Machine Learning Algorithms for High-dimensional Massive Data and Applications," Carnegie Mellon University, PhD Thesis 2006.
- [65] Y. Liao, "Neural Networks in Hardware: A Survey," University of California, Survey 2001.
- [66] D. H. Ballard, *An introduction to natural computation.*: The MIT Press, 1999.
- [67] P. W. Goldberg, "Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers," *Machine Learning*, vol. 18, no. 1, pp. 131-148, 1995.
- [68] C. C. Chang and Chih-Jen Lin. (2011) LIBSVM - A Library for Support Vector Machines. [Online]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [69] E. Kasutani and A. Yamada, "The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval," in *proceedings of International Conference on Image Processing (ICIP)*, 2001.
- [70] K. Wong, L. Po, and K. Cheung, "Dominant Colour Structure Descriptor for Image Retrieval," in *proceedings of International Conference on Image Processing (ICIP)*, 2007.
- [71] S. Chee, J. Soo, and D. Kwon, "Efficient Use of MPEG-7 Edge Histogram Descriptor," *ETRI Journal*, 2002.
- [72] L. Seneviratne and E. Izquierdo, "An Interactive Framework for Image Annotation through Gaming," in *11th ACM International Conference on Multimedia Information Retrieval (MIR)*, 2010, pp. 517-526.
- [73] S. Evaggelos, L. Hervae, M. Theoalos, C. Eddie, and A. Yannis, "Fusing MPEG-7 Visual Descriptors for Image Classification," *proceedings of*

- Artificial Neural Networks Formal Models and Their Applications (ICANN)*, 2005.
- [74] I. W. Selesnick, "Wavelet Transforms - A Quick Study," Polytechnic University, Physics Today Magazine 2007.
- [75] T. Piatrik, "Image Clustering and Video Summarisation using Ant-inspired Methods," Queen Mary University of London, PhD Thesis 2009.
- [76] M. S. Lew, *Principles of visual information retrieval*, 1st ed.: Springer-Verlag London, 2001.
- [77] C. C. Chen and D. C. Chen, "Multi-resolutional gabor filter in texture analysis," *Pattern Recognition Letters*, vol. 17, no. 10, 1996.
- [78] F. Mokhtarian and M. Z. Bober, *Curvature scale space representation: theory, applications, and MPEG-7 standardization*, 1st ed.: Springer, 2003.
- [79] T. Wu Lin and Y. F. Chou, "A Comparative Study of Zernike Moments for Image Retrieval," in *16th IPPR Conference on Computer Vision, Graphics and Image Processing*, 2003, pp. 621-629.
- [80] P. Salembier, T. Sikora, P. Salembier, and B. S. Manjunath, *Introduction to MPEG 7: Multimedia Content Description Language*, 2nd ed.: John Wiley and Sons, 2002.
- [81] Ka-Man Wong, Lai-Man Po, and Kwok-Wai Cheung, "Dominant Color Structure Descriptor for Image Retrieval," in *IEEE International Conference on Image Processing*, 2007, pp. 365 - 368.
- [82] B. Chanda and D. D. Majumder, *Digital image processing and analysis*, 1st ed.: PHI Learning Pvt. Ltd, 2004.
- [83] D. Messing, P. van Beek, and J. H. Errico, "The MPEG-7 Colour Structure

- Descriptor: Image description using colour and local spatial information," in *IEEE International Conference on Image Processing*, 2001.
- [84] M. R. Yong, K. Munchurl, Ho. K. Kang, B. S. Manjunath, and K. Jinwoong, "MPEG-7 Homogeneous Texture Descriptor," *ETRI Journal*, vol. 23, no. 2, 2001.
- [85] Wu. Peng, Y. Man Ro, C. Sun Won, and Y. Choi, "Texture Descriptors in MPEG-7," in *9th International Conference Computer Analysis of Images and Patterns*, 2001, pp. 21-28.
- [86] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: multimedia content description interface*, 1st ed.: John Wiley and Sons, 2002.
- [87] A. G. Barto, R. S. Sutton, and C. J. C. H. Watkins, "Learning and Sequential Decision Making," COINS Technical Report 1989.
- [88] R. Sun and C. L. Giles, "Sequence Learning: From Recognition and Prediction to Sequential Decision Making," in *IEEE Intelligent Systems*, 2001, pp. 67-70.
- [89] R. Begleiter, R. El-Yaniv, and G. Yona, "On Prediction Using Variable Order Markov Models," *Journal of Artificial Intelligence Research*, vol. 22, pp. 385-421, 2004.
- [90] B. Raj and E. U. D. Whitaker, "Lossless Compression of Language Model Structure and Word Identifiers," in *36th International Conference on Acoustics, Speech and Signal Processing*, 2003, pp. 388-391.
- [91] L. L. Ling and M. G. Lizarra, "Lossless compression of human static signatures," in *3rd International Conference on Signal Processing*, 1996, pp. 835 - 838.
- [92] A. Mehta and B. Patel, "DNA Compression Using Hash Based Data

- Structure," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 383-386, 2010.
- [93] H. Cai, S. R. Kulkarni, and S. Verdu, "An Algorithm for Universal Lossless Compression With Side Information," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4008-4016, 2006.
- [94] M. Zhu, F. Shen, and Z. Cheng, "Online Chinese Characters Recognition Based on Force Information by HMM," in *Human-computer Interaction: Interaction platforms and techniques*, Julie A. Jacko, Ed.: Springer-verlag Berlin Heidelberg, 2007, pp. 522-528.
- [95] Z. Jacob and N. Merhav, "On Context-Tree Prediction of Individual Sequences," *IEEE Transaction on Information Theory*, vol. 53, no. 5, pp. 1860-1866, 2007.
- [96] B. Knoll, "Ensemble Prediction by Partial Matching," University of British Columbia, Computer Science Course Project 2009.
- [97] J. Bulla, "Application of Hidden Markov Models and Hidden Semi-Markov Models to Financial Time Series," University Gottingen, PhD Thesis 2006.
- [98] P. Pathak, M. Sarwar, and S. Sohoni, "Markov Prediction Scheme for Cache Prefetching," in *Proceedings of 2nd Annual Conference on Theoretical and Applied Computer Science*, 2010, pp. 14-19.
- [99] B. Mobasher, D. Honghua, Tao Luo, and M. Nakagawa, "Using sequential and non-sequential patterns in predictive Web usage mining tasks," in *International Conference on Data Mining*, 2002, pp. 669 - 672.
- [100] G. Grimmett and D. Stirzaker, *Probability and random processes*, 3rd ed. USA: Oxford University Press, 2001.
- [101] L. Gerencser, "News from the World of Hidden Markov Models," *Special:*

- Space Exploration*, vol. 65, pp. 39-40, 2006.
- [102] G. D. Brushe, R. E. Mahony, and J. B. Moore, "A Forward Backward Algorithm for ML State Sequence," in *International Symposium on Signal Processing and its Applications (ISSPA)*, 1996, pp. 224-227.
- [103] M. Mitzenmacher and E. Upfal, *Probability and computing: randomized algorithms and probabilistic analysis*, 2nd ed.: Cambridge University Press, 2005.
- [104] W. Hoeffding and P. Kumar Sen N. I. Fisher, *The collected works of Wassily Hoeffding*, 1st ed.: Springer, 1994.
- [105] O. Watanabe, "Sequential Sampling Techniques for Algorithmic Learning Theory," in *Springer-Verlag Berlin Heidelberg*, 2001, pp. 27-40.
- [106] A. Harbitz, "An efficient sampling method for probability of failure calculation," *Structural Safety*, vol. 3, no. 2, pp. 109-115, 1986.
- [107] R. R. L. Kantam, K. Rosaiah, and R. G. Srinivasa, "Acceptance sampling based on life tests: log-logistic model," *Journal of Applied Statistics*, vol. 28, no. 1, pp. 121-128, 2001.
- [108] K. S. Stephens, *The handbook of Acceptance Sampling: Plans, Procedures, and Principles*. USA: American Society for Quality, 2001.
- [109] Department of Defence, "Sampling Procedures and Tables for Inspection by Attributes," Department of Defence, USA, Military standard 1963.
- [110] H. F. Dodge and H. G. Romig, *Sampling inspection tables: Single and double sampling*, Wiley classics library edition ed.: Wiley (New York), 1998.
- [111] H. Sackrowitz, "Alternative Multi-level Continuous Sampling Plans," *American Statistical Association and American Society for Quality*, vol. 14,

- no. 3, pp. 645-652, 1972.
- [112] W. E. Deming, *Some theory of sampling*, 1, Ed.: Dover Publications, Inc, 1966.
- [113] A. Wald, *Sequential Analysis*, Dover Phoenix ed.: John Wiley & Sons, 2004.
- [114] Z. Govindarajulu, *Sequential Statistics*.: Word Scientific Publishing, 2004.
- [115] R. Peck and J. L. Devore, *Statistics: The Exploration and Analysis of Data*, 7th ed.: Cengage Learning, 2011.
- [116] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*, 3rd ed.: Elsevier Academic Press, 2004.
- [117] K. Domicic, V. Bahovec, and N. K. Zivadinovic, "Studying an OC Curve of an Acceptance Sampling Plan:A Statistical Quality Control Tool," in *WSEAS International Conference on Mathematics & Computers in Business & Economics*, 2006, pp. 1-6.
- [118] C. D. Kemp and A. W. Kemp, "Generalized Hypergeometric Distributions," *Journal of the Royal Statistical Society*, vol. 18, no. 2, pp. 202-211, 1956.
- [119] D. Metzler, V. Lavrenko, and W. B. Croft, "Formal Multiple Bernoulli Models for Language Modeling ," in *27th Annual International ACM SIGIR Conference Proceedings of Special Interest Group on Information Retrieval*, 2004.
- [120] R. Ramaswamy and L. D. Servi, "The busy period of the migt1 vacation model with a bernoulli schedule," *Stochastic Models*, vol. 4, no. 3, pp. 507-521, 1988.
- [121] J. J. Heldt, *Quality Sampling and Reliability: New Uses for the Poisson Distribution*, 1st ed.: CRC Press, 1998.

- [122] A. K. M. Abdul and F. A. Burney, "Program for Item-by-Item Sequential Sampling by Attributes," King Abdulaziz University, Technical Report 1992.
- [123] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.
- [124] P. Sweetser and P. Wyeth, "GameFlow: A Model for Evaluating Player Enjoyment in Games," in *ACM Computers in Entertainment*, 2005.
- [125] R. Bernhaupt, M. Eckschlager, and M. Tscheligi, "Methods for Evaluating Games – How to Measure Usability and User Experience in Games?," in *Advances in computer entertainment technology*, 2007.
- [126] N. Beume et al., "Measuring Flow as Concept for Detecting Game Fun in the Pac-Man Game," in *IEEE Congress on Evolutionary Computation*, 2008, pp. 3448-3455.
- [127] H. Korhonen and E. M. I. Koivisto, "Playability Heuristics for Mobile Games," in *Human Computer Interaction with Mobile Devices and Services*, 2006, pp. 9-16.
- [128] B. Illowsky and S. Dean, *Collaborative Statistics*, 2nd ed.: Illowsky Publishing, 2008.
- [129] J. Vecer, T. Ichiba, and M. Laudanovic, "On Probabilistic Excitement of Sports Games," Columbia University, Department of Statistics, New York, 2007.
- [130] I. Steinwart and A. Christmann, *Support Vector Machines.*: Springer-Verlag, New York, 2008.
- [131] V. N. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*, 2nd ed. New York: Springer-Verlag, 2000.

Bibliography

- [132] S. R. Gunn, "Support Vector Machines for Classification and Regression," University of Southampton, Technical Report 1998.
- [133] H. J. Bierens, "Introduction to Hilbert Spaces," Pennsylvania State University, 2007.
- [134] M. I. Jordan and R. Thibaux, "The Kernel Trick," Berkeley, University of California, Technical Report 2004.
- [135] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Bell Laboratories, Lucent Technologies, Kluwer Academic Publishers 1998.

List of Author's Publications

1. **L. Seneviratne** and E. Izquierdo, "A Mathematical Approach Towards Semi-Automatic Image Annotation," in *European Signal Processing Conference (EUSIPCO)*, 2011.
2. **L. Seneviratne** and E. Izquierdo, "An Interactive Game for Semi-Automatic Image Annotation," in *35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.
3. **L. Seneviratne** and E. Izquierdo, "An Interactive Framework for Image Annotation through Gaming," in *11th ACM International Conference on Multimedia Information Retrieval (MIR)*, 2010.
4. **L. Seneviratne** and E. Izquierdo, "Image Annotation through Gaming (TAG4 FUN)," in *16th International conference on Digital Signal Processing (DSP)*, 2009.
5. **L. Seneviratne** and E. Izquierdo, "Image Annotation through Gaming," in *International conference of Collaborative Computing over Social Networking (CCSN)*, 2008.
6. **L. Seneviratne** and E. Izquierdo, "Image Annotation through Gaming," in *2nd K-Space PhD Jamboree Workshop*, 2008.

Appendix A

Support Vector Machines (SVM)

A.1 Introduction

Support Vector Machine [130] is a useful technique that is used widely in data classification. In SVM, the input vectors (features) are mapped non-linearly into a high-dimensional feature space through a Kernel, where a maximum separating line (hyperplane) is constructed. Then, on each side, two parallel hyperplanes are constructed. This process gives the maximum separation between two different classes. One of the key features of SVM is the non-linear decision regions that have better classification ability than traditional linear classifiers. SVM is designed based on the structural risk minimisation principle [131]. Figure A.1 shows an example of a binary classification problem.

In a binary separable learning problem, the set of indicator functions for defining separating hyperplanes can be represented as:

$$(\omega \cdot x_i) + b = 0, \quad \omega \in \mathbb{R}^n, \quad b \in \mathbb{R}, \quad i = 1, 2, 3, \dots, n \quad (\text{A.1})$$

where vector ω and scalar bias b define the actual location of the hyperplan.

It is said that when the distance between the closest data point to the hyperplane is maximal, then the data points are optimally separated by the hyperplane. There is some redundancy in (A.1), and without loss of the generality, it is appropriate to

A. Support Vector Machines

consider a canonical hyperplane [132], where the parameters ω and b are constrained by,

$$\min_i |\omega, x_i + b| = 1 \quad (\text{A.2})$$

A canonical hyperplane representation is obtained in the following form,

$$y_i((\omega, x_i) + b) \geq 1, \quad i = 1, 2, 3, \dots, n \quad (\text{A.3})$$

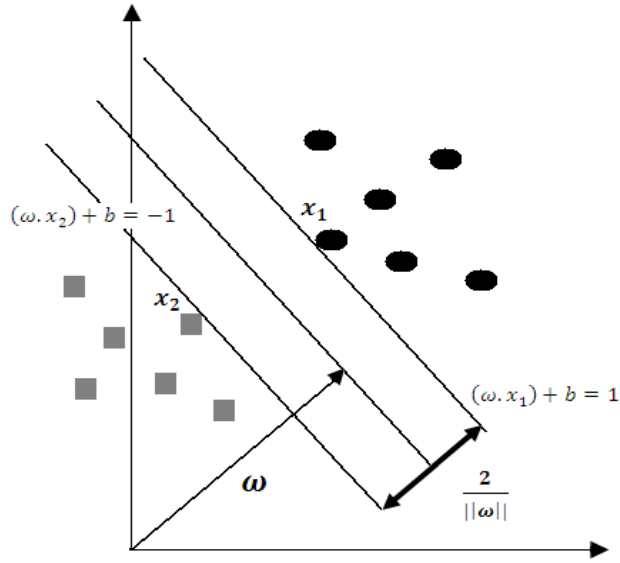


Figure A.1: Binary classification problem.

The distance $d(\omega, b; x)$ of a point x from the hyperplan (ω, b) is,

$$d(\omega, b; x) = \frac{|\omega, x_i + b|}{\|\omega\|} \quad (\text{A.4})$$

The optimal hyperplane is constructed by maximising the margin ρ . The margin is given by,

A. Support Vector Machines

$$\begin{aligned}
\rho(\omega, b) &= \min_{x_i: y_i = -1} (\omega, b; x_i) + \min_{x_i: y_i = 1} (\omega, b; x_i) \\
&= \min_{x_i: y_i = -1} \frac{|\omega, x_i + b|}{\|\omega\|} + \min_{x_i: y_i = 1} \frac{|\omega, x_i + b|}{\|\omega\|} \\
&= \frac{1}{\|\omega\|} \left(\min_{x_i: y_i = -1} |\omega, x_i + b| + \min_{x_i: y_i = 1} |\omega, x_i + b| \right) \\
&= \frac{2}{\|\omega\|}
\end{aligned} \tag{A.5}$$

Hence, the hyperplane that optimally separates the data is the one that minimizes,

$$\Phi(\omega) = \frac{1}{2} \|\omega\|^2 \tag{A.6}$$

This equation satisfies (A.3) and therefore it can show that (A.6) is independent of b (changing b will move in the normal direction to itself). Since the margin b remains unchanged, the hyperplan is not optimal, thus it may lie nearer to one class than other. Here, the structural risk minimisation (SRM) principle was used to minimise (A.6). Suppose that following bound holds,

$$\|\omega\| < A \tag{A.7}$$

Then from (A.3) and (A.4),

$$d(\omega, b; x) \geq \frac{1}{A} \tag{A.8}$$

Hence, the hyperplane cannot be near the $1/A$ to any of data points.

The VC dimension [67], h , of the set of canonical hyperplans in n dimensional space is bounded by,

$$h \leq \min[R^2 A^2, n] + 1 \tag{A.9}$$

where R is the radius of a hypersphere enclosing all the data points. Hence

A. Support Vector Machines

minimising (A.6) equals to minimising the upper bound on the VC dimension. The solution to the minimising problem is solved by introducing Lagrangian multipliers.

$$\Phi(\omega, b, \alpha) = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^1 \alpha_i (y_i |(\omega, x_i) + b| - 1) \quad (\text{A.10})$$

where α are the Lagrange multipliers. Here, the objective is to minimise the Lagrangian with respect to ω, b and maximised with respect to $\alpha \geq 0$. To make (A.10) easy to solve, it is transformed to its dual problem, which is given by,

$$\max_{\alpha} W(\alpha) = \max_{\alpha} (\min_{\omega, b} \Phi(\omega, b, \alpha)) \quad (\text{A.11})$$

The minimum with respect to ω and b of the Lagrangian, Φ , is given by,

$$\frac{\partial \Phi}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0 \quad (\text{A.12})$$

$$\frac{\partial \Phi}{\partial \omega} = 0 \Rightarrow \omega = \sum_{i=1}^n \alpha_i y_i x_i = 0 \quad (\text{A.13})$$

From (A.11), (A.12) and (A.13), the dual problem can be illustrated as,

$$\begin{aligned} \max_{\alpha} W(\alpha) &= \max_{\alpha} -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i, x_j) + \sum_{k=1}^n \alpha_k \\ \alpha^* &= \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i, x_j) - \sum_{k=1}^n \alpha_k = 0 \end{aligned} \quad (\text{A.14})$$

With constraints,

$$\alpha_k \geq 0 \quad i = 1, 2, \dots, n \quad (\text{A.15})$$

A. Support Vector Machines

$$\sum_{j=1}^n \alpha_j y_j = 0 \quad (\text{A.16})$$

The optimal hyperplane is obtained by solving (A.14) and (A.16),

$$\omega^* = \sum_{i=1}^n \alpha_i y_i x_i \quad (\text{A.17})$$

$$b^* = -\frac{1}{2}(\omega^*, x_r + x_s) \quad (\text{A.18})$$

where x_r and x_s are any support vector from each class satisfying,

$$\alpha_r, \alpha_s > 0, \quad y_r = -1, \quad y_s = 1 \quad (\text{A.19})$$

The hyperplane decision function can be written as,

$$f(x) = \text{sgn}((\omega^*, x) + b) \quad (\text{A.20})$$

Considering a complex non-separable dataset, a non-linear mapping of the input space into high dimensional space, H , may enable linear separation,

$$\Phi : \mathbb{R}^n \rightarrow H, x_i \rightarrow \Phi(x_i) \quad (\text{A.21})$$

Therefore,

$$(\Phi(x_1), y_1), (\Phi(x_2), y_2), \dots, (\Phi(x_n), y_n) \in H \times -1, +1 \quad (\text{A.22})$$

Here, the required number of samples increases as an exponential function of n . In SVM, the data is represented in a form of inner product $\langle x_i, x_j \rangle = x_i \cdot x_j$. The inner product in the input space is replaced with the inner product in Hilbert space [133]:

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi(x_i) \cdot \Phi(x_j) \quad (\text{A.23})$$

A. Support Vector Machines

where $K(x_i, x_j)$ is a kernel, a generalised non-linear similarity measure between two feature vectors. The inner product is evaluated directly in the input space by applying the non-linear function Φ . This is referred to as the kernel trick [134]. The goal is to embed data into the Hilbert space and then seek linear relations in both spaces.

The general form of inner products in Hilbert space is defined by the Mercer's condition [135]. If $K : X \times X \rightarrow \mathbb{R}$ is continuous and symmetric real value function on Hilbert space with a square integral function $f \neq 0, \int_x f^2(x) dx < \infty$. then:

$$K(x_i, x_j) f(x_i) f(x_j) dx_i dx_j \geq 0 \quad (\text{A. 24})$$

The appropriate condition to expand $K(x_i, x_j)$ as a uniformly convergent series on $X \times X$:

$$K(x_i, x_j) = \sum_{r=1}^{\infty} \lambda_r \Phi_r(x_i) \Phi_r(x_j), \lambda_r > 0 \quad (\text{A. 25})$$

If K is continuous kernel of a positive integral operator as defined by Mercer's condition, there exists a mapping Φ of an input space into a space where the kernel can be represented as an inner product. The corresponding problem is:

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (\text{A. 26})$$

$$\sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, 3, \dots, n \quad (\text{A. 27})$$

The decision function in higher dimensional feature space is:

$$f(x) = \text{sgn} \left(\sum_{i=1}^n \alpha_i y_i \cdot K(x_i, x_j) + b \right) \quad (\text{A. 28})$$

A. Support Vector Machines

It is said that if kernel function satisfied Mercer's condition, the solution of convex optimisation problem converges to optimal. Some widely used kernel functions that satisfies the previous conditions are illustrated in Table A.1.

Table A.1: Commonly use kernel functions

Kernel	Function
Linear kernel	$\langle x_i, x_j \rangle$
Polynomial	$(\gamma \langle x_i, x_j \rangle + r)^d, \gamma > 0$
Radial Basis Function	$\exp(-\gamma \ x_i - x_j\ ^2), \gamma = \frac{1}{2\alpha^2}, \alpha > 0$
Sigmoid kernel	$\tanh(\gamma \langle x_i, x_j \rangle + r)$

Appendix B

Image Databases

In this appendix, a list of all image databases used in the experiments is presented.

B.1 ESP Image Dataset

This dataset contains 100,000 images from the World Wide Web. These images contain complex scenes and scenarios with large numbers of objects present, such as busy streets, seaside, landscape, office environments etc. Therefore, they cannot be categorised into a particular semantic category. A selection of images used for testing is presented in Figure B.1 and B.2.

B.2 Caltech 101 Image Dataset

Caltech 101 dataset contains a higher level of ground truth based on semantic meaning. Images belonging to the same class illustrate the same concepts however, their visual appearance is different. This dataset consisted of 101 object categories which do not overlap with any other concepts. A selection of images used for testing is presented in Figure B.3 and B.4, which was selected from a number of object categories.

B.3 Corel Image Dataset

Corel image dataset contains a higher level of ground truth based on semantic meaning. Images belonging to the same class illustrate the same concepts; however, their visual appearance is different in practice. The dataset consists of seven concepts, namely, Car, Lion, Tiger, Cloud, Elephant, building and vegetation. A selection of images used for testing is presented in Figure B.5, which was selected from different object categories.

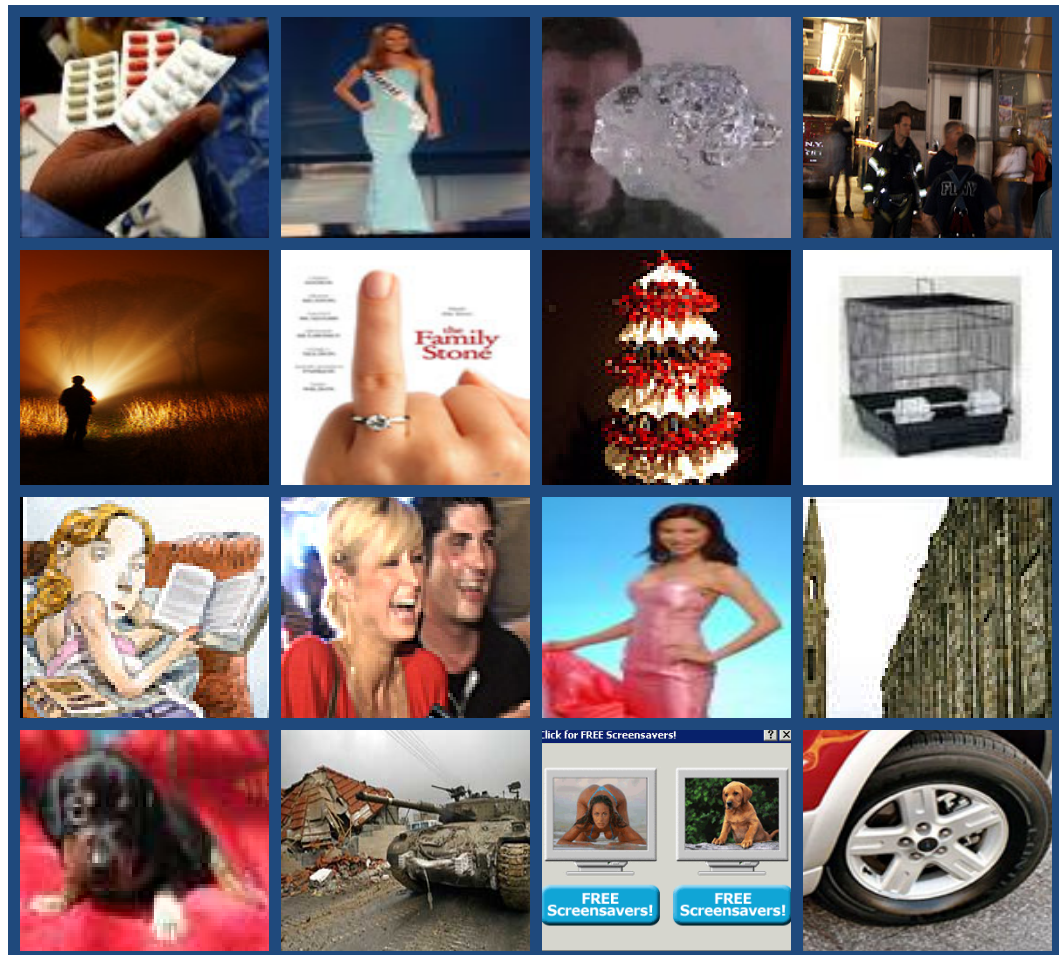


Figure B.1: Representative images for different categories taken from the ESP Image dataset (part 1).

B. Image Databases



Figure B.2: Representative images for different categories taken from the ESP Image dataset (part 2).

B. Image Databases

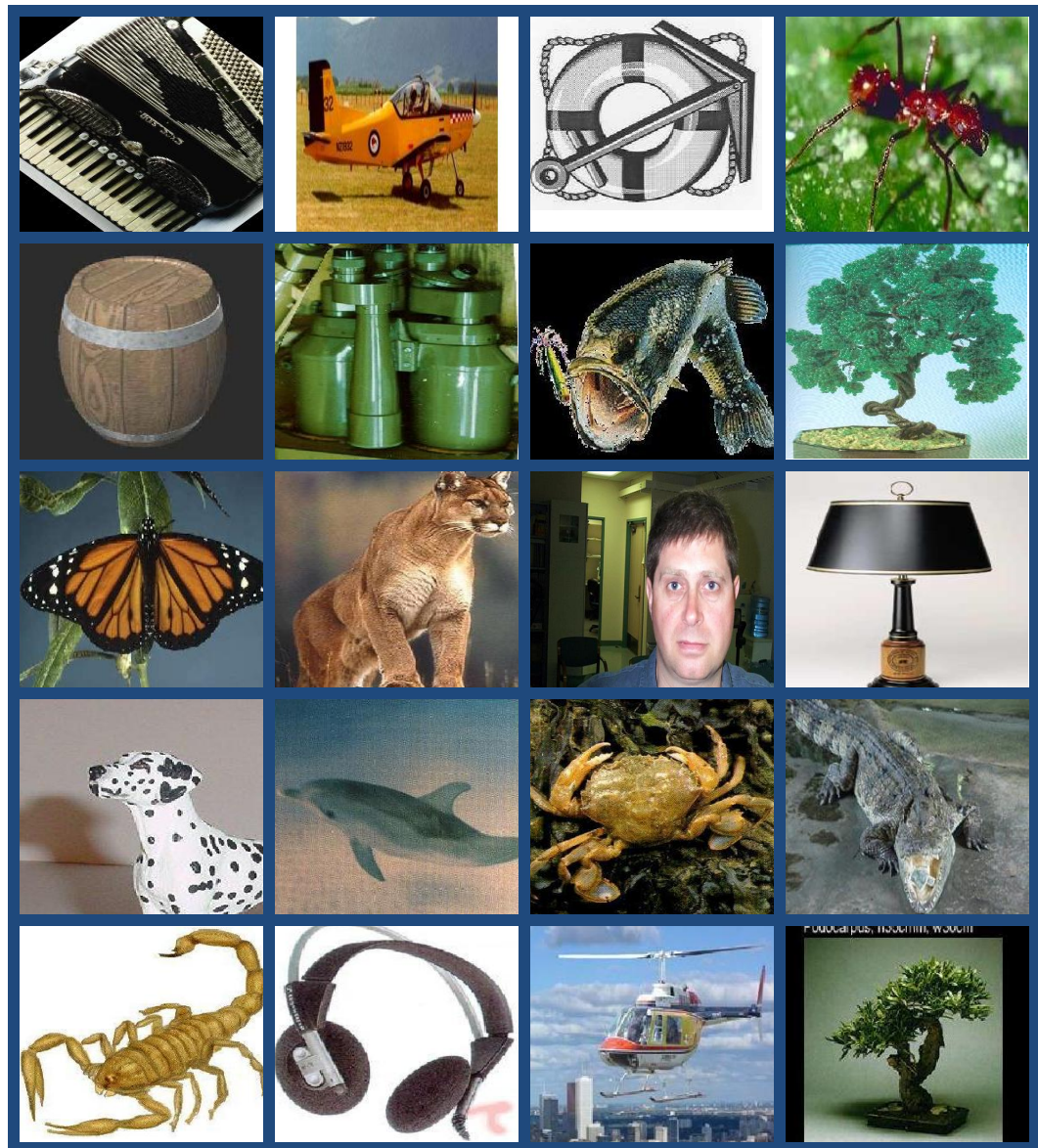


Figure B.3: Representative images for different categories taken from the Caltech Image dataset (part 1).

B. Image Databases

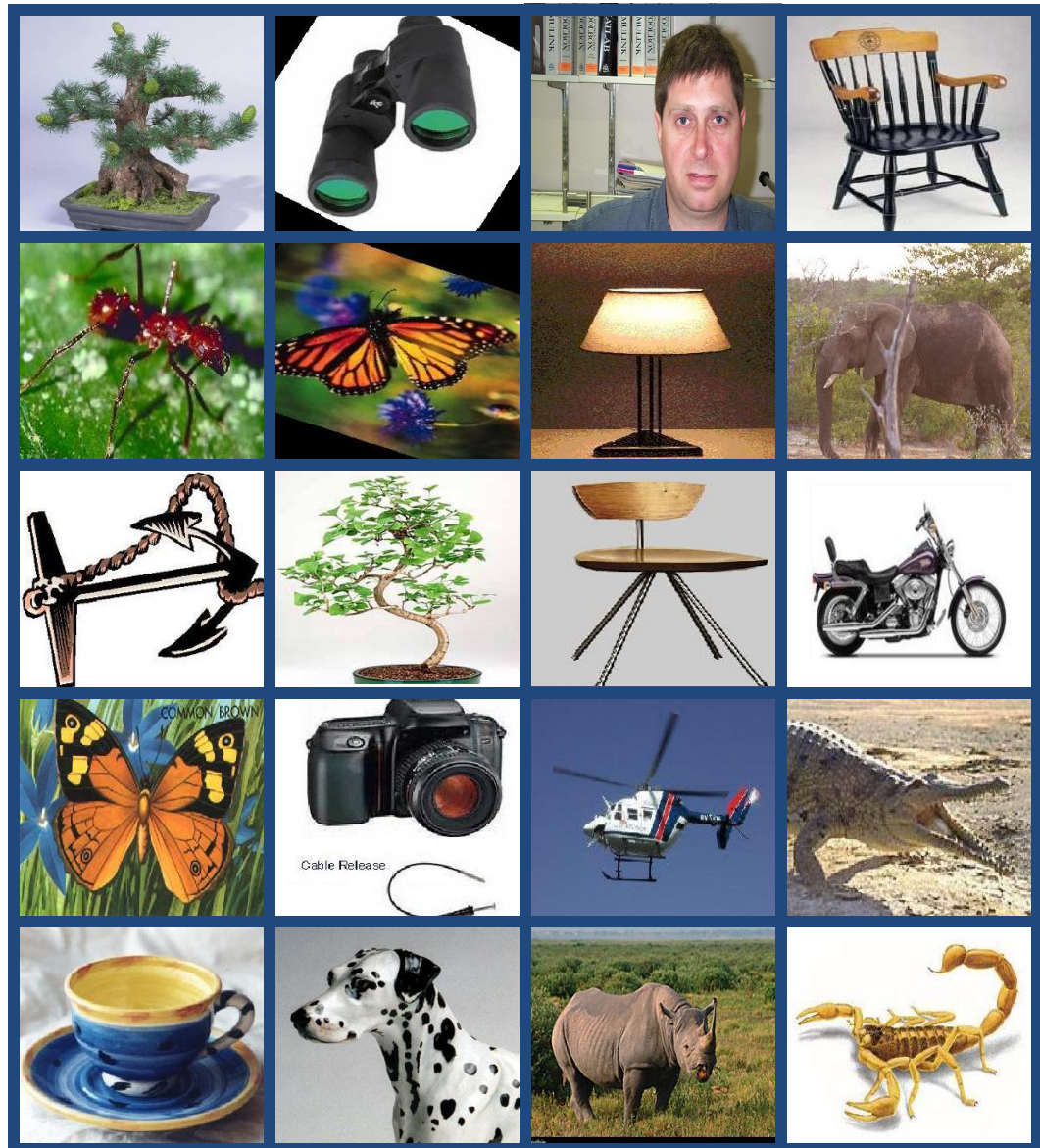


Figure B.4: Representative images for different categories taken from the Caltech Image dataset (part 2).

B. Image Databases



Figure B.5: Representative images for different categories taken from the Corel image dataset.

Appendix C

Questionnaire on Usability Test

In this appendix, a template of the questionnaire used for usability test is given.

C.1 Template of the Usability Test

Survey on ESP, Phetch, INT-1 and INT-2 game interfaces

Please take part in this very quick survey, as your responses will help us understand the players' views regarding different games and their interfaces. Please play the four games provided (ESP, Phetch, INT-1 and INT-2) before answering the following questions. The survey should take no more than 10 minutes to complete, after which all respondents will be entitled to play games for free as much as they want.

1. Age :
2. Sex :
3. Occupation :

C. Questionnaire on Usability Test

Please enter your scores into the table below, using the following values: 1 – Very low, 5 – Moderate and 10 – Very high

	Excitement factor	Attractiveness	Enjoyability	Difficulty in game play
ESP				
Phetch				
INT-1				
INT-2				

Thank you very much for your participation!

Appendix D

Outcomes of Analysis of Variance (ANOVA)

In this appendix, outcomes of all conducted ANOVA tests are presented and summarised in an ANOVA table. This table contains columns labelled as ‘Source of variation’, ‘Sum of Squares’, ‘ DF ’ for degrees of freedom, ‘Mean square’, ‘ F Statistic’ for F -ratio, and ‘ p ’ for significance among the data.

The tables given below present the outcomes of ANOVA tests conducted during the usability test.

D.1 Excitement

A one-way ANOVA was conducted with respect to the reported excitement levels across different age categories. The results are shown in Table D.1 below. The test was conducted using 1760 data samples, obtained from 440 game players in each of the 4 age categories (440×4).

D. Outcomes of Analysis of Variance (ANOVA)

Table D.1: ANOVA results related to excitement levels across different age categories.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
Excitement	520.4	3	173.5	32.71	$p < 0.0001$
Residual	9313.1	1756	5.3		
Total	9833.5	1759			

A one-way ANOVA was conducted with respect to excitement levels across all four games tested. The results are shown in Table D.2 below. The test was conducted with 1760 samples, yielded by 440 game players that tested each of the four different games (440×4).

Table D.2: ANOVA results related to excitement levels across the four games tested.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
All games	442.5	3	147.5	27.58	$p < 0.0001$
Residual	9391.0	1756	5.3		
Total	9833.5	1759			

D.2 Addiction

A one-way ANOVA was conducted with respect to addiction levels across different age categories. The results are shown in Table D.3 below. The test was conducted with 1760 data samples.

Table D.3: ANOVA results related to addiction levels across different age categories.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
Addiction	36.7	3	12.2	1.77	0.1513
Residual	12137.5	1756	6.9		
Total	12174.2	1759			

D. Outcomes of Analysis of Variance (ANOVA)

A one-way ANOVA was conducted with respect to addiction levels across the four games tested. The results are shown in Table D.4 below, based on 1760 data samples.

Table D.4: ANOVA results related to addiction outcomes across all four games.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
All games	334.2	3	111.4	16.52	$p < 0.0001$
Residual	11839.9	1756	6.7		
Total	12174.2	1759			

D.3 Enjoyability

A one-way ANOVA was conducted with respect to perceived enjoyment reported by different age categories. The results shown in Table D.5 below, are based on 1760 data samples.

Table D.5: ANOVA results related to perceived enjoyment reported by different age categories.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
Enjoyability	3.5	3	1.2	0.24	0.8675
Residual	8391.5	1756	4.8		
Total	8395.0	1759			

A one-way ANOVA was conducted for perceived enjoyment reported for the four games tested. The results are shown in Table D.6 below, based on 1760 data samples.

D. Outcomes of Analysis of Variance (ANOVA)

Table D.6: ANOVA results related to perceived enjoyment reported for the four games tested.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
All games	324.8	3	108.3	23.56	$p < 0.0001$
Residual	8070.1	1756	4.6		
Total	8395.0	1759			

D.4 Game difficulty level

A one-way ANOVA was conducted with respect to perceived game difficulty level across different age categories. The results shown in Table D.7 below are based on 1760 data samples.

Table D.7: ANOVA results related to game difficulty level across different age categories.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
Difficulty in game play	231.0	3	77.0	12.27	$p < 0.0001$
Residual	11014.2	1756	6.3		
Total	11245.2	1759			

A one-way ANOVA was conducted with respect to perceived game difficulty level across the four games tested. The results are shown in Table D.8 below, based on 1760 data samples.

Table D.8: ANOVA results related to game difficulty level across all four games.

Source of variation	Sum of Squares	<i>DF</i>	Mean square	<i>F</i> statistic	<i>p</i>
All games	4169.5	3	108.3	344.9	$p < 0.0001$
Residual	7075.7	1756	4.0	2	
Total	11245.2	1759			