國立臺灣大學電機資訊學院資訊工程學研究所
碩士論文
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

具有生產力的遊戲正確率分析
-以建築地標標註爲例

Accuracy Analysis in Productivity Games:
A Case Study on Landmark Annotation

陳麗徽
LI-HUI CHEN

指導教授：許永眞 博士
Advisor: Jane Yung-jen Hsu, Ph.D.

中華民國九十七年六月
June, 2008


國立臺灣大學
敎育學程研究所

Acknowledgments

這篇論文的完成，要感謝的人很多，首先要感謝的是我的指導教授：許永眞教授，在老師的指導下，我學習到了做研究的方法，老師總是親切的和我們討論、聊天，老師對研究的熱誠，以及許多生活上的觀點，教導了我、使我受益良多。

接下來我想要感謝實驗室的所有同學、學長姊、學弟妹們，身在一個實驗室這個大家庭中，同學之間都能互相幫助、互相監督學習，是我非常感謝的，其中我更要謝謝嘉涓和怡靜，謝謝妳們鼓勵我、回答我許多問題、以及在口試前辛苦的陪我改投影片、演練，幫助我釐清口試的重點; 謝謝跟我一起口試、口試前一起在博理館努力的好伙伴文芝和祖佑，謝謝你們和我討論以及幫助我解決了許多的問題; 謝謝育仲眞心的安慰及鼓勵我，謝謝家峻和現在身在新加坡的啓嘉的加油和鼓勵，我們是一起辛苦打拼的好伙伴。除此之外實驗室的學長姊和學弟妹們也都常常鼓勵我們、擔心我們、幫我們加油打氣，幫助我們處理了雜事，使得口試和論文能順利的完成，眞的很感謝他們。

最後，我想要謝謝我的家人、朋友、我碩二宿舍的室友姿呈、以及和我一樣辛苦的準備碩士論文的大學同學們，謝謝你們的陪伴、關心、和鼓勵，有你們在身邊我才能心無旁騖的準備論文的事宜，眞的非常非常感謝你們，你們是我心中永遠的依靠。


國立臺灣大學
敬品勵學・愛國愛人

Abstract

Despite impressive advancement in computer technology, there are still some problems that humans can solve efficiently but current computer programs can not. Image recognition and annotation are examples. *Human computation* is a new research area that focuses on this kind of problems. This thesis aims to explore the power of human computation and shows how humans could help solve problems that are hard for computers. We propose a two-player online human computation game, ImageHunter, to achieve the task of annotating for a collection of landmark images on the Internet. Moreover, we address on the quality analysis for the data collected by the game. We propose *confidence evaluation* instead of times accumulation to estimate the accuracy of the data. Experiments involving 28 players have been conducted. The preliminary results demonstrate that the game mechanism is reasonable and with confidence evaluation mechanism the accuracy improves effectively.

Keywords: Human Computation Games, Productivity Games, Accuracy Analysis, Confidence, Confidence Evaluation


國立臺灣大學
敎育學程研究所

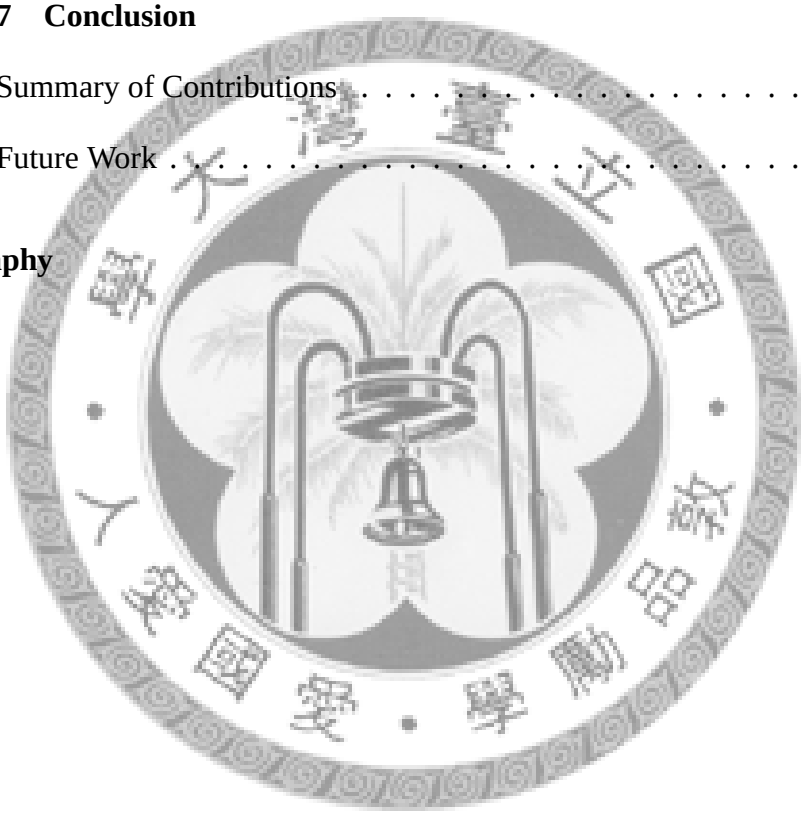
摘要

儘管電腦科學已有了相當長遠的發展，還是存在著一些問題是人們可以輕易解決，但是電腦卻沒有辦法的，像是圖片辨識(image recognition)和圖片標註(image annotation)。而人力計算(human computation)是專門在研究這類問題的一個新的研究領域，本篇論文即是利用人力計算的方法，讓人們來幫助解決這些電腦沒有辦法有效解決的問題。

在這裡我們提出了一個兩人網路線上遊戲ImageHunter 來幫助完成網路上建築地標圖片的標註(landmark image annotation)。除此之外，我們著重在分析遊戲中收集到的資料的品質分析。我們提出信任值估算(confidence evaluation)的方法取代了次數累積(times accumulation)的方法來評估資料收集的正確性。我們請來了28位受測者來進行實驗，實驗結果驗證了遊戲機制的合理性，以及信任值估算的方法可以有效地提升收集到的資料的正確性。

關鍵詞： 人機協力演算法、具有生產力的遊戲、正確性分析、信任值、信任值估算

Contents



List of Figures

List of Tables


國立臺灣大學
敎人學問品德
愛國愛學

Chapter 1

Introduction

1.1 Background

According to the development of Internet, great amount of information are exposed to people everyday. The content includes text, pattern, images, audios, and some other multimedia content. How to filter, manage and acquire ideal information from the vast is a big problem. Producing explanatory descriptions of the content would be a method. A description might be a short paragraph, a sentence, a string, or a word. A shorter and a more precise one would be more welcome. Tagging, annotation, and semantic annotation are famous techniques which usually produce short and meaningful comments for content. However, these works are difficult for current computers to solve automatically, but easy and trivial for humans.

Traditionally, many research aims on collecting these kinds of knowledge by human. They hired exports or expected volunteers to manually enter the information. However,

manually entering the data is quite a tedious job and needs a lot of efforts. In 2004, Luis von Ahn proposed the idea of human computation games to help solve the problems. He designed games which produce image annotation data while people playing. Fun is provided as the incentive to attract humans engage and therefore annotating images. However, since human computation games collect the information from mass, there might be some mistakes. Methods to measure the quality of the results are essential. In this thesis, we want to propose a feasible mechanism to measuring the results in human computation games.

1.2 Research Objectives

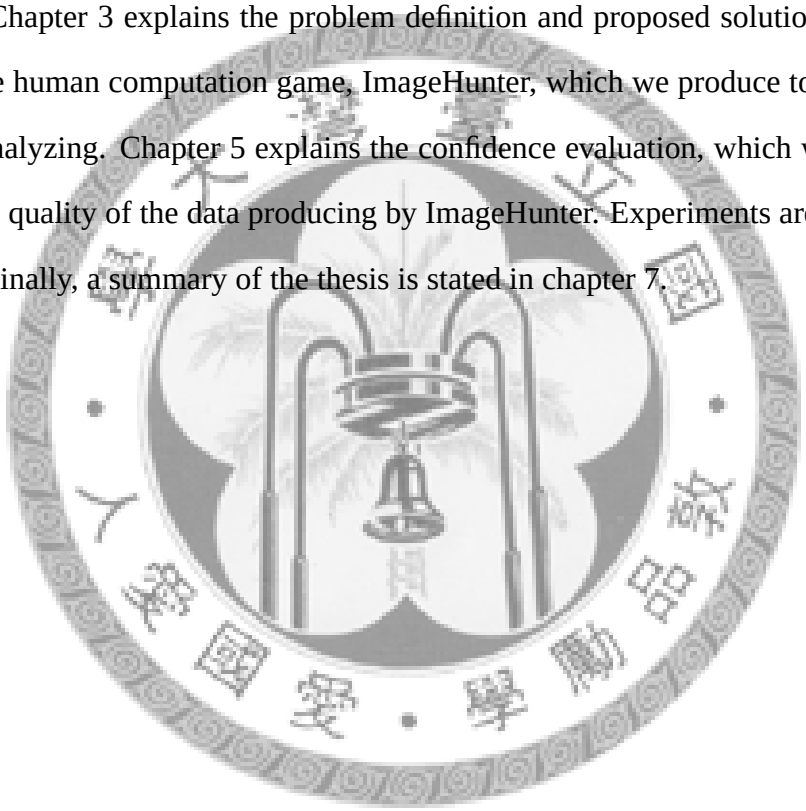
Previous works judge the result of the games only by the accumulated times. In ESP game, a pair of players is given an image and types the describing strings about the image in each round. When they type the same string, they get the points and the string would be recorded. For each image, if a string is attached to it twice, the string would be created as a label of the image. However, there might be some problems about this method. First, how many times a label is attached guarantees to be truth. Second, a label from different people might have different quality. For example, a label from an export and a label from an arbitrary people would be totally different. Without the consideration of this information could be very questionable.

In this thesis, we propose a mechanism avoiding the above situations. We propose a confidence measuring mechanism to estimate the quality of the data producing by the game. In this mechanism, we don't rely on the number of the accumulated times. More-

over, we estimate how much a player could be trusted and take that into the consideration.

1.3 Thesis Structure

This thesis is structured as follows. Chapter 2 gives a survey on collaborative information repositories, human computation and quality analysis for collaborative information repositories. Chapter 3 explains the problem definition and proposed solution. Chapter 4 describes the human computation game, ImageHunter, which we produce to collecting data for our analyzing. Chapter 5 explains the confidence evaluation, which we propose to estimate the quality of the data producing by ImageHunter. Experiments are described in chapter 6. Finally, a summary of the thesis is stated in chapter 7.




國立臺灣大學
敎學勵品
愛國人

Chapter 2

Related Work

This chapter will give a brief survey on three main fields, which includes collaborative information repositories, human computation and quality analysis for collaborative information repositories.

2.1 Collaborative Information Repositories

The collaborative, web-based creation of knowledge and information repositories are popular and become a trend. The pattern for building the information repositories is to attract many people do their effort to create and enter their common sense information and knowledge. The most successful and representative repository is Wikipedia. Wikipedia is a freely available, multilingual, open content encyclopedia, and it contains millions of articles that cover a wide swatch of knowledge. Wikipedia's articles are collaboratively written by volunteers and anyone with access to the Internet could edit and modify nearly

all of its articles. However, there are some issues for building such collaborative information repositories.

The first is how to encourage people spend their time and effort to contribute on building the knowledge system. Some research hires experts to annotate and enter the knowledge. One of the famous examples in AI field is the Cyc project, which attempts to assemble a comprehensive ontology and database of everyday common sense knowledge [6, 5]. Most of the other research relies on mass collaboration. Some of them provide financial incentives, and some of them provide the good for the mass. For example, the size and diversity of the information on Wikipedia is the cause that it could attract many people to visit and write articles on it. In addition, recently some research provides "fun" as an incentive to attract people contribute their knowledge willingly. This aspect will be discussed detailing in the following section 2.2.

Another issue is how to measure the quality of information created by the mass on the Internet. Without supervisors or detectors monitor on every items constructed by the people, there might be some mistakes in information repositories. As a result, in order to build up a useful information repository, how to ensure the quality of the information is an important issue. This will be discussed later in the section 2.3.

2.2 Human Computation

In 2003, Luis von Ahn proposed the ideas about human computation. He works on inventing novel techniques for utilizing the computational abilities of humans to complete tests that humans can easily pass, but current computer programs can't pass. A famous

technique is CAPTCHA, a challenge-response test in computing to determine that the response is not generated by a computer [14]. Since 2004, he has introduced several games that could create valuable outputs when people playing them [15, 16, 17, 18]. These games provide fun as an incentive for the people spending their time on it and achieve the goal of building information repositories.

2.2.1 The ESP Game and Peekaboom

The ESP Game is an online interactive game which intends to label all images on the web [15]. Two partners are randomly assigned among all the players in the game and are given an image every time. From the player's perspective, the goal of the ESP game is to guess what their partner is typing on the image. Once both players have typed the same string, they get some points and move on to the next image. In this process, the string that the players both type would become a new label of the image. This game help label the majority of the images on the World Wide Web by a funny way instead of tedious manual labeling.

By the ESP game, images are annotated with information about what objects are in the image. Moreover, Peekaboom is another online interactive game which achieves getting information about where an object is located in the image [18]. In Peekaboom, two random players from the Web participate by taking each of the two main roles in the game: ”Peek” and ”Boom”. Peek starts out with a blank screen, and Boom starts with an image and a word related to it. The goal for Boom is to reveal parts of the image relevant to the word and for Peek is to guess the word by observing parts of the image revealed by Boom. The reveal parts would be recorded and the information about where the object

Figure 2.1: The ESP Game is a two-player interactive game. A label of an image is created when two partners type the same string on it.

is located is therefore acquired while playing. These two projects don't need volunteers to manually labeling all images on the Web; they expect all images to be labeled because people want to play the games.

2.2.2 Other Applications

After the success of the ESP game and Peekaboom, many other research have been proposed under the idea of human computation. Examples include works on collecting semantic annotations of music [13], Common Consensus [7] and Verbosity [17] for collecting commonsense information, Phetch [16] which aims on collecting explanatory descriptions of images, and PhotoSlap [3, 1] which achieves semantic clustering and therefore accomplishes photo annotation. The tasks aimed by these works are all difficult for com-

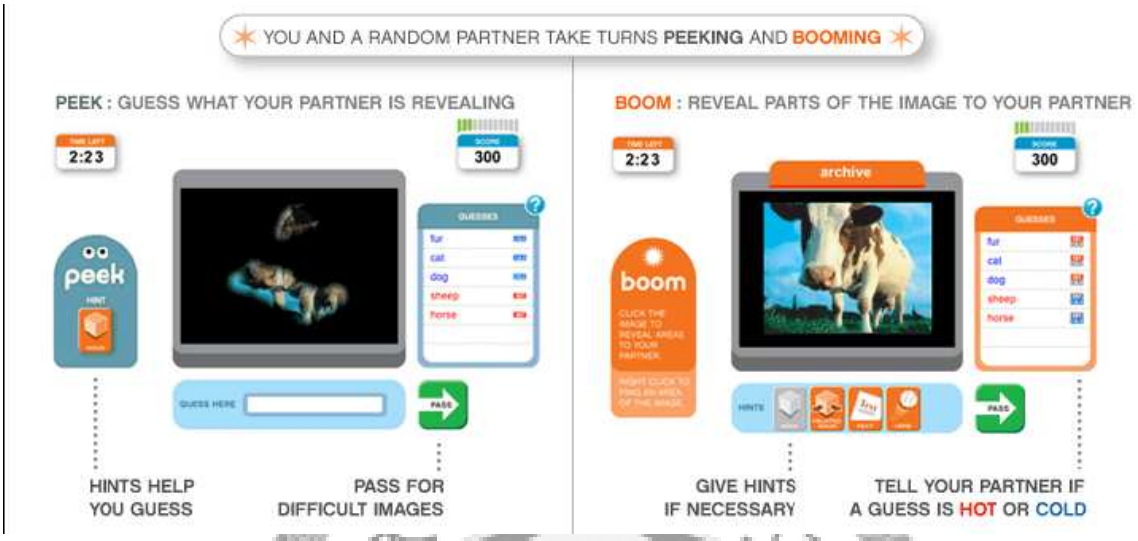


Figure 2.2: Peekaboom is a game which achieves getting information about where an object is located in the image.

puters to complete. By the idea of inventing games, these works could successfully attract people spending times on them and completing the tasks of collecting meaningful information.

2.3 Quality Analysis for Collaborative Information Repositories

The Internet offers people new opportunities to share their knowledge, personal thoughts, opinions, and experiences to the global community of Internet users. This behavior could be fun, informative, but might be risky. Is the advice of a self-proclaimed expert reliable? To solve the problem, reputation Systems are proposed [11]. Most reputation systems

are based on users' feedback on other's contributions or behavior [11, 2]. A very famous example is the rating system of eBay[1], where each buyers and sellers rate for each other after the transactions. By the use of feedback, these reputation systems could provide some reasonable reference for the future activities.

Besides building reputation systems by users' feedback, some research about reputation systems for Wikipedia are based on the fare of the articles or contributions [19, 10, 12]. The amount of text being added and deleted would be used for evaluated the trust value of an author and the article. In some works, Dynamic Bayesian Networks are used to model the evolution of trust level over the versions [19]. This approach could get the prediction of an article's quality. In addition, the reputation value of an author could also be predicted by analyzing the revision history [12]. According to these online reputation mechanisms, revealing the result could guide trust decisions for activities on the Internet.

[1] http://www.ebay.com

Chapter 3

Semantic Annotation

Many kinds of metadata could be gathered from images including information about people, objects, locations, time, and activities. This semantic information is the meaningful and interesting part of images. With the information, image managing, sharing, and searching could be improved significantly. Furthermore, research on computer vision and personal social network analysis[4] could also benefit from it. However, automatically generating the annotations of these semantic metadata is still a problem that is difficult for computers to solve. Therefore, this thesis would like to propose a solution for generating these semantic annotations for images.

3.1 Annotating Landmark Images on the Web with their Proper Names

Considering the metadata about objects, there are common objects and also unique objects which are nominated with proper names. When people want to search for images of unique objects such as a famous landmark, usually people would query by using the proper name of it. As a result, gather this information would help a lot and better the search results. Therefore, we address on collecting the metadata of unique objects. In this thesis, we deal with collecting the metadata of landmarks. We want to annotate images of landmarks on the web with their proper name. For instance, annotating images of the official home of the President of the United States by using "The White House" is the goal in this thesis.

3.2 Problem Definition

In this thesis, we want to annotate images of landmarks on the web by using landmarks' proper names. Here we annotate the images according to the landmark image clues which are provided as input of the problem. Followings we will give a definition of what is a landmark image clue and defines the problem.

Definition 1 Landmark Image Clue

A landmark image clue $c = (i, w)$ *is a semantic pair where* i *is an image contains a landmark whose proper name is* w*.*

Definition 2 The Problem

Consider the set of images on the web $\Delta = \{i_1, ..., i_N\}$*. Given a set of landmark image clues* $C = \{c_1, ..., c_n\}$*, and a set* L *as the union of the proper names* w *in* C*. The problem is to extract a set* $\Psi_i \subseteq L$ *for each image* i_i*, where each element represents the proper name of a landmark appearing in* i_i*.*

The image clues are necessary in our project. The annoation produced by our system are basically from the proper names of the clues. Second, the clues are the ground truth in our system. It would be used for the scoring mechanism in ImageHunter and for the confidence evaluation proposed here. The details would be described in later sections.

3.3 Proposed Solution

Image recognition and annotation are trivial for humans, but continue to be a challenge for computer programs. As a result, utilizing human processing power is essential to solve this problem. Manually annotation is widely used for solving this kind of problems before; however, it is boring and will take a lot of time to annotate a great amount of images. Therefore, we follow the idea of "Human Computation" and propose a game to achieve the task.

In addition, since there might be some errors among knowledge repositories built up by the mass, we analyze the accuracy of the data collecting in our system. We use confidence evaluation to measure the quality of the data. The process of our system are shown in figure 3.3.

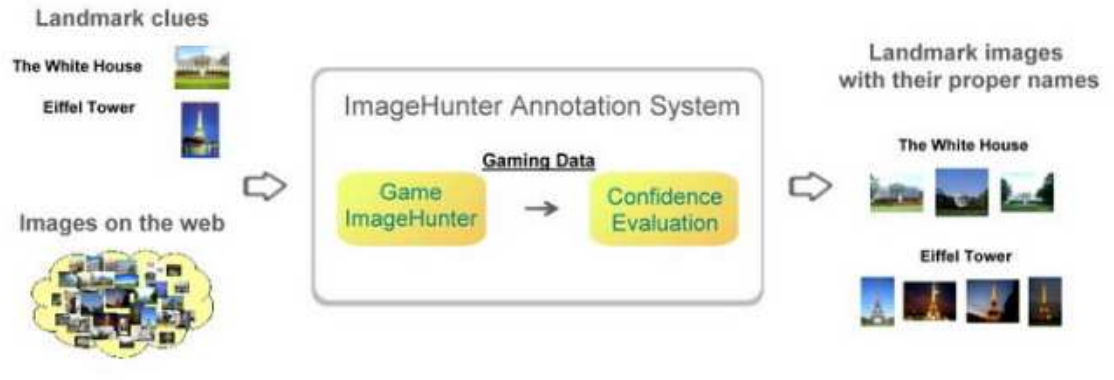


Figure 3.1: The process of our system: ImageHunter Annotation System

3.3.1 A Human Computation Game: ImageHunter

Here we want to let people annotate landmarks in images by using the proper names. An intuitive procedure is presenting each of the landmark images to people and requiring them to directly type the proper names of the landmarks. However, this would not be easy for everyone because many people do not know a lot of landmarks and remember the names of them. Since the game is proposed to be played by the mass, we need to design another procedure which could be completed by most people straightforwardly. Accordingly, landmark image clues are necessary as an input of the game. We reveal the visual appearance of a landmark for people and translate the task into an image recognition and comparison task.

We design a game “ImageHunter” which could achieve annotating landmarks on im-

ages. ImageHunter is an on-line two-player game. While playing ImageHunter, a pair of players is given one specimen image with a landmark and many other candidate images. The goal is to hunt among candidate images for the ones which contain the same landmark as in the specimen image. Several features are used to make the game captivating, including score-keeping and timed-response. There is a brief explanation about the design of ImageHunter in chapter 4.

3.3.2 Confidence Evaluation

To ensure the accuracy of the final annotation produced by our system, we proposed a confidence measuring mechanism to analyze the quality of the gaming data produced by ImageHunter. We calculate confidence values for each round of the game. Gaming data produced by trustworthy players and gaming data produced by pranksters would be evaluated separately. The detailed explanation about the mechanism of confidence evaluation would be described in chapter 5.

Chapter 4

Human Computation Game Design

Follow the idea of human computation games proposed by Luis von Ahn [15, 16, 17, 18], we design a game “ImageHunter” to help annotate landmark images with the proper names. The design and the mechanism of ImageHunter are introduced in this chapter. First we will give a brief description about the game rules and the scoring mechanism, and talk about the general design principles of games applied on ImageHunter.

4.1 Game Mechanism

ImageHunter is an on-line two-player game. Partners are randomly assigned from among all the players; players can’t know who is their partner and also can’t communicate with the partner. In each round, each pair of players is given one specimen image with a landmark and many other candidate images. The goal is to search and “hunt” for images which contain the same landmark as in the specimen image.

4.1.1 Game Rules

The view of our game ImageHunter are shown in figure 4.1.1. Players are asked to choose among sixteen candidate images for the ones that contain the same landmark as in the specimen one. Different time limits are set for different game levels. Players should hunt for images within the time limit. A level with a shorter time limit would be more difficult but more exciting. Partners are given points for every image chosen by both of them. The permutations of the sixteen candidate images among players are different. Without the permit of communication, players could only choose the correct ones in order to gain the points. 4.1.1

In ImageHunter, since a specimen of the appearance of a landmark is given to the players, players only need to have the ability of pattern recognition or pattern comparison in order to choose the target images. Players don't need to have the knowledge of these landmarks but help annotate the proper names of these landmarks. As a result, ImageHunter could be played by almost the mass.

4.1.2 Scoring Mechanism

In ImageHunter, two partners are given points for every image that is chosen by both of them. However, there would be some images with ground truth hidden in the candidate images. If a player chooses an image with ground truth that it is an incorrect one, the player would loose points owning to the punishment.

Except getting points for the match images, we will give player some bonus points according to the time they leave over and the ratio of match images to all chosen images.

Figure 4.1: The view of our game ImageHunter. Players are asked to choose the images which contains the same landmark as the specimen image.

Figure 4.2: Partners are given the same candidate images but with different permutation.

Three kinds of scoring mechanisms are listed below:

- **Basic Score (s_{basic})**

  A player are given points for choosing images that are also chosen by his/her partner or is ground truth positive except for the images that is ground truth negative; these choices are defined as correct choices $C_{correct}$ here. Moreover, a player will loose a lot of points for choosing images that is ground truth negative; these choices are defined as wrong choices C_{wrong} here. $C_{correct}$ and C_{wrong} are defined as follows:

  $$C_{correct} = I_{player} \cap (I_{partner} \cup \mathcal{I}_{positive}) - \mathcal{I}_{negative}$$

  $$C_{wrong} = I_{player} \cap \mathcal{I}_{negative}$$

  Where $\mathcal{I}_{positive}$ and $\mathcal{I}_{negative}$ are the set of images that is ground truth positive and negative respectively; I_{player} is the set of images chosen by the player and $I_{partner}$ is the set of images chosen by his/her partner.

  The basic score s_{basic} is defined as follows:

  $$s_{basic} = |C_{correct}| * s_{correct} - |C_{wrong}| * s_{wrong}$$

  Where $s_{correct}$ is the score players would get for each correct choices, and s_{wrong} is the amount of score players would loose for each wrong choices.

- **Time Left Score (s_{time})**

  A time limit is set in each round of the game. Players who finish the task within a shorter time would be given more bonus points. The time left bonus points are:

  $$s_{time} = s_{basic} * \frac{t_{left}}{t_{total}}$$

Where t_{left} is the time left by the player and t_{total} is the time limit in the round.

The time left score is based on $s_{time} = s_{basic}$ to prevent players select images arbitrary. For example, a player might select images arbitrary to finish quickly and wants to get higher time left score. However, since the time left score is based on $s_{time} = s_{basic}$, the score might be very low without selecting correct images. As a result, players who want to get high score must select images correctly.

- **Match Ratio Score (**$smatch$**)**

  Match Ratio is defined as

$$R_{match} = \frac{|C_{correct}|}{|I_{player} \cup I_{partner} \cup \mathcal{I}_{positive}|}$$

  The match ratio bonus points are:

$$s_{match} = s_{basic} * \frac{|C_{correct}|}{|I_{player} \cup I_{partner} \cup \mathcal{I}_{positive}|}$$

  Why we want to consider this? Because the basic score depends on how many images the partners both chosen, players might choose as many images as possible. Nevertheless, we want them to choose the correct ones. By considering match ratio, more precise answers as their partners would get more bonus point. As a result, this could lead them to choose answers that will also be chosen by their partners.

4.2 General Design Principles of Games

Challenge is a key aspect of many successful games. This could be translated into different things, including obvious goals, variable difficulty level, multiple level goals, hidden

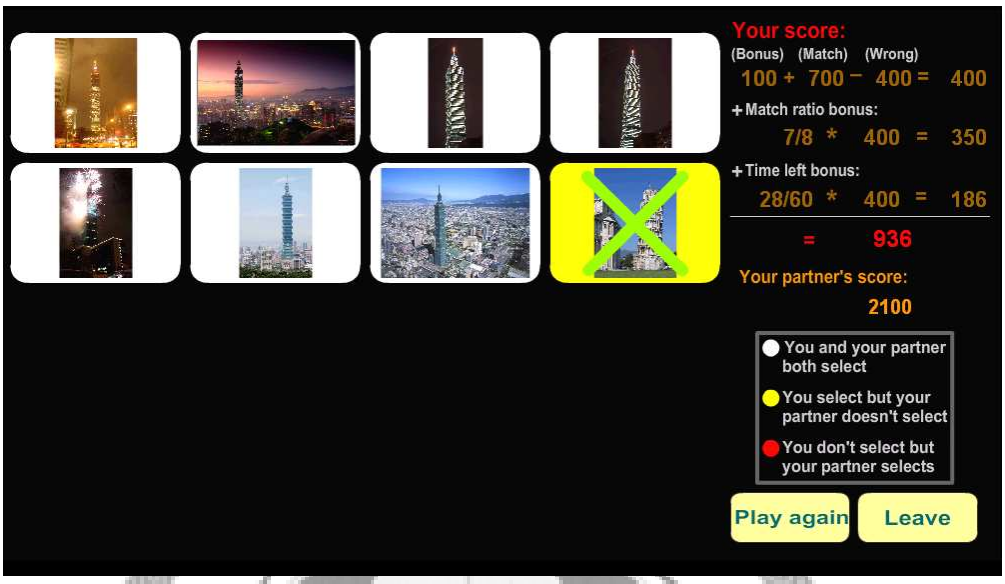


Figure 4.3: After the playing of each round, players are given points for the match images and loose points for choosing the images which is ground truth incorrect. Moreover, players are given bonus points according to the time left and the match ratio.

information, and uncertain outcome like randomness [8, 9]. These are all elements that could make games be captivating. We could take advantage of these methods for introducing challenge into a game, in turn encouraging continued play and enjoyment.

In ImageHunter, we adopt several methods to make it be interesting and enjoyable. The methods we adopt are as follows:

- **An Obvious Goal**

  The goal in ImageHunter is pick the images with the same landmark as the specimen image. To achieve the goal only need to contain the ability of pattern recognition or pattern comparison. Moreover, players should take their partners' behavior into consideration. Since there is no way to communicate with their partners, the goal

of players would be pick the most precise and possible images that would also be picked by their partners.

- **Speed Response**

  In our game, a time limit is set in each round. Players should complete the task within the time limit. Moreover, players who complete the task earlier would have more bonus points. With this method, players would feel more interested and excited while playing the game.

- **Score Keeping**

  Players would get points according to their performance in each round. Simultaneously, we will record the highest score a player have ever had and the accumulated score of a player. Also, we maintain billboard of top scores and tip accumulated scores which would encourage players play again to get higher scores.

- **Variable Difficulty Levels**

  Different levels with different time limits are provided in our game. After players become more familiar with the game and improve their skills, they could challenge a higher level and the competing would be more intense. Players could get much honor for winning in higher levels.

- **Randomness**

  A game is boring if the outcome for a player is always the same, for example, certain to win or certain to lose. As a result, we use randomness to make our game uncertain and more interesting. First, partners are randomly assigned from among

all the players. Second, for each round, the target landmark is randomly chosen from all the landmark clues. Third, even for the same landmark, the candidate images are randomly assigned from all the possible images. These uncertainties would make the outcome of games be different and not always fixed.

Chapter 5

Confidence Evaluation

First, we will show that players would do the desired action under the mechanism of our game. However, players might still be incautious and make some mistakes while playing games. Therefore, we propose a confidence evaluation mechanism to measure the actions of players, and we will describe the mechanism detailed in the second part.

5.1 Rationality of ImageHunter

The objective of productivity games is to attract human beings play games and produce useful information simultaneously. Therefore, the mechanism of the game should lead players do the actions that could produce correct information. To achieve this goal, images with ground truth are concerned while choosing candidate images. Moreover, the scoring mechanism is also adjusted in ImageHunter.

5.1.1 Candidate Images

In each round, a target image and a set of candidate images are given to a pair of players. How to generate the candidate images is an important issue. If we just randomly choose images from database, the game would become just a statistical one: more people select the images, than more possible to be the true ones with the target landmark. Besides, how to give the score would also be a problem. If we give players points for each image that they both select, they could select all images and have the highest expected value of score. This would become a big problem that we couldn't produce any useful information through the game.

To avoid the above situation, we use the input image clues which contain the ground truth as the resource of the candidate images. We generate the candidate images $I_{candidate}$ as follows:

- **Images which are Ground Truth Positive**

  A set of images $\mathcal{I}_{positive}$ with ground truth that contains the same landmark as the target image.

- **Images which are Ground Truth Negative**

  A set of images $\mathcal{I}_{negative}$ with ground truth that don't contain the same landmark as the target image.

- **Images without Ground Truth**

  A set of images $I_{unknown}$ that don't have ground truth and are randomly chose from database. These are the ones that need to be distinguished whether contain the target landmark or not.

We contain images with ground truth in candidate images, and this information could be used to avoid invalid actions. With the information, we could punish players while they select images which are ground truth negative. We could also give players points for selecting ground truth images even if their partners don't select the images. With the information, we could check whether players do the right things and evaluate the collected data.

5.1.2 Scoring Mechanism

As mentioned in section 4.1.2, we give players points on three aspects. Some aspects are for interesting, some aspects are for directing rational players to do the desired actions. Here we will give a brief explain on how to direct them to do desired actions.

Basic Score

First, we will discuss about the basic score s_{basic}. For each correct selection(the same selection as the partner or ground truth positive images but not ground truth negative images), players would get score $s_{correct}$. For each wrong selection(selecting ground truth negative images), players would loose score s_{wrong}. As a result, for a random selection, the probability of selecting a wrong one is

$$p_{wrong} = \frac{|\mathcal{I}_{negative}|}{|I_{candidate}|}$$

The probability of selecting a correct one is at most

$$p_{correct} = 1 - \frac{|\mathcal{I}_{negative}|}{|I_{candidate}|}$$

Therefore, the expected value of the random selection is

$$E(s_i) \leq p_{correct} * s_{correct} - p_{wrong} * s_{wrong}$$

To avoid random selection, the we adjust $s_{correct}$ and s_{wrong} to make the expected value of a random selection less than or equal to zero. In this way, rational players who want to get high scores would not choose randomly and do the acitons we desired: choose images which really contain the target landmark.

Time Left Score

Second, since this is a game with the challenge of reaction time, a player who completes tasks quickly should be given some bonus to be fun. Therefore, time left score s_{time} which would give scores according to the time left after playing is included in ImageHunter. In many games, time left score is just depends on how much time left and give points for each second for instance. However, there would be some problems for this kind of mechanism in ImageHunter. In ImageHunter, players might unreasonably and quickly select images and press finish to gain time left score. This would not be desired by us because this would lead to incorrect collections. As a result, we give time left score according to the basic score.

$$s_{time} = s_{basic} * \frac{t_{left}}{t_{total}}$$

In this way, players who arbitrary select images and gain no basic scores also wouldn't get any time left score. Players should follow the restriction that they should select images correctly and then could try to accelerate their recognition speed. As a result, players would do the actions that are desired by us.

Match Score

The goal of considering match score is to let players consider their partners' action and increase some fun activities. While a player usually being conservative/active and his/her partner being active/conservative, they might change their actions. For instance, if his/her partner could select images which contain target landmarks in a small and unclear region, it might make the player pay more attention next round. In the case, if two players could get high match score, they would be very excited and satisfied.

5.2 Confidence Evaluation Mechanism

After explained that our mechanism would make rational players who want to get high scores do the desired actions: select the images that really contain the target landmark, we also provide a confidence measuring mechanism to eliminate mistakes produced because of the incautiousness of players. We use image clues as the ground truth images within candidate images not only for scoring but also to measure the performance of players. Before introducing the confidence evaluation mechanism, we define gaming data first.

Definition 3 Gaming Data

A gaming data is defined as $(I_{player}, w_{target}, I_{candidate}, \mathcal{I}_{positive}, \mathcal{I}_{negative})$. *Where* I_{player} *is the images selected by the player,* w_{target} *is the proper name of the target landmark,* $I_{candidate}$ *is the set of candidate images, and* $\mathcal{I}_{positive}$ *and* $\mathcal{I}_{negative}$ *are sets of ground truth positive and negative images within the game respectively.*

In each round of ImageHunter, players are given a target image and a set of candidate images. Players are asked to select from candidate images for the ones with the same landmark as the target one. The selection of a player and the original settings are recorded as a record of gaming data. For each round of game, each of two players' selections would be recorded as one gaming data. Although they play together, their gaming data is recorded separately. Afterwards, we will analyze each gaming data for confidence evaluation.

5.2.1 Confidence Measurment

First, we measure a confidence value for each gaming data. This confidence value differs according to the performance of the players in the game. The confidence value represents how much proportion the selections of a player in a game are correct, and we assume it is also the probability that the selected images actually contain the same landmark as the specimen image.

Second, since each image appears in different rounds of the game, we would measure the final confidence that represents whether an image contains the landmark by analyzing these gaming data. We propose two kinds of mechanisms for measuring the final confidence.

Definition 4 Confidence of Gaming Data

For each gaming data g_i, a correct confidence $\Phi_{correct}(g_i)$ represents how much we believe the images whithin the selections of the player (I_{player}) really contain the target landmark.

Confidence of gaming data are measured as follows:

$$\Phi_{correct}(g_i) = \frac{|I_{player} \cap \mathcal{I}_{positive}| + \frac{1}{2}|I_{player} - \mathcal{I}_{positive} - \mathcal{I}_{negative}|}{|I_{player}|}$$

Since we don't know the answer of images without ground truth, we assume half of the selected unknown images are correct. Therefore, the above equation represents the proportion of the selected images that are correct. This is the probability that the selections are correct conditioned on being selected in gaming data g_i.

After rounds of playing, an image might have been selected several times in different gaming data, and each gaming data has its own confidence value. As a result, we summarize these confidence values to calculate final confidence value $conf_{correct}(i,j)$ for measuring how much possibility that an image i contains a landmark j in it.

Definition 5 Image with Landmark Confidence

An image with landmark confidence $conf_{correct}(i,j)$ is the confidence value of how much we believe an image i contains landmark j.

We propose two mechanisms to measure $conf_{correct}(i,j)$. For each image i and each landmark j, $G_{select}(i,j)$ is the set of gaming data where the target landmark is j and image i is one of the candidate images and selected by the player.

In mechanism 1, $conf_{correct}(i,j)$ is measured as follows:

$$conf_{correct}(i,j) = \begin{cases} 0 & \text{if } G_{select}(i,j) = \phi \\ 1 - \frac{\prod_{g \in G_{select}(i,j)}(1-\Phi_{correct}(g))}{{p_{i,wrong}}^{(n-1)}} & \text{otherwise} \end{cases}$$

${p_{i,wrong}}^{(n-1)}$ is the probability that image i doesn't contains the target landmark. Since $\Phi_{correct}(g)$ could be think of as the probability that the selection is correct conditioned

on the behavior of selecting in gaming data g. $conf_{correct}(i,j)$ could be think of as the probability that the selection is correct conditioned on having been selected in a set of gaming data $G_{select}(i,j)$. Therefore, the equation of mechanism 1 is formulated from the theorems of conditional probability.

Otherwise, $conf_{correct}(i,j)$ is measured by mechanism 2 as follows:

$$conf_{correct}(i,j) = \begin{cases} 0 & \text{if } G_{select}(i,j) = \phi \\ 1 - \alpha^{\sum_{g \in G_{select}(i,j)} \Phi_{correct}(g)} & \text{otherwise} \end{cases}$$

Here $0 < \alpha < 1$. If the game needs more verification and stricter answers, α could be set bigger. On the other hand, if the game only needs a little verification and doesn't need a very strict answer, α could be smaller.

The idea of mechanism 2 is that higher confidence value would be more trustworthy. Confidence values of the set of gaming data $G_{select}(i,j)$ are accumulated to acquire the final confidence value. Confidence values which are higher and trustworthy would let the final confidence value increase faster. On the contrary, low confidence values would only increase the final confidence value a little.

From the above equation, we have measured the confidence values of how much probability an image contains a target landmark. With this we could make the final judgment that whether an image contains a landmark and whether to annotate the proper name of the landmark on an image or not. Both confidence results measured by two mechanisms are values from 0 to 1 no matter how many times an image appears and is selected. This is an advantage that we could set a fixed threshold for the final judgment.

In chapter 6, we conduct several experiments to show up that the confidence measurements are useful and beneficial. First we will show up that the measurement of gaming

data confidence is reasonable. Second, we will compare the differences between two $conf_{correct}(i,j)$ (image with landmark confidence) measuring mechanisms.

Chapter 6

Experiment

6.1 Data Collection

There are 35 landmarks being included in ImageHunter, and we collect 3221 images from Google Image search results (querying by the proper names of these landmarks.) Moreover, there are a set of 93 images which without contain no landmarks in them (for the conducting of ground truth negative images). Therefore, there are total 3314 images in our database.

We conduct experiment includes 28 players, where there are 5 groups where each consists 4 players and 4 groups where each consists 2 players. For each group, players are asked to play ImageHunter for 30 minutes continuously. Totally 593 games are produced. In each game, what images in it, which ones are the ground truths, and the selections of the players are recorded.

6.2 Evaluation

Here we proposed a game ImageHunter to produce annotation on landmark images. First, we need to verify the mechanism of the game could lead to correct actions. Player would do their best for annotating the images while they consider there is the same landmark as the target image. Moreover, even if they desire to act correctly, there still might be some errors according to uncertainties or incautiousness. For the reason, we propose confidence evaluation to analyze the gaming data from players. Therefore, afterward we will verify that the confidence evaluation mechanism proposed here is feasible.

6.2.1 Player's Action

We want to verify whether the mechanism proposed here will lead to correct action here. In ImageHunter, pattern recognition is the ability which is necessary to play the game and almost every people have the ability. If players really select images correctly, we could say that we achieve the goal that producing a game collecting information we want.

Here we analyze the actions of players by the statistics of true positive, false positive, true negative, and false negative probabilities. Define A_{select} as the actions that players select the images according to the target landmarks, and $A_{correct}$ as the actions that players should select the images according to the target landmarks. $A_{select} \cap A_{correct}$ are the set of correct selections that players have done and this is the true positive actions here. $A_{select} \cap {A_{correct}}^c$ is the wrong selections that players have done and this is the false positive actions here. ${A_{select}}^c \cap {A_{correct}}^c$ are the set of correct actions that players didn't select the images that are incorrect and this is the true negative actions here. ${A_{select}}^c \cap A_{correct}$ is the wrong

actions that players should select but didn't and this is the false negative actions. The probability of these four situation are present in Table 6.1.

	True	False
Positive	92.01%	7.99%
Negative	88.98%	11.02%

Table 6.1: The statistics of players' actions.

The correct actions true positive and true negative exceeds 88% which are high enough to confirm feasibility of our game. However, the true negative is a little lower than true positive. There are two reasons that could lead to the result. First, some images are uneasy to be selected because landmarks are very small and in obscure regions. Second, because time limits are set, players might be nervous and pass over some images incautiously. However, this is one of the interesting features in the game and the error is not so large that the result is still convincing.

6.2.2 Confidence Measurement for Gaming Data

First, we will demonstrate that the confidence measuring for gaming data is feasible. In subsection 6.2.3, we will compare confidence evaluation with mechanisms in previous works to demonstrate the advantages of confidence evaluation.

We calculate the ratio that how many actions are correct in the set of gaming data with a specific confidence value. Here we calculate the ratio of true positive actions. The results are as figure 6.2.2.

We could see the ratios of true positive actions among gaming data with high confi-

Figure 6.1: The ratio of true positive actions among the sets of different confidence values gaming data.

dence values are higher than the ratios among gaming data with low confidence values. Therefore, we could conclude that the confidence measurement reflects the correctness of the actions accurately. The confidence measurement for gaming data is therefore demonstrated rational and effective.

6.2.3 Performance among Different Mechanisms

Here we compare three mechanisms in judging the result of the game respectively. First is the mechanism used in previous works: judging the result by accumulated times. The others are two mechanisms proposed in our thesis, both judging the result by the confidence value but different integration mechanisms.

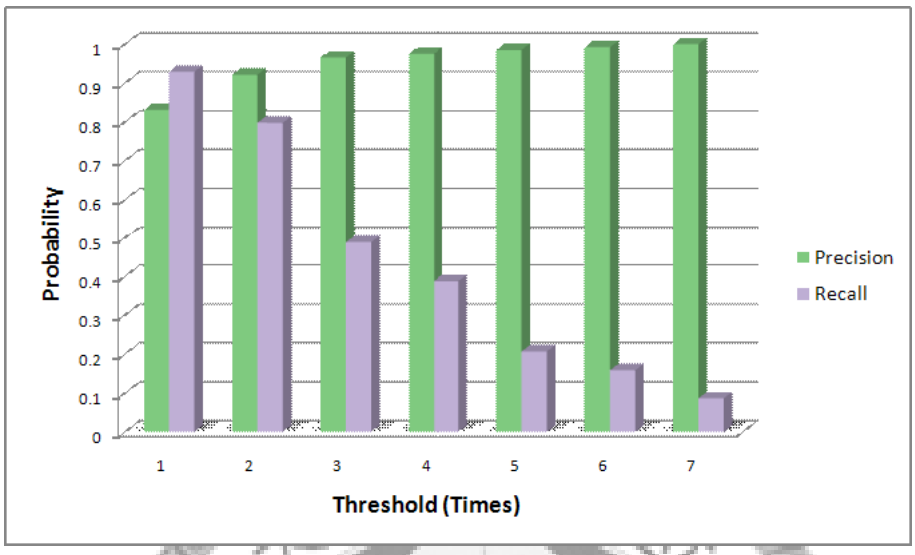


Figure 6.2: The precision and recall of the annotation results deciding by accumulated times.

Times Accumulation

Previous works judge the result of the game by times accumulation. They accumulated the times for each data and accept the data if the accumulated time exceeds the threshold. Here we conduct the experiment for observing this mechanism. We calculate the selecting times of each image corresponding to each target landmark. The annotations are produced while the accumulated times exceed the threshold t_{accept} times. The precisions and recalls of the annotations with different thresholds are computed respectively. The results are as figure 6.2.3.

However, there are some problems about this method. A game played by a few players and a game played by many players needs different threshold; therefore, how to set the threshold for accumulated times is a big problem. Moreover, the mechanism doesn't

measure how much trustworthy for each data. However, there must be a lot of difference.

In figure 6.2.3, we could see that the precision is 83% if we annotate the images which have ever been chose once. The precision of the results increase while the threshold increase. However, how to decide the threshold is still a difficult problem. It seems more times would produce more accurate results; however, the recall shrinks quickly. Since this is a productivity game, the production of the game is still a concern. In the data collection producing now we could set a temporary threshold to get a result (here might be two); however, this is because now we have the answer and the real performance value. When the game published, with more and more people engaged in, twice choices among ten players and twice choices among hundreds of players must have different performance. How to set up the threshold must be a problem.

Image with Landmark Confidence Measurement

We proposed two mechanisms to measuring image with landmark confidence values as mentioned in chapter 5.2. After getting the confidence values, we use it as the concern for the annotation decision. The annotations are accepted with at least c_{accept} confidence value. The result of the precisions and recalls of two mechanisms are as figure 6.2.3 and figure 6.2.3.

These two mechanisms first solve the problem of setting the threshold. No matter how many players playing the game, the confidence values are usually within the range [0,1]. Besides, the recall is not shrink fast as the times accumulation mechanism. Moreover, generally two mechanisms could have high precision and not bad recall.

Second, we will figure out the differences between two mechanisms. In mechanism

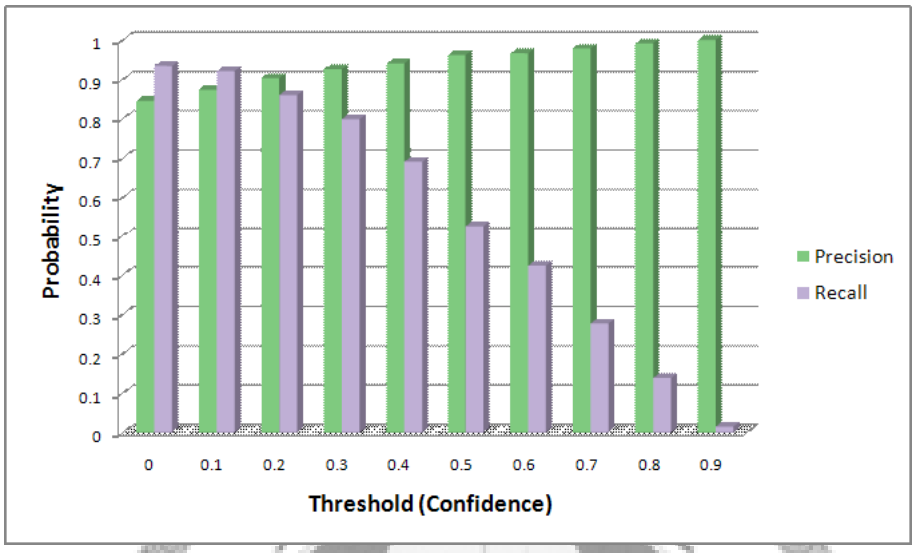


Figure 6.3: The precision and recall of the annotation results producing by confidence mechanism 1.

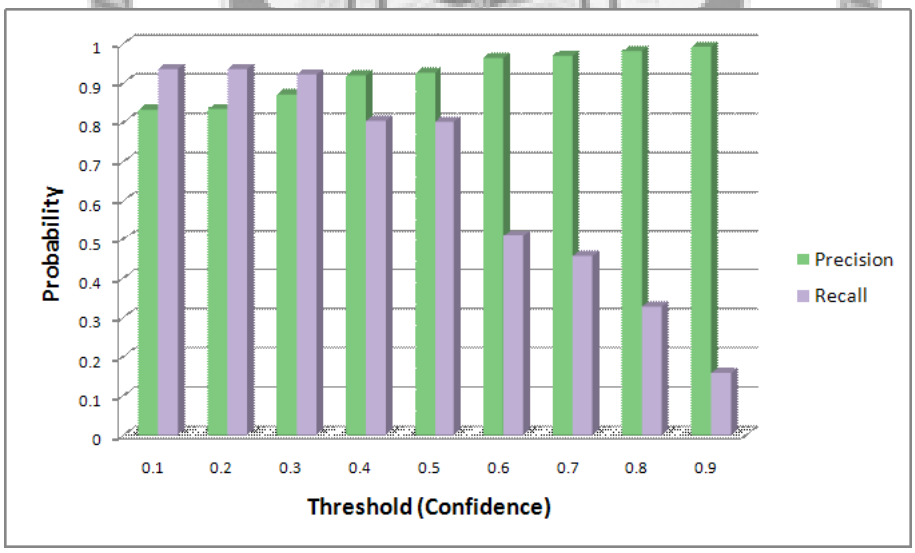
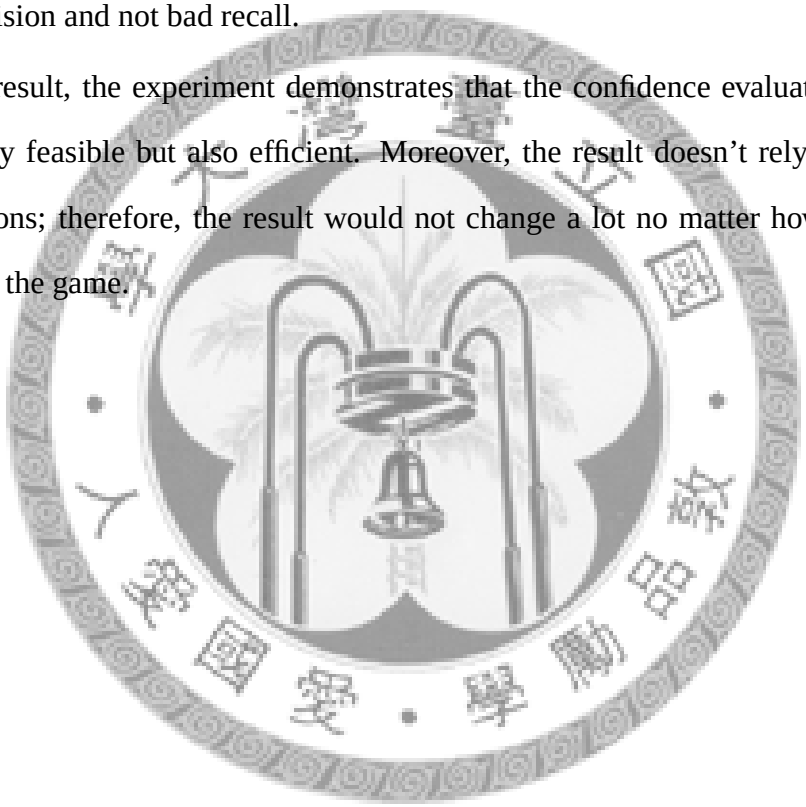


Figure 6.4: The precision and recall of the annotation results producing by confidence mechanism 2.

1, precisions are a little higher than in mechanism 2 and the recalls are low when the precision is high. This shows that mechanism 1 is a stricter mechanism on measuring the confidence. The confidence values in mechanism 1 increase slow but guarantee to be accurate. Thus we could conclude that experiments which need high precision and more careful verifications could use mechanism 1 as the evaluation mechanism and set up a higher threshold. However, mechanism 2 also performs well. Two mechanisms both have high precision and not bad recall.

As a result, the experiment demonstrates that the confidence evaluation mechanism is not only feasible but also efficient. Moreover, the result doesn't rely on the number of selections; therefore, the result would not change a lot no matter how many players engage in the game.



Chapter 7

Conclusion

7.1 Summary of Contributions

In this thesis, we address the problems about human computation games. Previous works trust the data produced by players very much; they confirm the data only if the data is produced twice from players. However, this is questionable that is a result produced by players twice would be accurate? Therefore, here we proposed another new mechanism for confirming the data. We propose a mechanism for measuring the confidence values of the data.

Here a human computation game ImageHunter proposed by ourselves is used to produce necessary gaming data in confidence measuring. The idea of ImageHunter is to collect data that could help annotating landmark images by using the proper names. By the analysis of gaming data, confidence values are produced to present the possibility that an image with a specific landmark in it. Afterwards making the final decision that whether

an annotation could be attached on an image or not.

We proposed two confidence measuring mechanisms. Experiments are conducted to evaluate the performance of the mechanisms. The result shows that using the confidence measuring mechanisms could avoid the problem of relying on number of times. Moreover, the system could produce the annotation results in high precision and recall with confidence measuring mechanisms.

7.2 Future Work

In our confidence measuring mechanism for gaming data, we assume images without ground truth have 50% probability to be correct ones. However, there could be some modifications be applied on the measuring function. For example, the probability that unknown images are correct could be measured by the confidence value they have already had. It is also likely that it could be measured by content base analysis. With these modifications provides more possibilities and might produce better results.

In another side, in our system we produce the unknown candidate images randomly. There might be some changes could lead to better production. For example, we may generate candidate images exclude what is confirmed to have the specific landmark with high probability. We could also generate candidate images from disputed ones to achieve better judgments about these images. There are a wide variety of aspects that could be researched on.

Since human computation is a new research area. Beside the issues we address, there are still a lot worthy to be investigated. Whether there are some general principles to de-

sign the mechanism of human computation games? What kinds of problems are worthy to be completed by human computation games? How to guarantee the production of human computation games could keep on while there are more and more players engaging? With the researches exploring the issues, we hope human computation games could be developed more and more mature and be used for solving great and meaningful problems.

Bibliography

[1] T. H. Chang, C. J. Ho, and J. Hsu. The photoslap game: Play to annotate. In *Intelligent System Demonstration, Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, Vancouver, British Columbia, July 2007. AAAI Press.

[2] C. Dellarocas. The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. In *MIT Sloan Working Paper No. 4296-03*, March 2003.

[3] C. J. Ho, T. H. Chang, and J. Hsu. Photoslap: A multi-player online game for semantic annotation. In *Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, Vancouver, British Columbia, July 2007. AAAI Press.

[4] T. H. Huang and J. Hsu. Beyond memories: Weaving photos into personal social networks. In G. Kaminka, D. Pynadath, and C. Geib, editors, *Modeling Others from Observations: Papers from the 2006 AAAI Workshop*, volume Technical Report WS-06-13, pages 29–36. The AAAI Press, Menlo Park, California, 2006.

[5] D. B. Lenat. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[6] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. Cyc: toward programs with common sense. *Communications of the ACM*, 33(8):30–49, 1990.

[7] H. Lieberman, D. Smith, and A. Teeters. Common consensus: a web-based game for collecting commonsense goals. In *In Workshop on Common Sense for Intelligent Interfaces ACM*

International Conference on Intelligent User Interfaces (IUI 2007), Honolulu, Hawaii, January 28 2007.

[8] T. W. Malone. What makes things fun to learn? heuristics for designing instructional computer games. In *SIGSMALL '80: Proceedings of the 3rd ACM SIGSMALL Symposium and the First SIGPC Symposium on Small Systems*, pages 162–169, New York, NY, USA, 1980. ACM.

[9] T. W. Malone. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, pages 63–68, New York, NY, USA, 1982. ACM.

[10] D. L. McGuinness, H. Zeng, P. Pinheiro da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*, May 2006.

[11] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[12] B. Thomas Adler and L. de Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th International World Wide Web Conference(WWW2007)*, Banff, Alberta, Canada, May 2007.

[13] D. Turnbull, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. In *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[14] L. von Ahn, M. Blum, N. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *Proceedings of Eurocrypt*, pages 294–311, 2003.

[15] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2004)*, pages 319–326, New York, NY, USA, 2004. ACM.

[16] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 79–82, New York, NY, USA, 2006. ACM Press.

[17] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 75–78, New York, NY, USA, 2006. ACM Press.

[18] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2006)*, pages 55–64, New York, NY, USA, 2006. ACM.

[19] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, October 2006.