KissKissBan: A Competitive Human Computation Game for Image Annotation

Chien-Ju Ho[1], Tao-Hsuan Chang[2], Jong-Chuan Lee[3],
Jane Yung-jen Hsu[4], and Kuan-Ta Chen[5]

[1,5]Institute of Information Science, Academia Sinica
[2,4]Department of Computer Science and Information Engineering, National Taiwan University
[3,4]Graduate Institute of Networking and Multimedia, National Taiwan University

{[1]kinkin,[5]ktchen}@iis.sinica.edu.tw
{[2]r97055,[3]r97944029,[4]yjhsu}@csie.ntu.edu.tw

ABSTRACT

In this paper, we propose a competitive human computation game, KissKissBan (KKB), for image annotation. KKB is different from other human computation games since it integrates both collaborative and competitive elements in the game design. In a KKB game, one player, the blocker, competes with the other two collaborative players, the couples; while the couples try to find consensual descriptions about an image, the blocker's mission is to prevent the couples from reaching consensus. Because of its design, KKB possesses two nice properties over the traditional human computation game. First, since the blocker is encouraged to stop the couples from reaching consensual descriptions, he will try to detect and prevent coalition between the couples; therefore, these efforts naturally form a player-level cheating-proof mechanism. Second, to evade the restrictions set by the blocker, the couples would endeavor to bring up a more diverse set of image annotations. Experiments hosted on Amazon Mechanical Turk and a gameplay survey involving 17 participants have shown that KKB is a fun and efficient game for collecting diverse image annotations.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work; H.5.m [**Information interfaces and presentation (e.g. HCI)**]: Miscellaneous; I.2.6 [**Learning**]: Knowledge acquisition; I.2.1 [**Applications and Expert Systems**]: Games

General Terms

Human Factors, Design

Keywords

Amazon Mechanical Turk, ESP Game, Games With A Purpose, Human Computation, Image Annotation

1. INTRODUCTION

Since the introduction of the ESP Game, Games With A Purpose (GWAP) [1] successfully demonstrates the potential of using human efforts to accomplish tasks that are intractable for computers. Useful information is generated as a side effect of game play. Many applications have been developed to solve different computationally hard problems, such as commonsense collection [2] and music annotation [3].

Current GWAP implementations rely on player collaborations to gather desired information. For example, players in the ESP Game are given the same image, and the descriptions which they both agree become the image labels. Players will be rewarded, e.g., gaining game points, for achieving consensus opinions. The matching mechanism encourages players to contribute relevant information and provides a validation for the information contributed by players. By changing the objects presented to the players (e.g., images or music) and the descriptions players are required to provide (e.g., labels or preferences), this mechanism can be applied to solve various AI-hard problems.

However, there are some known issues in this kind of collaborative mechanism. First, players can benefit from cheating by forming coalitions. If players agree on typing "a" for every image, no useful information will be collected. This issue is usually dealt by system-level approaches, e.g. pairing players from different places or using pre-recorded bots to break the coalition. Second, players tend to give easier and more generic descriptions [4], and therefore the diversity of the output becomes limited. While taboo words [5] have been proposed to alleviate this problem, they can only be constructed after an image has been played for a couple of times. Moreover, taboo words present additional information besides the image and may cause a potential problem in biasing player behavior. For instance, players may tend to guess "guy" when seeing an image with a taboo word "man".

This research proposes a novel human computation game, KissKissBan (KKB), which addresses these issues by introducing competitive element into the game. Besides the typical two-player matching mechanism, KKB introduces an

additional player, the blocker, whose objective is to stop the matching from happening. By entering labels prior to the matching process, the blocker provides a blocked word list. Other players, the couples, would get penalties for guessing the blocked words. In contrast to the taboo words, the blocked words are not visible to the couples. While the blocker is in competition with the other two players, he/she is motivated to break the coalitions between the couples. Therefore, KKB naturally provides a *player-level cheating-proof mechanism*. Besides, the invisible blocked words set the restrictions of the available words, thus the couples are encouraged to provide more *diverse* labels to evade the restrictions.

This paper starts by briefly reviewing human computation and the image naming process in a psychological view. We then introduce the design and implementation of KKB. Finally, we explain the evaluation results and outline the contribution and future work of this research.

2. RELATED WORK

In this section, we introduce the recent developments in human computation games and give a brief overview on picture naming process in psychology.

2.1 Human Computation Games

Human computation aims to solve problems that are hard for computers by utilizing human brain powers. For example, Amazon Mechanical Turk[1] provides a marketplace for the developer to outsource human intelligence tasks. Human computation games, also called Games With A Purpose (GWAP) [1], propose that using computer games can gather human players and solve open problems as a side effect of playing. The GWAP approach has been shown to be useful and is widely used in various domains, such as image tagging [5, 6], commonsense collection [2], and music annotation [3].

To ensure the quality of the collected labels, most GWAP implementations adopt consensus opinions as the correctness measure. Taking the ESP Game as an example, image labels are generated by collecting descriptions which both players agree in the game. In [7], Luis von Ahn proposed three game-templates, which summaries their successful experiences in deploying GWAPs. The three templates, namely input-agreement games, output-agreement games, and inversion-problem games, rely on player's collaborations to collect consensus opinions. In contrast to previous work, we demonstrate the game design integrating competitive elements.

2.2 Picture Naming

In the ESP Game [5], players perform the process of giving descriptions to the images. This process usually involves naming objects in the image and is well studied as picture-naming process. In psychology, an important measure in picture-naming process is how fast a person can name a given picture correctly. The naming latency is determined by two main factors, namely the word frequency and age of acquisition (AoA) [8]. Word frequency is the times of occurrence in large word corpus, and age of acquisition is the age at which player learned the word. Typically, players name faster when the words are learned earlier or have higher fre-

quency. While there are discussions about which factor is more important [9, 10], we know that the word properties, i.e. the word frequency and AoA, affect the player efforts and the responsive time in naming the pictures.

In our game design, we adopt two time intervals, 7 seconds and 30 seconds, in different game stages. Investigating the differences between the image labels collected in two stages may provide additional information besides the matching events.

3. GAME MECHANISMS

KissKissBan (KKB) is designed to be played by three online players. One of the players is the "blocker" and the other two players are the "couples". 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). Actually, KissKissBan is named by combining the objectives of game players.

The game rules are described as follows. In the begining of each round, the blocker has 7 seconds to provide blocked word list, which is the list of words he/she thinks couples might match on. These blocked words are not visible to the couples. After the 7 seconds of entering blocked words, the couples have 30 seconds to match with each other. The game time will decrease by 5 seconds if any couple types the blocked word, i.e., being blocked. Also, agreeing on the blocked word does not count as matching. The couples win the round if they successfully match with each other within the time limit, otherwise the blocker wins. Players switch roles every 5 rounds in 15 rounds of the game.

3.1 Different Roles

There are two different roles in KKB:

- Blocker: Each player will play the blocker for 5 rounds in the 15 rounds of the game. Though the blocker only has 7 seconds to act in each round, 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 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 founds the couples try to match on "a" in every round.

- Couple: The objective of the couples is the same as the players in the ESP Game: 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.

3.2 Incentive Structure

KKB is a three-player zero sum game. In the current implementation, the blocker loses 200 points and each couple gains 100 points when the couples win; the blocker gains 200 points and each couple loses 100 points when the couples lose. While the blocker and the couples are strictly competitive, the blocker is motivated to prevent the couples from cheating. Therefore, KKB provides a player-level cheating-proof mechanism.

Another difference between the blocker and the couples is their available time period for entering words. In order to

(a) The user interface for the blocker after the blocker finish typing the blocked words. The blocker can monitor all the words the couples are typing.
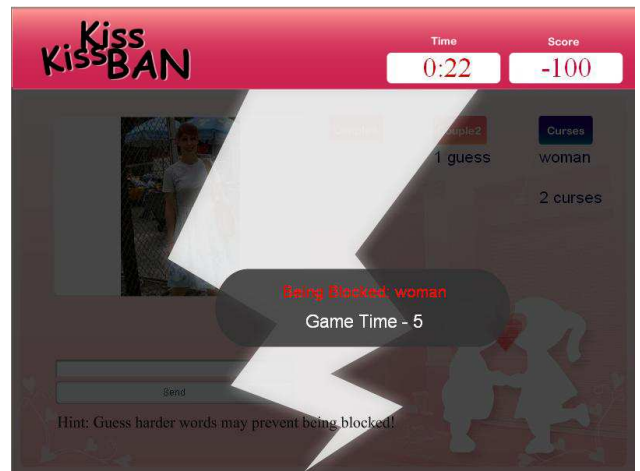
(b) The user interface for the couple when the couple is blocked. When the couple guesses the blocked word, he/she will be punished by decreasing the available game time.

Figure 1: Screenshots of the game.

increase the possibility of blocking, the blocker will have to type as many words as possible in the short period of time, i.e., 7 seconds. Studies [9, 10] indicate that the time latency in picture naming is related to the properties of words, e.g., the word frequency or age of acquisition. Therefore, the blocked words given by the blocker may be more common and easier. On the other hand, the couples have sufficient time to consider their strategies for matching. Under the belief that the blocker would give more common and easier words as blocked words, the couples are motivated to give harder words. Therefore, the mechanism of KKB promotes a diverse set of output.

3.3 Collected Labels

In the ESP Game, image labels are generated when two players match with each other. In KKB, there are two kinds of labels, namely the matching label and the blocking label. The matching label is defined as the label agreed by both couples, and the blocking label is the label that is generated when any couple guesses the blocked words. Actually, the blocking label is the matching between the couple and the blocker.

According to this definition, it is possible to generate more than one label in each round since the couples might be blocked for several times in the same round. This mechanism increases the efficiency of labeling.

3.4 Why Not Use Taboo Words

The blocked words are similar to the taboo words, but there are two main differences between the mechanisms. First, the taboo words are generated from the statistics of past play history, but the blocked words are entered by the blocker in each play. Second, the taboo words are visible by players but the blocked words are not.

Generating the blocked words which are invisible to the couples provides more uncertainties, and the couples would be motivated to provide more diverse labeling in order not to match the blocked words. Besides, the presence of the taboo words may have a potential problem of biasing player behavior. When seeing an image with a taboo word "sky", players may tend to guess some related words, such as "blue" or "cloud", instead of analyzing the image content. This phenomena can be illustrated by the auto-playing bot written by Weber et al [4]. They designed a bot which plays the ESP Game according to their trained language model. Even without analyzing the image, they can match with normal player at the rate of 81% for images with taboo words presented.

3.5 Implementation

KissKissBan follows the client-server architecture. To facilitate the deployment of the game, the client is implemented by Adobe Flash, and the server is written in pure Python. During the game, the server logs all the player activities into the database for future analysis. We also implement bots, which replay the actions performed by previous players, for games with less than three players.

4. EVALUATION

To evaluate the efficiency and quality of the collected data, we published KissKissBan on Amazon Mechanical Turk to collect game data. In addition, we conducted a user survey involving 17 players to evaluate the gameplay of KKB.

The evaluation is base on the 125 images randomly chosen from the dataset released by the ESP Game. The labels in the dataset are served as partial ground truth in the evaluation.

4.1 Efficiency

In our experiments, KissKissBan has been played for 537 times. 5521 labels, including 3296 blocking labels and 2225 matching labels, have been generated in total 4955 rounds of play. On average, there are 1.11 labels generated in each round, and 3.20 labels are collected in each minute for one game. Though we introduce a new player to decrease the chance of matching, the game is still efficient in generating image labels.

The quality of the labels is evaluated using the descriptions provided by the ESP Game Dataset. In average, there are 13.89 descriptions per image for the images we used. Taking these descriptions as "partial" ground truth for image labeling, the precision of our collected labels is 78.84% and the recall is 70.02%. The precision is the ratio of our labels to be correct. This result demonstrates the data quality of our collected labels. Actually, the correct ratio should be much higher since we only use "partial" ground truth for evaluation. The recall of 70.02% shows the diversity of KKB labels.

4.2 Property of Collected Labels

To evaluate the data properties of collected labels, we re-implemented the ESP Game for comparison. We will call this re-implemented version as K-ESP. To simplify the comparison, there is no taboo word in K-ESP. After publishing on Amazon Mechanical Turk, K-ESP has been played for 5977 rounds, and 4994 labels are collected.

In 4994 labels generated by K-ESP, 6.56 distinct labels are collected for each image, whereas KKB has collected 11.54 distinct labels per image out of 5521 labels in total. The data entropy calculated from K-ESP is 4.79, while the data entropy in KKB is 7.18. These statistics suggest that KKB motivates players to give more diverse labels. Table 1 gives an example of different labels generated by KKB and K-ESP.
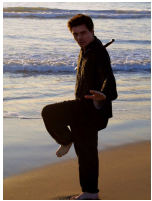
K-ESP	ML-KKB	BL-KKB
man * 21	beach * 3	sea * 9
beach * 10	water * 3	man * 8
karate * 5	sand * 3	ocean * 3
water * 1	sea * 2	black * 1
	ninja * 1	china * 1
	kungfu * 1	sand * 1
	ocean * 1	

Table 1: Labels produced by KKB and K-ESP. ML-KKB is the list of the matching label produced by matching between couples, and BL-KKB is the list of blocking label by matching blocker and one of the couples.

4.3 Gameplay Survey

To evaluate the gameplay, we conduct a questionnaire survey involving 17 participants, consisting of 11 males and 6 females. The participants are all college students, whose ages are ranging from 21 to 25. Before answering the questions, they are required to play the game at least 2 times, i.e. 30 rounds. Some of them have played the game over 10 times.

In the anonymous survey, the average enjoyability rating is 3.76 out of 5, and 88 percent of the subjects claim that they will play the game again. In addition, over 60% of the players think it's more fun and challenging to play as the blocker.

5. CONCLUSION

We have presented KissKissBan, a competitive human computation game for image annotation. While the ESP Game and other GWAPs present using player collaborations for collecting useful information, the main contribution of this paper is to demonstrate the integration of competitive elements into human computation games. In the

design of KKB, we have two advantages over traditional human computation games: 1) KKB provides a player-level cheating-proof mechanism which can alleviate coalition between players; 2) KKB motivates players to contribute more diverse labeling and therefore collects a broader set of data. Evaluations on Amazon Mechanical Turk and a gameplay survey have shown KKB to be an efficient and fun game for collecting diverse image annotations.

In our future work, we will explore how players behave when they play as different game roles. The differences between KKB labels, namely the matching labels and the blocking labels, is also an important subject to investigate.

6. REFERENCES

[1] Luis von Ahn. Games with a purpose. *IEEE Computer Magazine*, 39(6):92–94, 2006.

[2] Luis von Ahn, Mihir Kedia, and Manuel 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. ACM Press, 2006.

[3] Edith L. M. Law, Luis von Ahn, Roger B. Dannenberg, and Mike Crawford. Tagatune: A game for music and sound annotation. In *International Conference on Music Information Retrieval (ISMIR' 07)*, pages 361–364, 2003.

[4] Ingmar Weber, Stephen Robertson, and Milan Vojnovic. Rethinking the ESP game. Technical report, Microsoft Research, September 2008.

[5] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM Press, 2004.

[6] Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 55–64. ACM Press, 2006.

[7] Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[8] J.B. Carroll and M.N. White. Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1):85–95, 1973.

[9] C.M. Morrison, A.W. Ellis, and P.T. Quinlan. Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory and Cognition*, 20:705–705, 1992.

[10] C. Barry, K.W. Hirsh, R.A. Johnston, and C.L. Williams. Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language*, 44(3):350–375, 2001.