A Novel Algorithm for Improving the ESP Game

Mohamed Sakr
Department of Computer Science
Faculty of Computers and
Information, Menoufia University
Shebin El-Kom, Egypt

Hany Mahgoub
Department of Computer Science
Faculty of Computers and
Information, Menoufia University
Shebin El-Kom, Egypt

Arabi Keshk
Department of Computer Science
Faculty of Computers and
Information, Menoufia University
Shebin El-Kom, Egypt

***Abstract*—one of the human-computation techniques is games with a purpose (GWAP) and microtask crowdsourcing. These techniques can help in making the image retrieval (IR) be more accurate and helpful. It provides the IR system's database with a rich of information by adding more descriptions and annotations to images. One of the systems of human-computation is ESP Game. ESP Game is a type of games with a purpose. In the ESP game there has been a lot of work was proposed to solve many of the problems in it and make the most benefit of the game. One of these problems is that the ESP game neglects player's answers for the same image that don't match. This paper presents a new algorithm to use neglected data to generate new labels for the images. We deploy our algorithm at the University of Menoufia for evaluation. In this trial, we first focused on measuring the total number of labels generated by our Recycle Unused Answers For Images algorithm (RUAI). In our evaluation of the RUAI algorithm we present a new evaluation measure we called it quality of labels measure. This measure identifies the quality of the labels in compared to the pre-qualified labels. The results reveal that the proposed algorithm improved the results in compared to the ESP game in all cases.**

Keywords—ESP game; Games with a purpose; Human computation; crowdsourcing

I. INTRODUCTION

There are problems which are difficult to be processed by computers such as those related to artificial intelligence. These problems are easy to be solved by the human brain power. Human computation is the idea of solving difficult problems using human intelligence. Some of these problems are related to artificial intelligence (AI) or image recognition.

Games with a purpose (GWAP) are one of the human computation [1,2]. GWAP are a way to make useful of the human desire to be entertained. Several GWAP systems have been proposed for image annotation and commonsense reasoning. Von Ahn and Dabbish [3] classified GWAP into three game-structure templates that generalize successful instances of human computation games: output-agreement games, inversion-problem games, and input-agreement games. Yuen et al. [4] added output-optimization game to these three templates.

ESP game is one of the GWAP systems. ESP game was the first systems to clarify the advantages of using human computation and GWAP systems. It is example of output-agreement games and is a two player's game for labeling images [5]. Barnard et al. [6] reported that labeling images has proven to be a hard problem for computer vision, but it is something that humans can do easily. It has been shown that the image labels collected through the ESP game are usually of good quality. Moreover, the game results allow more accurate image retrieval, help users block inappropriate images (e.g., pornographic content), and improve web accessibility (e.g., the labels can help visually impaired people surf web pages [7]). In order to humans to label images there must be some sort of motivation. One type of motivation is entertainment, which is achieved in the ESP game. In the ESP game the players are chosen randomly and are assigned the same image. Each player doesn't know the other player and the two players can't communicate with each other. The only thing they have in common is the image that they play with. Each player is asked to give description to that image and has to guess what the other player is typing for each image to win the game and go to the next image. Once the two players have entered the same word, this word becomes the label for the image. The easiest way for both players to type the same string is by typing something related to the common image. The round lasts for 2.5 minutes.

During the round the players try to describe as many images as they can. The players get number of points for each image they label. If the players agree on 15 images they get a large number of bonus in points. Once there is a difficult *image* that the players can't agree on they both can press the Pass button. The game is attached with a scoreboard, with the names of players with the highest scores. Empirical studies of other peer-production systems have shown that points are a key feature in motivating users [8].

One type of GWAP systems and output-agreement games is ESP game. During the play in the ESP game it appears anecdotally that people coordinate on the same words, but the other words are neglected. In this work we are concentrating on ESP game and on solving one problem of the ESP Game that the player's answers for the same image that don't match in the same game are neglected. This paper presents a new algorithm to use these neglected data to generate new labels for the images.

The rest of this paper is organized as follows. Section II presents a review of related works on problems of the ESP game. Section III presents our algorithm to solve the problems of neglected data in ESP game. The results and simulation analysis of our proposed algorithm are presented in section IV. Section V provides conclusions and future work.

II. RELATED WORK

ESP game is one of the successful applications of the games that harvest human intelligence and time to solve tasks,

which is difficult by computer. In this work we will show that, although the idea underlying the game is an extremely powerful one, more care needs to be taken in the design as the game uses only the answers of the players that match and neglects the other answers.

After analyzing the ESP game we notice the following problems:

- Informative labels: many of the labels from the ESP game are redundant and not very informative ("man" and "guy"). Many labels are generic and not descriptive ("building" and "terraced house"). Many labels can be expected and generated automatically ("water", "blue", "sky" and "clouds" are all related)
- How to measure the system's productivity: Test if the system is productive and give informative labels with good quality and acceptable quantity.
- How to select the next puzzle to play with: How to select the next puzzle (next image to play with) select one with most descriptions or with least descriptions

In the ESP game there has been a lot of work was proposed to solve many of the previous problems in it and make the most benefit of the game. In the next paragraph we try to present some of the previous work that is proposed to solve the problems of the ESP game.

Weber et al. [9] notice that the ESP game failed to collect informative labels so they proposed a language model to generate probabilities to the next labels to be added given the pre-added labels as training data. Chen et al. [10] proposed anew metric called system gain, use analysis to study the properties GWAP systems and implemented a new puzzle selection strategy to improve the GWAP systems. Jain and Parkes [11] presented game theoretic analysis for the ESP game, and they investigated the equilibrium behavior under different incentive mechanisms and provided guidelines to design incentive mechanisms. Von Ahn and Dabbish [3] suggested a set of evaluation metrics, such as throughput, lifetime play, and expected contribution, to determine whether ESP-like GWAP systems are successful. Ho et al. [12] also notice that the set of labels determined from the ESP game for an image, are not very diverse, and develop a three-player version of the ESP game that involves the addition of a "blocker" to type in words that the other two players cannot use to match. In this work we address the informative labels problem and how to generate new labels with no need to extra un-useful game rounds between players.

III. Recycle Unused Answers For Images Algorithm (RUAI)

After analyzing the ESP game, the previous problems and there solutions, we found that in some time when the players play the game they enter informative labels and these labels when they are not agreed upon they are neglected and trashed. So the ESP game throws away the unused answers. In this section, we present the (RUAI) algorithm which recycles the player's answers to make use of these informative answers in the situations where they are neglected as shown in Fig.1 and Fig.2.

RUAI Algorithm

Input: images (E), labels (L), answers (A)
Output: labels with new words (L)
For i = 1 to count (E) **Do**
  q = count (distinct A for E_i)
  For n = 1 to q Do
    C= $A_{i,n}$;
      If(count(C) >= *threshold*) **Do**
        isExist = false;
        For a=0 to count(L_i) **Do**
          If(C == $L_{i,a}$)
            isExist = true;
            Break;
          ENDIF,
          a++**;**
        ENDFOR
        If(*isExist* == false)
          Do add C to L_i
        END IF
      END IF
    n++;
  ENDFOR
  i++;
ENDFOR
Stop

Fig. 1. The RUAI Algorithm.

There are two types of algorithms offline and online. Online algorithm runs during the game while the players are playing. Offline algorithm runs after the players finish playing the game. The RUAI algorithm is categorized as offline algorithm as it runs on the data from the players after they play the game not during the game (the algorithm starts to run after the players finish playing). Also, the algorithm runs on all the players and their data not only on the two players of the game. The scenario of algorithm work is done as follow:

Step 1: get all the images, its corresponding answers and its labels from the database.

Step 2: for each image get all the distinct answers.

Step 3: for each answer calculate the count of its occurrence and test it against a given number which represent the number of players that agree on that answer. This number ranges from 2 to *m* (threshold). Threshold is decided by the user when searching for a given query in the database. When the threshold is increased it means that the resulted images will be more relevant to the search query (give me the images related to the query Q with accuracy X where Q is the query that the user entered and X is the threshold).

Step 4: if the count is bigger or equal than the threshold we will check if the answer is in the labels for that images if no we will insert the answer as a new label for that image. If yes go to the next answer.

Step 5: if the count is smaller than the threshold we will go to the next answers.

Step 6: after iterating between all the answers for this image we will go to the next image and redo the steps from 2 to 5.

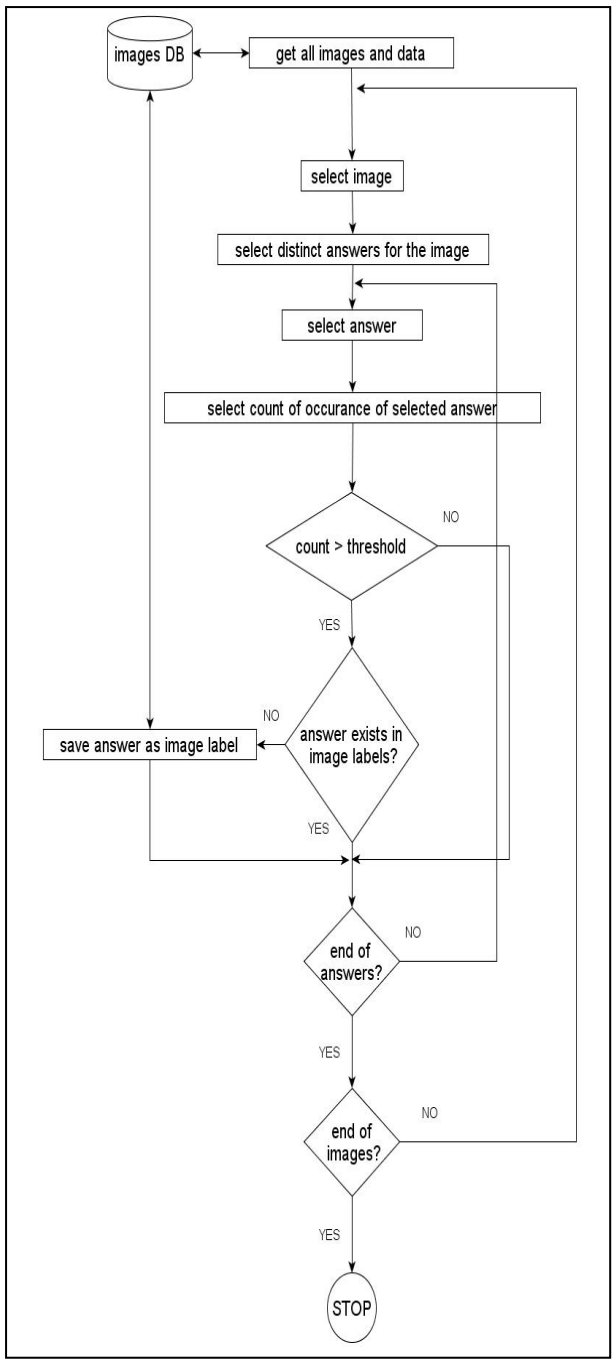

Fig. 2. Flowchart of the RUAI algorithm.

First we will get all the images E, all it's corresponding answers and all its labels from the database. Second, for each image E_i we will get all distinct answers A. For each answer A_n we will calculate the count of its occurrence. After that, we will test the count if it is bigger or equal the threshold or not. If yes we will look through the labels L for the image E_i to check if that answer is exist or not. If it doesn't exist it will be inserted to the labels L of image E_i then go to the next answer, repeat the previous steps and so on until the end of all image's E_i answers. BUT if it exists it will go to the next answer for image E_i, repeat the previous steps and so on until the end of all image's E_i answers After the end of all image's E_i answers it will go to the next image, repeat the steps and so on until the end of images E.

A. *Case study of the RUAI algorithm*

This case study illustrates the work of the (RUAI) algorithm. Suppose there are four players P1, P2, P3 and P4 that are playing the game as shown in Fig.3. P1 plays with P2 and P3 play with P4. The four players are playing on the same image I1. P1 entered the words B and D. P2 entered the words A, B and C. As the ESP game the label for the image will be B and the answers A, C and D will be neglected. The other game is between P3 and P4. P3 enters the words Z and D. P4 enters the words A and Z. As the ESP game the label for the image will be Z and the answers A and D will be neglected. From the ESP Game the labels for the image I1 will be B and Z as shown in Fig.4 (a). Now we will perform our algorithm which will iterate on every image as shown in Fig.4.

In this case study the RUAI algorithm would be performed on I1. For each distinct answers on that image which will be (A, B, C, D and Z) test count of occurrence of each distinct answer against the threshold for now it will be 2. Count of A, B, C, D and Z will be 2, 2, 1, 2 and 2 as shown in Fig.4 (b). The labels A, B, D and Z all their counts are bigger than or equal threshold 2 so all will be labels for the image. But B and Z are already inserted as labels by the ESP game so the new labels A and D will be inserting into the database as shown in Fig.4(c) (d).

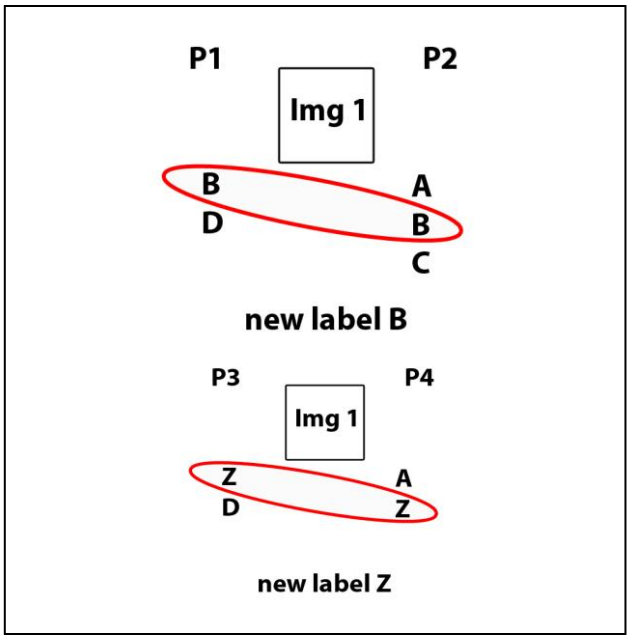

Fig. 3. Case study Game scenario

IV. Results And Simulation Analysis

At the beginning we start to get a data-set of images and labels to work on. A number of crowdsourcing data-sets are available for research. For example, von Ahn et al. contributed a list of 100,000 images with English labels from their ESP Game [13]. We used von Ahn et al. data-set. First, we integrate only the images to our system and not the labels.

Image	labels
I1	*B,Z*

(a) The image and its labels from the ESP game

Distinct answer	count
A	2
B	2
C	1
D	2
Z	2

(b) Distinct answers for the image I1 and the count of each answer.

Distinct answer	count
A	2
D	2

(c) Distinct answers for the image I1 with count >=2(threshold) and not in fig.4(a)

Image	labels
I1	B,Z,A,D

(d) image and new labels after run RUAI algorithm

Fig. 4. RUAI data table

We deploy our system at the University of Menoufia for evaluation using Java 1.6 as programming language, Mysql [14] server as database management server and run the server over Intel core i7 with 4 GB Ram PC on windows 7-64 operating system.

In this trial, we first focused on measuring the total number of labels generated by our RUAI algorithm. We sent emails to the staff of the University. We advertised the system as a free game and provided a link to the system. Sixteen users signed up initially (many of them were research students and demonstrators). The results of the proposed algorithm and the ESP game are presented in Table 1.

The results show that the total number of images that were described is 56 images. Total number of answers users entered was 736 answers. Total number of labels that our prototype of the ESP Game generated was 155 labels. Total numbers of labels that our proposed algorithm generated were 198 labels. So our algorithm generated new 43 labels of the images.

TABLE I. THE RESULTS OF THE RUAI ALGORITHM AND THE ESP GAME FOR SIXTEEN USERS.

Method	Total Images	Total Answers	Total Labels
ESP Game	56	736	155
RUAI algorithm	56	736	198

In our evaluation of the RUAI algorithm we present a new evaluation measure we called it *quality of labels measure*. This measure identifies the quality of the labels compared to pre-qualified labels.

To compute the quality of labels measure we first compare the labels results from our RUAI algorithm with the labels in von Ahn et al. data-set and compute the total number of labels resulted from the RUAI algorithm which is exists in von Ahn et al. data-set then calculate the percentage of them. This computation is done for each image then at the end we compute the average. The mathematical formula of the quality of labels measure is illustrated as in (1)

$$Quality\ of\ labels\ measure = \frac{\sum_{i=1}^{N} \frac{length(L_{iRUAI} \cap L_{idata-set})}{length(L_{idata-set})}}{N} \quad (1)$$

$L_{idata-set}$: Set of labels for image *i* from von Ahn et al. data-set

L_{iRUAI} : Set of labels for image I from RUAI algorithm

N : Total number of images

Length: is a function to calculate the length of a given set

The algorithm result's shows that about 78% of labels in von Ahn et al. data-set were generated by our RAUI algorithm. We noticed that in the previous work of evaluating the quality of the labels is done manually by asking group of people to describe images and compare the results or by giving them questions for a given image to know if the labels for that image are correct. In this paper the evaluation of the label's quality is done automatically by using the quality of labels measure formula. The advantages of using this formula is to reduce the time and cost.

Due to the time limitation, we did not observe the users for longer period to give more answers. However, we believe if we deploy the system to a larger demographic, our algorithm would produces even more promising results fuelled by the network effect.

V. CONCLUSION

This paper presents a new algorithm to generate new labels for the images with no need to extra game rounds. The algorithm overcame some of the problems in the ESP game by using neglected player's answers for the same image that don't match. Also we present a new evaluation measure which called quality of labels measure. This measure identifies the quality of the labels compared to pre-qualified labels. The using of this measure improved the time and saved the cost. The results of comparing RUAI algorithm and the ESP game reveal that the RUAI algorithm is much better than the ESP game in all cases.

In the future work we intend to generate new labels for the images using the data mining techniques. Furthermore we intend to evaluate the performance of our algorithm in terms of its efficiency and scalability.

REFERENCES

[1] C. Harris and P. Srinivasan , "Comparing Crowd-Based, Game-Based, and Machine-Based Approaches in Initial Query and Query Refinement Tasks," Lecture Notes in Computer Science, vol. 7814, pp. 495-506, 2013.

[2] S. Thaler, E. Simperl and S. Wolger , "An Experiment in Comparing Human-Computation Techniques," Internet Computing, IEEE, vol. 16(5), pp. 52-58, October 2012.

[3] L. Ahn and L. Dabbish, "Designing games with a purpose," Communications of the ACM, vol. 51(8), pp. 58-67, August 2008.

[4] M. Yuen, L. Chen and I. King, "A Survey of Human Computation Systems," Proceeding CSE '09 Proceedings of the 2009 International Conference on Computational Science and Engineering, vol. 4, pp. 723-728, August 2009.

[5] L. Ahn and L. Dabbish, "Labeling images with a computer game," Proceeding CHI '04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319-326, 2004.

[6] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei and M. Jordan, "Matching Words and Pictures," Journal Of Machine Learning Research, vol.3, pp. 1107–1135, 2003.

[7] J. Bigham, R. Kaminsky, R. Ladner, O. Danielsson and G. Hempton, "WebInSight : Making Web Images Accessible," Proceeding Assets '06 Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility, pp. 181 - 188, 2006.

[8] K. Nam, M. Ackerman and L. Adamic, "Questions in , Knowledge iN ? A Study of Naver ' s Question Answering Community," Conference on Human Factors in Computing Systems, CHI 09, ACM, April 2009.

[9] S. Robertson, M. Vojnovic and I. Weber, "Rethinking the ESP Game," Proceeding CHI EA '09 CHI '09 Extended Abstracts on Human Factors in Computing Systems, pp. 3937-3942, September 2009.

[10] L. Chen, B. Wang and K. Chen, "The Design of Puzzle Selection Strategies for GWAP Systems," Journal Concurrency and Computation: Practice & Experience, vol. 22(7), pp. 890-908, May 2010.

[11] S. Jain and D. Parkes, "A Game-Theoretic Analysis of Games with a Purpose," Proceeding WINE '08 Proceedings of the 4th International Workshop on Internet and Network Economics, pp. 342-350, 2008.

[12] C. Ho, T. Chang, J. Lee, J. Hsu and K. Chen, "KissKissBan: a competitive human computation game for image annotation," Proceeding HCOMP '09 Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 11-14, 2009.

[13] ESP Game dataset: http://server251.theory.cs.cmu.edu/ESPGame100k.tar.gz

[14] MySQL Server http://dev.mysql.com/downloads/mysql/