PhotoSlap: A Multi-player Online Game for Semantic Annotation

Chien-Ju Ho and **Tsung-Hsiang Chang** and **Jane Yung-jen Hsu**
Department of Computer Science and Information Engineering
National Taiwan University
{b90090,vgod,yjhsu}@csie.ntu.edu.tw

Abstract

Multimedia content presents special challenges for the search engines, and could benefit from semantic annotation of images. Unfortunately, manual labeling is too tedious and time-consuming for humans, whereas automatic image annotation is too difficult for the computers. In this paper, we explore the power of *human computation* by designing a multi-player online game, PhotoSlap, to achieve the task of annotating metadata for a collection of digital photos. PhotoSlap engages users in an interactive game that capitalizes on human ability in deciphering quickly whether the same person shows up in two consecutive images presented by the computer. The game mechanism supports the *objection* and *trap* actions to encourage truthful input from the players. This research extends human computation research in two aspects: game-theoretic design principles and quantitative evaluation metrics. In particular, PhotoSlap can be shown to reach *subgame perfect equilibrium* with the target strategy when players are rational and without collusion. Experiments involving four focus groups have been conducted, and the preliminary results demonstrated the game to be fun and effective in annotating people metadata for photo collections.

Introduction

As digital cameras and other image capturing devices have become ubiquitous, the way people manage their photos has changed dramatically from the era of film cameras. We are experiencing an exponential growth in the volumes of digital content available over the web for sharing. Despite impressive advances in Internet search technologies, multimedia content still presents significant challenges for the state-of-the-art search engines.

Semantic annotation of images can greatly improve the accuracy and efficiency of image search. Annotated metadata of a personal photo collection can help profile a person and facilitate photo sharing (Huang & Hsu 2006). In general, we can define *photo metadata* to include the following key attributes:

- People: *Who* are in the picture?
- Objects: *What objects* are in the picture?
- Event: *What* was the event?
- Time: *When* was the photo taken?
- Location: *Where* was the picture taken?

While such information may be gleaned from a picture by humans with ease, automatic annotation is quite difficult for the computers. The standard JPEG image format embeds useful metadata in the EXchangeable Image File format (EXIF). Attributes such as time, resolution, ISO, aperture and shutter speed are recorded at the time of image capture, and can be easily extracted from EXIF. With the growing availability of GPS-equipped devices, it becomes possible to capture location information with a camera. However, there's no information about *who* and *what*. Despite some limited success of vision algorithms in specialized domains, no general solution can match the performance of humans in image recognition and understanding.

Unfortunately, manually annotating large collections of photos can be quite tedious and time-consuming for most people. Proper incentives are in order. Monetary rewards are effective but too costly except for images of high commercial value. As evidenced by Wikipedia, Flickr, and steve (the art museum social tagging project[1]), promoting a common good may provide incentives for people to contribute their time, energy and knowledge. Last but not least, incentives can be offered in the form of *fun*. The concept of *human computation* proposed in (von Ahn & Dabbish 2004; von Ahn, Liu, & Blum 2006) successfully demostrated turning games into productivity tools. Players are brought together to interactive games for fun, and their actions can be used to generate image annotations as a result.

This research explores photo annotation with multi-player online games by extending research on human computation in two aspects: disciplined game design and meaningful evaluation metrics. PhotoSlap is a web-based variation of Snap, a popular card game. Players are engaged in an interactive game to decipher whether the same person shows up in two consecutive images presented by the computer. The mechanism for *objection* and *trap* encourages truthful input from the players. Using game theoretic analysis, we show that PhotoSlap reaches *subgame perfect equilibrium* with the target strategy for rational players. That is, players

[1] http://www.steve.museum

would take the actions prescribed by the strategy in order to maximize their scores in the game. Experiments involving four focus groups have been conducted, and the preliminary results showed the game to be fun and effective in annotating people metadata for photo collections.

This paper starts by surveying recent work on both automatic image annotation and human computation for photo-annotation. The design of PhotoSlap is introduced in terms of its game mechanism, game strategy analysis, and gameplay. The system overview and implementation details are then presented. Finally, the paper summarizes the results from our experiments with four focus groups, and outlines the contributions and future extensions of this research.

Related Work

We have surveyed relevant research on automatic image annotation as well as recent developments in human computation, and in particular for photo annotation.

Automatic Image Annotation

In recent years, research on automatic image annotation has been quite active. Most of the approaches require training the annotation system with a large collection of (manually) annotated images. Training images are analyzed and transformed into feature vectors. The relationships between the feature vectors and the annotations are captured by various probabilistic models, such as the co-occurrence model (Mori, Takahashi, & Oka 1999), translation model (Duygulu *et al.* 2002), or cross-media relevance model (Jeon, Lavrenko, & Manmatha 2003). ALIPR [2](Li & Wang 2006) is an example of applying automatic annotation to a real-world domain. It can annotate any online images in real-time and the highest ranked word in the annotation has an accuracy of 51%.

While automatic annotation has impressive functionalities, its performance depends heavily on the collection of training data. Annotations are often selected from a restricted vocabulary of limited size. Given our research focus on annotating people metadata, we have found both collecting representative data (face photos) for training and annotating photos with all possible human identities present significant bottlenecks in practice.

A more related domain to our implementation is face recognition. However, as pointed out by the survey in (Zhao *et al.* 2003), current face recognition approaches still encounter the problems of changes in illumination and pose. The performance of current face recognition systems are still far away from the capability of human perception.

Human Computation

Research has shown that humans do remarkably well in identifying faces, even under various kinds of degradations (Sinha *et al.* 2006). To utilize human brain power, several tools have been developed to annotate images via some form of human-computer collaboration over the web.

LabelMe (Russell *et al.* 2005) is a web-based tool for annotating images and sharing those annotations in the community. It provides an easy-to-use interface for manual labeling the object information, including position, shape, and object label. Another example is Riya[3], a visual search engine with a personal album. It provides a semi-automatic procedure that combines manual labeling and machine learning to achieve effective photo annotation.

The ESP Game(von Ahn & Dabbish 2004) presents an impressive work on how to take advantage of human desire to be entertained. It is an interactive game in which a player attempts to label a given image presented by the system to match any label given by his online partner within a predefined time limit. This simple yet innovative game turns the tedious manual-labeling process into entertainment. Users are motivated to contribute their image labeling skills while enjoying themselves.

The research reported in this paper is inspired by von Ahn's research on human computation, which utilizes human computing power by designing interactive games to solve problems not yet solvable by computers. While ESP and its variations have generated impressive responses, we would like to further explore the design principles underlying such productivity games and identify meaningful evaluation metrices. Players engage in rational decision making rather than second-guessing each other. Instead of labeling one photo at a time, clusters of face images are annotated together. There's no discipline to ensure the game rules will produce the desired outcome in human computation. Therefore, we proposed game theoretic analysis to ensure truthful player behavior, thereby producing quality annotations. While we explain the idea with a specific gameplay, PhotoSlap, it is essential in designing any serious game.

Game Mechanism

PhotoSlap is designed as a multi-player online game, with the rules similar to the popular card game *Snap*[4]. In this game, cards of photos are dealt to each player in face-down stacks. Players take turns to take the top card from their stacks and place it face-up in a central pile. If two cards placed consecutively on the pile are *matching* in that they contain photos of the same person (alternatively, object, event, or location), then the first player to slap on the central pile wins the round. To ensure data quality, random slapping is discouraged by the machanism of objection and setting traps. The game actions and the scoring mechanism are explained in more details below.

Game Actions

Each player in PhotoSlap may perform four possible actions:

- **Flip**

  Each player flips a single card in turn. The photos are chosen by the game server adaptively.

- **Slap**

[2]http://alipr.com

[3]http://www.riya.com/

[4]The name PhotoSnap has been taken by another software.

Given the last two consecutive cards on a central pile, players may choose whether to slap. To achieve high scores, a rational player should slap as soon as he/she recognizes two consecutive photos of the same person.

- **Object**
  When a player slaps, the other players have the option to challenge the slapped result by flagging an "objection." If the objection is successful, the objector would gain points while the slapper would lose points. If the objection fails, i.e., falls into the trap, the objector is penalized with a large point loss.
- **Trap**
  While the "object" action is used to prevent random slapping, the "trap" mechanism is designed to prevent random-objection. At the beginning of a new game, each player is presented with a subset of all photos, in which he/she can set one or more traps by identifying photos containing faces/heads of the same person.

To address a potential problem of random-trap, in our game design, the "trap" and "slap" mechanisms serve as mutual validation given that both actions can be applied to the same target pairs, i.e. photos with the same person. The trap photos identified in the trap stage will be randomly selected and presented in the game stage. The player who sets the trap is not allowed to slap on it. On the other hand, he/she can get points if another player slaps on it, but will lose points if no player slaps. The game mechanism is designed to encourage players to slap and trap as accurately as possible.

Scoring Mechanism

- A player gets points under the following conditions:
  - First to slap without drawing any objection.
  - First to object without falling into a trap.
  - Another player slaps on the trap he/she sets.
- A player loses points under the following conditions:
  - First to object and fall into the trap.
  - No player slaps on the trap he/she sets.

Intuition Behind the Game Rules

PhotoSlap can be viewed as a system that continuously questions all players whether the same person is shown in two consecutive images. "Slap" is the user action to label two photos of the same person, while "objection" is the user action to eliminate such labelings. When a pair of photos presented by PhotoSlap are slapped without meeting any objection, the group of users are said to have reached an agreement on the same person relation. To encourage truthful input from users, the "trap" mechanism is designed not only to prevent random-objection but also to provide a mutual validation with "slap".

Game Strategy Analysis

In this section, we plan to use game theoretic analysis to show that rational players will take the action prescribed by the target strategy in PhotoSlap. For example, players tend to set a trap and to slap when they believe the two photos do match, i.e. images of the same person. On the other hand, players tend to *object* when they believe the photos do not match. The desired choices of action per the player's belief on whether the photos match are summarized in Table 1.

Belief	*Match*	*No Match*
Trap	Set	Stay
Slap	Slap	Stay
Object	Stay	Object

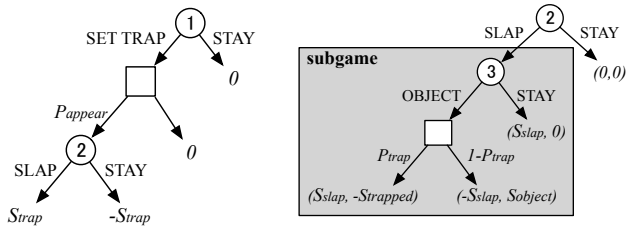
Table 1: The target strategy.

For simplicity, let us assume that all players are rational, striving to maximize their scores for each game. Each player also believes that the other players are rational. All players have the ability to identify if the two photos are of the same person. Besides, the players do not communicate with one another about their playing strategies. We model the whole process of the game as a multi-player extensive game. With the exception of the player who sets the trap, the players do not know whether any given pair of cards is a trap or not. As a result, the game should be modeled as an extensive game *without perfect information*.

There are two stages in the process of each PhotoSlap game. The first is the trap stage in which players have the opportunity to set trap over a given subset of photos. The second is the game stage in which the snap-like game is played. The player strategy for each stage is analyzed as follows.

Trap Stage

During the trap stage, each player may choose to set any pair of photos as a trap. The pair may be selected by PhotoSlap during the game stage according to a given probability P_{appear}. The game tree is illustrated in Figure 1(a). For simplicity, only the payoff of player 1 is shown.



(a) The trap stage. (b) The game stage.

Figure 1: Game trees.

It is clear from the game tree that player 1 should set the trap if he/she believes player 2 will slap, and stay otherwise in order to get the highest score.

Game Stage

The players take turns to flip a card from their stacks. Whenever a new photo is displayed, the players can choose

whether to slap on the last two flipped cards. Given that the player who sets the trap cannot slap on it, and a rational player will not object to his/her own trap, the game stage can be modeled as a three-player game. Player 2 is defined to be the player who slapped first, and is said to stay (do nothing) if no player slapped. Whenever a pair is slapped, the other players can choose whether to object. Player 3 is defined to be the first player who objected, and is said to stay if no player objected. The game tree is drawn in Figure 1(b). The payoff of player 2 and player 3 is shown in the leaf node.

Consider the subgame in which player 3 is the first one to act. Let P_{trap} be the probability that the slapped pair is a trap. The expected payoff for player 3 to object is $P_{\text{trap}}(-S_{trapped}) + (1 - P_{\text{trap}})S_{object}$, where $-S_{trapped}$ is the penalty for falling into the trap and S_{object} is the score for successful objecting. The expected payoff to stay is 0.

Player 3 should object if P_{trap} is smaller than $S_{object}/(S_{object} + S_{trapped})$, and stay otherwise. By carefully setting the value of $S_{trapped}$ and S_{object}, player 3 would choose to object if he/she believes player 1 did not set the trap, and stay otherwise. Now consider the whole game tree as in Figure 1(b). Player 2 will choose to slap if he/she believes player 3 will stay, and choose to stay otherwise.

Subgame Perfect Equilibrium

According to the analysis above, the action of player 1 depends on his/her belief about the action of player 2. The action of player 2 depends on his/her belief about player 3. And the action of player 3 depends on his/her belief about player 1. It's clear that the target strategy, as shown in Table 1, is a subgame perfect equilibrium of this game. Thus, we can inform the players of a default strategy,the target strategy, to satisfy subgame perfect equilibrium. Since the players do not communicate with others about their strategy, they will tend to keep following the strategy being told.

Gameplay

In the electronic game industry, game developers have come up with several design principles (Kramer 2000) that promote deep and persistent engagement. *Fun* is the most important element that motivate players, and gameplay is all the activities and strategies game designers employ to get and keep players engaged (Prensky 2002).

The core elements of gameplay in PhotoSlap are *pattern-recognition* challenge and *reaction time* challenge. Players compete with each other to slap on a matching pair as soon as possible. To keep players engaged, PhotoSlap is designed to be self-adaptive in that the competitors are carefully chosen so that the difficulty of challenges, the competitors' reaction time, and their ability can be balanced to enable them to stay in the flow state (Csikszentmihalyi 1990). Furthermore, to ensure the players experience challenges in a proper tempo, the appearance of a matching pair will be dynamically adjusted to follow a tension curve that has proper frequent peaks. Each time a card is flipped, PhotoSlap makes a decision about whether to present a challenge by a probability function $p(t) = Ne^{gt}$, where N is the normalizing constant, t is the interval from the last hit to the current flipping, and g is a constant for adjusting the growing speed of tension.

System Overview

Figure 2 shows the overview of the PhotoSlap system. The system can be splited into three layers: detection layer, game layer, and annotation layer. In the detection layer, the face/head images are extracted automatically using the face detection module or by manual annotation for any new photo collection. To ensure that the matching cards can appear in each game, the image set for playing is limited to a photo set of the registered users. Furthermore, the selected image set can be applied with any pattern classification algorithms so that matching pairs have more chance to be slapped or set as a trap.

To enhance the fun element of the game and void coalition formation, players are matched by PhotoSlap according to their ability and no communication among players is allowed. After the game, the images are clustered based on the player actions. Each pair of images is given a confidence value C by $C = W_1(C_s - C_o) + W_2 \cdot C_t$, where W_1 and W_2 are weights and C_s, C_o, and C_t are the counts of slap, object, and trap actions respectively. The clusters are built by linking matching pairs if the confidence value of a pair is greater than a pre-defined threshold. After the process in the game layer, the cluster can be labeled for semantic annotation.

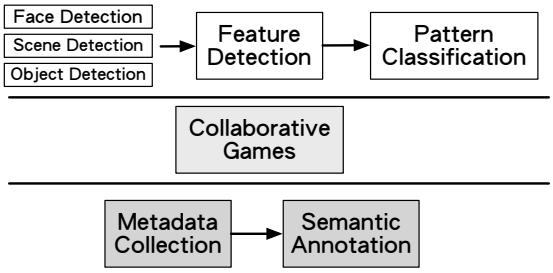


Figure 2: System overview.

System Implementation

The PhotoSlap system is implemented as a web application consisting of the following components:

1. Photo uploading tool with semi-automatic face detection
2. PhotoSlap game
3. Bulk annotation tool

The whole system can be viewed as a collaborative tool for classifying a large collection of photos into several clusters. Users upload their photos as inputs, and it produces several individual clusters. Thus, the clusters can be labeled by users easily. The processing flow of the system is illustrated in Figure 3.

PhotoSlap provides a tool for users to upload their personal photo collections from either Flickr[5] or the local file

[5]http://www.flickr.com

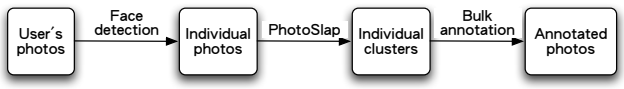


Figure 3: Process flow.

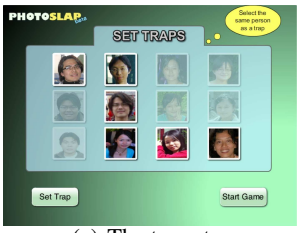
system. During the uploading process, the system performs automatic face detection using OpenCV (Open Computer Vision Library) (Bradski, Kaehler, & Pisarevsky 2005). Since the face detector does not have perfect accuracy, we create a WYSIWYG (What You See Is What You Get) interface (Figure 4) for adding, removing, and editing the detected face regions.

Figure 4: WYSIWYG face editing interface.

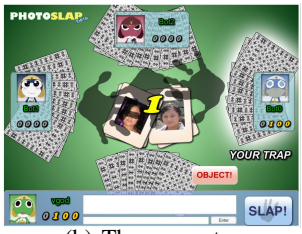
Images of faces extracted semi-automatically are utilized by PhotoSlap in presenting potentially matching images to game players. In turn, images are classified into clusters for different individuals in the photos based on the human actions logged during the game. After the face clusters are generated, users can then use the bulk annotation interface to click on a specific cluster to give it a label.

PhotoSlap

PhotoSlap is implemented in the client-server architecture. To increase the accessibility of PhotoSlap, we adopted Adobe Flash as the client, and implemented the server purely in Python. Upon the completion of a game, the game server logs all activities of each player in the database for future analysis.



(a) The trap stage.



(b) The game stage.

Figure 5: The screenshots of the game.

Evaluation

We have conducted small-scale experiments using 4 focus groups, consisting of 4 users each. For each focus group, users played PhotoSlap for a 30-minute session continuously. Each session produced about 11 games. The test dataset used in the experiments contains 572 faces of various poses and illumination from 24 different persons, and all faces were manually labeled and annotated by the authors. Given the test dataset with the ground truths, the game is evaluated in the following aspects.

Is The Game Fun

At the conclusion of each focus group session, the users were requested to answer a set of survey questions providing feedbacks about playing the game and to write down any specific comments. Based on the data collected from the game play survey, PhotoSlap received an average score of 7.6 points on a 10-point scale. All users claimed that they would like to play again.

How Good Is The Game Strategy

To validate the game strategy analysis, precision and recall are measured. In addition to three player actions (slap, object, and trap), we define and analyze an extra action called "slap-object" which means slapping without any objection. Let $S1_{action}$ be the set of photo pairs applied with the *action* and $S2_{action}$ be the set of photo pairs that should be applied with the *action* according to the target strategy. For example, $S1_{slap}$ is the set of photo pairs being slapped and $S2_{slap}$ is the set of photo pairs which appear consequently in the game and contain the photo of the same person. The precision and recall of a specific action are defined as follows:

$$Precision = \frac{|S1_{action} \cap S2_{action}|}{|S1_{action}|}$$

$$Recall = \frac{|S1_{action} \cap S2_{action}|}{|S2_{action}|}$$

According to the definition, the corresponding precision and recall for each action in the experiments is shown in Table 2.

	Slap	Object	Slap-Object	Trap
Precision	90.20%	77.91%	99.84%	99.70%
Recall	99.05%	98.53%	96.04%	81.73%

Table 2: Precision and recall of the user strategy.

The photo pair is considered to be applied with slap/object if any of the players performs the action. That's the reason why the recall values of them are much higher than their precision values. According to the intuition of the actions, "slap-object" means the players have an agreement on whether the same person is in the two photos. Therefore, the precision and recall of it are both over 96%. In the trap stage, because the players are given 12 photos in sufficient time for recognizing, the precision of "trap" is much better than its recall value.

The experimental results confirmed our intuition and served as a validation of the game strategy analysis. Besides, the result of the high precision and recall value for "slap-object" suggests that the matched photos can be linked in a precise and quick manner.

Is The Game Productive

By combining the results of the actions being taken, the links between face photos will be built and the face clusters can thus be formed. Therefore, the *productivity* of the game is measured by the links being built and the percentage of the correct links. In the focus-group study (8 person-hours), 1480 links are formed in which 1455 links are correct. In other words, each game can produce 12.3 links per minute and 98.31% of them are correct.

Discussion

In the survey of the first version of PhotoSlap, many players mentioned the implementation of the game deeply influences how people enjoy the game and how accurate the data can be collected. The user interface, sound effects, and hot-key specification should be carefully considered. By refining the UI design and hot-key specification, the precision of user strategy increases 5 to 10 % in the latest version.

Conclusion

This paper presented our design of a multi-player online game called PhotoSlap, which explores the power of human computation for semantic annotation of a collection of digital photos. While ESP and its variations have demonstrated the potential of human computation, this research aims to explore the design principles underlying such productivity games and to identify meaningful evaluation metrics. Our contributions can be summarized below.

- PhotoSlap, a multi-player on-line game based on the rules of Snap, has been designed, implemented, and beta released.
- PhotoSlap is complete with a tool for photo uploading (from Flickr or locally) , a WYSIWYG face editing interface with automatic face detection and a bulk annotation tool.
- PhotoSlap supports the *objection* and *trap* steps to encourage truthful input from players. Using game theoretic analysis, we showed that PhotoSlap reaches subgame perfect equilibrium with the target strategy when players are rational.
- Quantitative evaluation metrics are proposed. Experiments involving four focus groups have been conducted, and the results showed PhotoSlap to be fun, conforming to target strategy, and productive in annotating people metadata for personal photo collections.

At this point, PhotoSlap is under limited release with a global release planned for next month. The wider participation will enable larger-scale experiments for comprehensive evaluation of the game. The evaluation metrics can be further refined to provide performance measurements for productivity games in general. The design principles identified in PhotoSlap should generalize to human-computer collaboration in similar tasks. For example, new games may be designed to annotate objects and events as long as people perform better than computers in such tasks. Whether competance/familiarity enhances fun is another important subject of our ongoing investigation.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments. This research was supported by the National Science Council (NSC-95-2622-E-002-018) and by the Program for Promoting Academic Excellence, National Taiwan University (95R0062-AE00-05).

References

Bradski, G.; Kaehler, A.; and Pisarevsky, V. 2005. Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal* 09(01).

Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. Harper and Row, 1st edition. chapter 4.

Duygulu, P.; Barnard, K.; Freitas, N.; and Forsyth, D. A. 2002. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, volume 4, 97–112.

Huang, T.-h., and Hsu, J. Y.-j. 2006. Beyond memories: Weaving photos into personal social networks. In *Modeling Others from Observations: Papers from the 2006 AAAI Workshop*, volume Technical Report WS-06-13. Menlo Park, California: The AAAI Press. 29–36.

Jeon, J.; Lavrenko, V.; and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 119–126. New York, NY, USA: ACM Press.

Kramer, W. 2000. What makes a game good? *The Games Journal*.

Li, J., and Wang, J. Z. 2006. Real-time computerized annotation of pictures. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, 911–920. New York, NY, USA: ACM Press.

Mori, Y.; Takahashi, H.; and Oka, R. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*.

Prensky, M. 2002. The motivation of gameplay: or, the real 21st century learning revolution. *On The Horizon* 5(1).

Russell, B. C.; Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2005. Labelme: A database and web-based tool for image annotation. Technical Report MIT-CSAIL-TR-2005-056, Massachusetts Institute of Technology.

Sinha, P.; Balas, B.; Ostrovsky, Y.; and Russell, R. 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE* 94(11):1948–1962.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proc. CHI '04*, 319–326. ACM Press.

von Ahn, L.; Liu, R.; and Blum, M. 2006. Peekaboom: a game for locating objects in images. In *Proc. CHI '06*, 55–64. New York, NY, USA: ACM Press.

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM Comput. Surv.* 35(4):399–458.