Human Computation: Image Metadata Acquisition based on a Single-player Annotation Game

Jakub Šimko[a], Michal Tvarožek[a], Mária Bieliková[ac]

[a] *Institute of Informatics and Software Engineering,*

Faculty of Informatics and Information Technologies,

Slovak University of Technology in Bratislava,

Ilkovičova 3, 842 16, Bratislava, Slovakia

[jsimko,tvarozek,bielik]@fiit.stuba.sk

[c] Corresponding author:
bielik@fiit.stuba.sk
tel: +421 2 602 91 514
fax: +421 2 654 20 587

Abstract

Effective acquisition of descriptive semantics for images is still an open issue today. Crowd-based human computation represents a family of approaches able to provide large scale metadata with decent quality. Within this field, games with a purpose (GWAP) have become increasingly important, as they have the potential to motivate contributors to the process through entertainment. However, the existing solutions are weak, when specific metadata are needed. In this work, we present a game with a purpose called PexAce, which utilizes manpower to collect tags characterizing a set of given images. Using novel game mechanics, the game is single-player, less prone to cold-start problems and suitable for deployment in the domain of personal imagery. As our experiments show, the game delivers tags that characterize images with high precision (using a posteriori expert evaluation and evaluation against the gold standard: the extended Corel 5k dataset). We also employ the game in the domain of personal images, where very specific metadata are needed for their proper organization (person names, places, events) and show, that the game is able to collect even these kinds of metadata. We show that the key to higher quality metadata lies in combining the fun factor of the game with motivation for personal gain.

Keywords

human computation, game with a purpose, metadata, crowdsourcing, image, semantics

1 Introduction

The proliferation of digital multimedia (especially images) in recent years has been dramatically accelerated by technological advancements in mobile devices and network infrastructure. Consequently, the almost ubiquitous availability of multimedia recording devices brings challenges for effective multimedia creation and delivery. Amongst others, *multimedia metadata authoring* stands out as the enabler of many advanced (semi-)automated approaches supporting multimedia authoring. These include personalized, dynamic and/or interactive multimedia presentation, (semantic) search, runtime customization and tailoring of multimedia to meet specific requirements (e.g., constructing time limited multimedia presentations). The multimedia metadata used by such approaches have to be both of high quality (e.g., descriptive and accurate) and sufficient quantity (i.e., detailed and of high granularity). Consequently, a lot of metadata is often required for a single resource, and metadata for many resources must be acquired.

Personal images represent a multimedia type which is characterized by its limited public availability and high specificity. Due to the availability of personal multimedia creation devices,

personal images grow in numbers and also need to be properly decorated with metadata. This would enable effective search, browsing or even automated storytelling (Zsombori, 2011) for their owners and social circles. While the acquisition of image metadata in general domains (i.e., for general concepts) is problematic, for personal images, it gets even worse: the desired metadata must include information such as person names, places and events, meaningful and specific to the image owners.

Based on these requirements, we identified *image metadata authoring* (for personal images in particular) and its *scalability* as two key challenges, which we address via *collaborative end-user image metadata authoring based on the human computation paradigm*.

Typically, metadata have been created manually by content creators (e.g., photo annotations), provided by devices as subsymbolic metadata (e.g., EXIF metadata supplied by digital cameras) or extracted by automated analysis of the content itself (Duygulu and Barnard, 2002; Feng et al., 2004; Ke et al., 2011; Kuric and Bieliková, 2011; Lavrenko, 2003; Wang and Khan, 2006). Automated approaches however have limits in terms of types and level of detail of the metadata they provide. In the image domain, photo annotation cannot so far be effectively performed via automated approaches and requires manual image annotation, which provides quality metadata but does not scale due to the *lack of motivation* of humans to do it.

Despite continuous effort to improve automated image annotation, some approaches focus on *increasing human motivation* to annotate images manually by providing non-monetary incentives. Web 2.0 and crowdsourcing introduced approaches based on social interaction and collaboration, where users are motivated by the social experience and describe, comment and tag multimedia they share, which in turn enables annotation of a much larger portion of multimedia that would otherwise never be annotated by humans. The drawbacks of social annotation lie in the varying quality of metadata provided by users (e.g., insufficient level of detail, non-descriptive comments, user specific meaning of terms) and more importantly in the limited scope of the acquired metadata since users only interact with content attractive to them. Therefore, there are no existing mechanisms to control what content gets annotated.

To address these drawbacks (i.e., to provide control over the scope of annotated resources and over the quality of the metadata), game-based approaches coined as *games with a purpose* (GWAP) emerged. In this human computation paradigm, various approaches transform *human intelligence tasks* into appealing games (Ho et al., 2009) by aligning the winning conditions of a game with task solving. Games with a purpose serve as an alternative approach for achieving large-scale data processing by humans in cases, when the task is hard to perform by machines (von Ahn and Dabbish, 2008). Unlike in social tagging, GWAPs can control what resources get annotated and how many times.

As their downside, GWAPs have several limitations, related mostly to *cold-start problems* (e.g. many players are needed to play at the same time) and *dishonest player behavior* (as in all games, some players aim to cheat, destroying the experience of other players and also "the purpose"). Despite that, GWAPs represent the most advanced form of human-oriented metadata acquisition (Roman, 2009) and an attractive research field.

To address the image metadata needs and also the existing problems with GWAP design, we devised PexAce – a GWAP with a novel interactive image annotation approach. It is based on the popular visual memory game Concentration, where the players' goal is to find pairs of cards (images) by continuously inverting cards on a game board. The game PexAce serves as a means for:

- *collaborative image metadata authoring* via collecting and evaluating player assigned annotations into metadata – the key aspect of the game is that players are allowed to help their memory by making annotations for images they have seen. In our experiments we have deployed the game as a web application for 107 players and evaluated the output of the game (image annotations) experimentally to show its correctness;
- *dynamic interactive presentation of images*, typically photo albums. The approach may also be tailored to video and audio;
- *entertainment* by engaging players in friendly competition in a game.

This work primarily focuses on the semantics acquisition capabilities of PexAce. We use and evaluate the game in two major use case scenarios:

- *General image annotation*. Using the Corel 5k dataset - a standard dataset for evaluation of image annotation approaches - we collect image tags through the game (with a large player group) and measure their accuracy against a gold standard (Corel 5K dataset tags, extended further by tags provided by experts) and by a posteriori expert evaluation.
- *Using personal images as input*. For the use in this specific domain, we made several modifications to the game and collected tags for images from albums belonging to small social groups, from which the players were recruited. Different members of these groups then a posteriori evaluated the produced tags. We have also conducted a qualitative study on the player experience with the game in this scenario.

Our approach retains all advantages of the GWAP paradigm: It is scalable (by means of available human participants), provides a control over what resources get annotated and also how specific the annotations are, while optimizing the use of human labor (the task solving redundancy is pursued only until a task is solved). Furthermore, we bring new ideas to the GWAP domain:

- We introduce the concept of "helper artifacts", which mitigate the cold start problem of most GWAPs, i.e. a traditional multiplayer GWAP needs a critical mass of players or data upon deployment in order to start working. Our concept, however, is based on the idea that players create the useful value (image annotations) only to help them in the game, but do not get scored directly for their quality. This allows our game to be single-player, alleviating the cold-start problem.
- We show that the quality of artifacts produced by the player is further increased when additional motivation (other than to play a fun game) is present within the game. In our case, when we deployed the game for personal image annotation, the motivation of "*doing the work for themselves*" increased the quality of annotations from individual users and allowed us to set the annotation validation rules to a less restrictive level and achieve the same metadata quality, sparing the manpower for other tasks.

The paper is structured as follows: An overview of related approaches in metadata authoring and games with a purpose (with respect to their design aspects) is provided in section 2. In section 3, we describe our game and metadata extraction approach. We evaluate the effectiveness and scalability of our approach, and the validity of the acquired image metadata for general domain in section 4. In section 5, we discuss the modifications and use of the game for personal image repositories. Lastly, we discuss our findings and experience with games with a purpose as an end-user image metadata authoring platform.

2 Image metadata authoring approaches

2.1 Traditional metadata authoring approaches

The majority of multimedia today is authored without suitable metadata, which have to be acquired later either by automatic or human-oriented means. Despite their heterogeneous nature in terms of quality, automated metadata acquisition approaches are generally used for annotation of large resource collections.

Many approaches aim to identify semantics relevant to content of static images via identification of visual features. All of these approaches involve some degree of supervision. Duygulu and Barnard (2002) employed segmentation of the image and associated identified features within individual segments with words from a large vocabulary. The vocabulary was used afterwards to identify the semantics of the image. Their evaluation over Corel 5K dataset yielded 70% correct prediction. Better results were achieved when a probabilistic model was employed by Lavrenko et al. (2003). Feng et al. (2004) proposed enhancement to the segmentation approach, which employed the co-occurrence of terms related to images (e.g., tiger – grass occurring more frequently than tiger – building), which also improved output correctness but was more bound to the training data set of images. Improvements were also achieved when information about global and local features were used together (Kuric and Bieliková, 2011). Various approaches use machine learning for image or image region categorization. Techniques such as SVM (Cusano et al., 2004) or Bayes point machine (Chang et al., 2003) perform well (precisions over 90% in Corel 5K dataset), but are limited to a small number of categories and lack of training sets to be used effectively for acquisition of more specific metadata. The acquisition of the semantics of multimedia content (visual or aural) may also involve OCR or speech recognition approaches

(Bolettieri et al., 2007). Due to its non-textual nature, metadata acquisition for multimedia resources is often performed via analysis of their context (e.g., in the web environment) which may contain text or already annotated resources (Papadopoulos et al., 2006; Verborgh et al., 2011; Wang et al., 2009). Generally, automated approaches introduce some inaccuracy which makes them difficult to apply to heterogeneous resources. More or less, they also need large training sets of already annotated images, which stress the initial requirement of human labor to create them.

Compared to automated approaches, human-oriented image metadata creation results in high quality annotations. Humans author metadata either as their primary activity (e.g., paid experts annotating press photos for commercial use) or just as a positive side effect of another activity (e.g., tagging photos while socializing on Facebook thus organizing online image galleries or commenting videos). Metadata validity is implied by contributors' trustworthiness (expert work) or via agreement of multiple contributors who produce metadata for the same resource independently (crowdsourcing).

High metadata quality is achieved by dedicated experts, who are aware of the purpose of their activity and annotate resources as their primary job. However, such experts must be properly paid for their work. Although human expert work does not scale, the crowdsourcing paradigm can be applied to many human intelligence task instances in parallel. However, crowdsourcing approaches suffer from the inability to deliver structured and specific metadata, and the prevailing opinion among crowd members may not correspond to the truth (Roman, 2009).

2.2 Games with a Purpose

Games with a purpose (GWAPs) emerged in the crowdsourcing approach family as an interesting research field, first coined and popularized by Luis von Ahn in his ESP Game (2008). The premise of GWAPs lies in the fact that humans often perform non-trivial reasoning and problem solving (i.e., many "human computation cycles") in normal games and have fun. GWAPs take advantage of this human behavior and harness human computation to solve real-world problems in exchange for entertainment instead of monetary values. They are typically used for solving *human intelligence tasks*, which can be characterized as formal problems that cannot be efficiently solved by machines but are easily solvable by humans.

The motivation of playful experience is often used to engage users into solving the human intelligence tasks in a field called gamification (Zichermann, 2010) which is related to GWAPs. The gamification-based approaches involve game elements like leaderboards and achievements into existing working processes. The GWAPs on the other hand, create a new working process as a game: they transform the problem solving into game rules that force players to disclose their knowledge or solve an instance of a problem. Game logs usually serve as raw material, containing potential solutions for the given problem. Through a secondary filtering, only correct solutions are passed as results. This is usually based on agreement of multiple players, typical for crowdsourcing. The GWAPs have in fact became the bright side of crowdsourcing as they offer a more controlled environment for information acquisition than is available in regular, uncontrolled crowdsourcing (Roman, 2009).

Although there are no known limits in terms of domains or problem types suitable for GWAPs, most are used in the Web semantics domain as

- *domain modeling games*, which enrich knowledge bases such as ontologies or taxonomies by collecting fragments of knowledge supplied by players (von Ahn and Dabbish, 2008; Krause et al., 2010; Markotschi and Völker, 2010; Siorpaes and Hepp, 2008)
- *metadata acquisition games*, which create metadata beyond the capabilities of automated approaches for textual pronoun co-reference finding (Hladka et al., 2009; Chamberlain et al., 2009),images (von Ahn and Dabbish, 2008; Ho et al., 2007; Seneviratne and Izquierdo, 2010) or even people (Guy, 2011).

Outside the Semantic Web domain, less conventional GWAPs address FPGA circuit layout optimization (Terry et al., 2009) or protein design (Cooper et al., 2010).

Much attention is paid to multimedia annotation games specifically targeted at images. In the pioneering ESP game, two players have to blindly agree on the same term describing a given image (von Ahn and Dabbish, 2008). Several other games were devised for a similar purpose (Seneviratne and Izquierdo, 2010). The KissKissBan game is a modification of the ESP Game which introduces a third player as an opponent to the remaining two (partnered) players, either to

increase specificity of tags retrieved from the game and also due to anti-cheat purposes (Ho et al., 2009). Another of Ahn's games, Peekaboom, focuses on identification of the exact contours of features in images (von Ahn and Dabbish, 2008).

All games with a purpose must address several common design issues in order to be successful which makes GWAP design non-trivial (i.e., there is no methodology for transforming a formal problem into a game).

Artifact validation. All GWAPs must address one design paradox: they have to give players immediate feedback during (or shortly after) the game to maintain player attention; however, if the feedback (score) is not based on the purpose of the game, players might try to win the game in some other way next time (and not generate useful game logs). But how can the game evaluate the artifacts the player has produced if these are by definition products of a human intelligence task, which cannot be performed by a machine? Some GWAPs use multiple players to validate each other's outputs at the same time (von Ahn and Dabbish, 2008; Ho et al., 2009; Markotschi and Völker, 2010) which introduces a serious *cold start problem* since few players play the game at the beginning and it is impossible to match them effectively together (especially if a condition of the game is, that the players do not know each other, as in the ESP Game). Others *bootstrap* the knowledge from players by validating their behavior with already annotated resources or known facts (Seneviratne and Izquierdo, 2010). In unique cases, an existing exact (Peck, 2007) or approximate (Šimko et al., 2011) automatic validation method can be used to verify artifacts.

For GWAPs, the resolution of cold-start issues is relevant for several reasons:

- Many GWAPs do not survive the initial deployment phase when few players play them – their design is prone to cold start problem which prevents that. Only once a GWAP has enough players, does game design preventing the cold-start problem become irrelevant. While the attraction of more players is a matter of other design aspects, if GWAP design mitigates this problem, it is able to function with fewer players and has better chances to establish itself in the long term.
- If the game is used in a specific problem domain, the pool of potential players shrinks (because only few are able to solve the specific problem), rendering the need for solving the cold start more severe.
- Cold-start prone design may also become impractical for researchers and practitioners who want to experiment with their designs on smaller scale.

Our approach solves the artifact validation issue via what we call "helper artifacts". In our game, useful artifact creation (e.g., annotation of images) is *optional* for the player, i.e. the game is playable without it. However, the player is allowed to *help* himself during the game by creating "helper artifacts", which make it easier for him to succeed in gameplay. For instance to aid his memory, he may write notes about what to do, or as in our case, describe a certain object in the game he interacts with (the images). These additional artifacts are then used as a product of the GWAP. The player needs to create these artifacts with a certain level of quality, so they really help him in the game (which also implies their potential quality from the GWAP's purpose point of view). The main desired characteristic of helper artifacts is that they keep the GWAP single player and significantly diminish the cold-start problem of the game.

Popularity and game throughput. The game popularity is crucial in order to gain enough human power to solve web scale tasks. Currently, only Ahn's ESP game has gained a more significant role (von Ahn and Dabbish, 2008), other games tend to remain in the domain of research and experimentation. However, the popularity of a GWAP is determined mainly by the entertainment it can provide (von Ahn and Dabbish, 2008). The authoring process must be simple and appealing (Foss and Cristea, 2010), i.e. players must not perceive the game as work. For this reason, some games encapsulate their purpose (e.g., common fact creation) into game story, sometimes using more advanced graphics (Krause et al., 2010).

Cheating vulnerability. Cheating and dishonest player behavior is a phenomenon that must be considered in all computer games, including GWAPs. A fraction of players always wants to exploit game rules, various "holes" and bugs or directly interfere with a game's implementation in order to acquire advantages. This damages the fairness of the game and discourages other players from playing the game. In case of GWAPs, it may also damage the problem solving capability of the game. Multiplayer GWAPs mostly use mutual player supervision to detect cheating individuals which appears to work well (Ho et al., 2009). However, this is not applicable for single player

games. In our game, we address cheating via preventive rule restrictions, which some GWAPs implement to prevent known threats (Šimko et al., 2011) and also by a posteriori cheating detection based on detection of inverse correlation between player scores and the value of the artifacts they create.

3 Method for game-based metadata authoring

To address the aforementioned issues of efficient image metadata creation and scalability, we devised a two-stage game-based approach (see scheme in Figure 1) for acquisition of tags characterizing a given set of input images:

1. Users provide textual annotations for untagged images via a game with a purpose called *PexAce* which motivates them to play via fun and friendly competition in a public ranking ladder (see Figure 1, top).
2. The acquired textual annotations are translated into English, lemmatized via WordNet and processed using multiple user agreement rules into meaningful metadata for images (tags) (see Figure 1, bottom).

While the game is run mainly as a client application, the game server is responsible for collecting and refining annotations and also selects the images that are going to be played by new players from a large image pool.

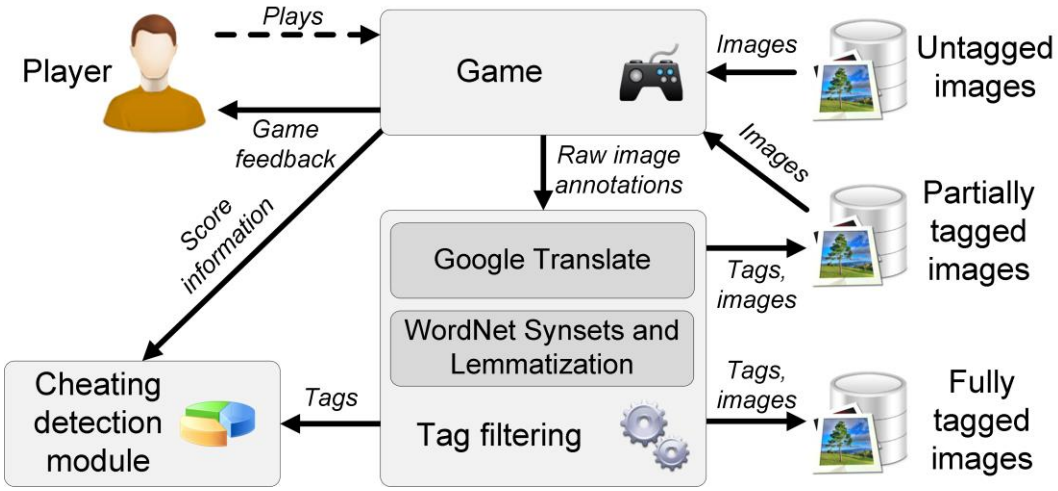

Figure 1 Approach overview. In the PexAce game (top center), players play and create raw logs (image annotations). The processing of these logs to image tags is done in the analysis module (bottom center), which performs basic natural language processing (e.g. lemmatization, translation) over the created annotations and produces tags which undergo the validation procedure based on multiple-player agreement (voting procedure). The game keeps the lowest possible number of partially annotated images that need further processing in-game before being sufficiently tagged.

3.1 PexAce game design overview

PexAce[1], whose purpose is to provide annotations for input images, is a computer adaptation of the popular *Concentration* game. *Concentration* is a turn-based board game for two or more players, who compete to collect the most of the image card pairs, which are mixed and placed on a table facing down (a standard game comprises a board of 8x8 cards). Each turn a player may flip two arbitrary cards to see whether they form a pair. If yes, the player keeps the pair (and receives points for it), otherwise the cards are flipped back and the next player continues. The key to success is to memorize the positions of the cards that have been flipped during unsuccessful attempts (one's own or those of other players) and retrieve them during one's turn.

We devised *PexAce* as a computer modification of the *Concentration* game with a key difference: it is a single player (to address the cold start problem often encountered in other GWAPs). The scoring function has been therefore redesigned and is based on:

[1] Available at: http://mirai.fiit.stuba.sk/pexace

- the *number of card flips* needed to find all pairs, where fewer flips equal more points;
- the *total number of card pairs* present in the game, where more pairs equal more points but also a higher difficulty due to increased board size;
- the elapsed *game time* where shorter games result in more points.

A second key difference makes the game useful: In contrast to *Concentration, we allow* players to "cheat" by adding textual annotations to images acting as "helper artifacts". Whenever a player flips a card, he is allowed to write a short text on it. This annotation is visible for the rest of the game without the need of flipping the card again. This way, the player can take advantage of his annotation when he encounters the second card of a pair for finding the first one. Therefore, he needs fewer card flips to finish the game and receives more points. After the game, the annotations the player has created during the game describe the respective images correctly with high probability. The key to the descriptiveness of the created annotations lies in the fact, that only a correct description of the image truly helps the player to determine, whether a particular (concealed) card contains the image he currently seeks. *The option of using textual annotations motivates the player to annotate images to improve his score.*

Of course, the challenge of memorizing card positions is lowered in *PexAce* (compared to the original *Concentration*). For larger game board sizes however, *PexAce* still presents a challenge of card positions memorizing since a simple "blind" annotation scanning of many image annotations takes a lot of time and that time negatively affects the score (the sizes of game boards are from 4x4 to 10x10, gradually increasing the difficulty). On top of this, a new challenge in writing *effective* annotations is introduced and may become very tricky. For instance, if there are several similar images of mountains present in the game, a simple "mountain" annotation would be insufficient to properly differentiate between them forcing players to either make extra card flips (i.e., lose points) or use *more specific annotations*. This effect may be achieved purposefully: the game can group similar images to game sessions according to already existing tags acquired in previous game sessions.

The PexAce has a competitive aspect which lies in a ladder ranking system. The players compete in achieving the highest possible score and rank in the game ladder, where their total score is cumulative through multiple games. The score contribution for each game is decreased via a hyperbolical function as the number of a player's completed games grows and is also influenced by the number of annotations provided. These factors are public and known to players in order to motivate them to annotate and play as many games as possible.

We implemented PexAce as a web application suitable for mass public deployment on the Web which is the usual choice for most GWAPs. The game interface is shown in Figure 2. A typical gaming session can be summarized as follows:

1. The game starts by initializing the game board with hidden cards facing down and starting the game timer.
2. The player continuously makes turns, flipping two cards at each one.
3. When the first card (in a turn) is flipped, the player can display the annotations on the rest of the cards using the mouse. He moves the mouse over undisclosed cards to display annotations as tooltips and eventually picks a second card to flip.
4. When two cards are flipped, the annotation display is disabled and the game timer is stopped (so the player is not constrained by time during annotation creation). Then, image annotations can be optionally entered for each image. Ending a turn hides the cards, resumes the timer and re-enables annotation display.
5. If an identical card pair is flipped, points are awarded, and the pair remains visible on the board permanently.
6. The game ends once all card pairs are discovered. The player is shown the final score and rank in the game ladder based on the time, number of card flips and total number of cards.

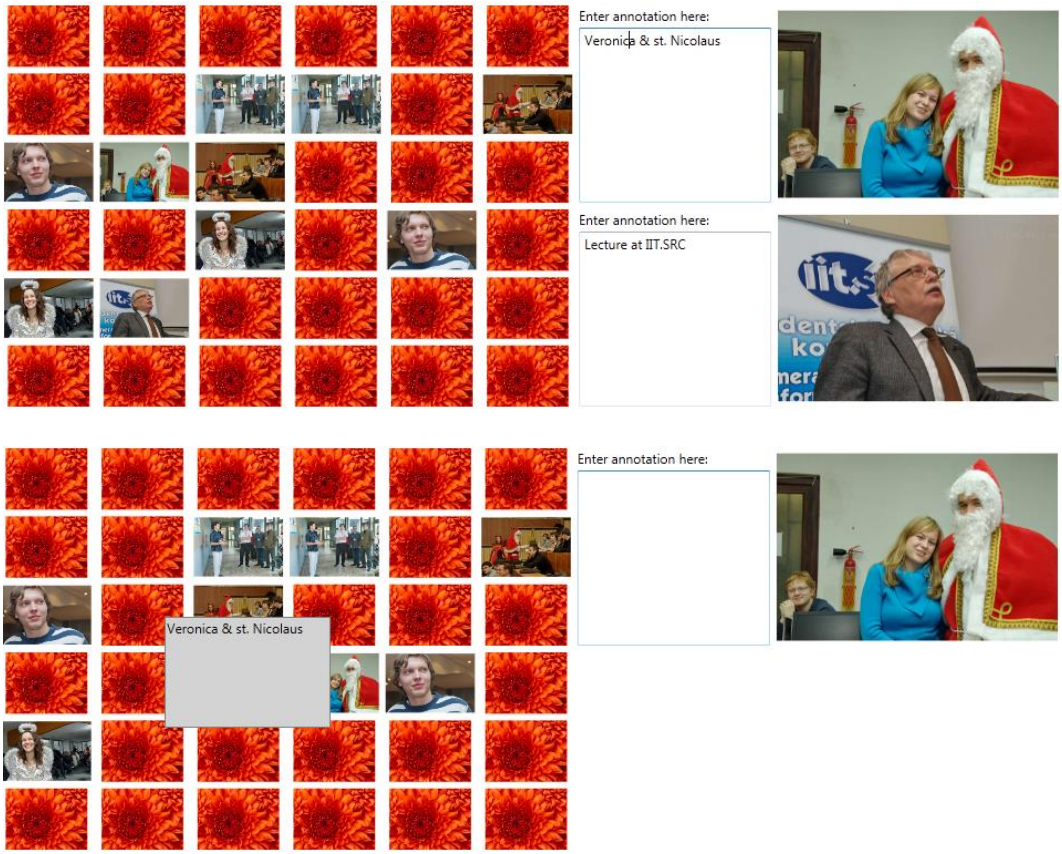

Figure 2 Two examples of the PexAce game interface. In the top screenshot, the player flips two image cards on the game board (left), examines the larger image previews (right) and enters annotations into two text fields (center). Later, when the player comes across an image he has seen before elsewhere (bottom screenshot) he may review his annotations by moving the mouse cursor over hidden cards to acquire the position of the second card of the pair.

3.2 Extraction of tags from image annotations

Textual annotations for images, provided by players, are used as raw material by our tag extraction procedure. Due to their completely freeform nature, user annotations contain both meaningful words but also a lot of noise, i.e., they may be meaningful to their author only, comprise words in various languages (sometimes within the same annotation) and declinations, misspelled words or words with/without accents. We employ multiple player agreement principles and filter tags from user annotations:

1. Each raw annotation is processed via the Google translate service to obtain purely English texts and parsed into word tokens, removing all spaces, commas etc.
2. All resulting tokens are transformed into basic morpheme terms (e.g., plurals to singulars) via WordNet. Occurrence of a particular term is considered only once.
3. Each term then represents a vote – a triplet of *image*, *player* and a *suggested tag*. If the same tag is suggested for an image by multiple users (at least two), we assume that the tag is valid and assign the tag to the image.

This procedure is performed at regular intervals and recursively influences the set of images that are used in the game: images are used in the game until they acquire enough (i.e., a given number) valid tags. However, in case of simpler images players simply use one or two obvious words describing the image. In such cases instead of imposing the valid tag limit, we stop using an image after it was annotated by a certain number of players (in our experiments described later, our goal was collecting either 5 or reaching a maximum of 15 annotations for a single image). The same player may play with the same images again unless they are annotated at least once. Using this strategy, we gradually annotate images from the dataset, maximizing the annotation potential.

3.3 Gameplay extensions

Primarily focusing on the semantics acquisition, we kept the game mechanics plain in our experiments. However, the player experience in PexAce could be improved by several game mechanics extensions, with some also introducing interesting options for image tag creation.

In the original design, player success depends heavily on how images are annotated, marginalizing the memory mechanics (which is scored via time elapsed between card flips, i.e. if the player's memory is bad, he spends more time on seeking annotations). To focus more importance on memory mechanics a time limit period for displaying image annotations could be introduced to the game. After flipping the first card in a turn, the player would only have a few seconds to review existing annotations. Therefore only few annotations could be reviewed forcing players to use their memory to remember card positions at least approximately. Alternatively, the number of times an annotation is displayed to a player in one turn can be limited.

To confront the player with more challenges, an offline competition could be imposed not just through the overall score ladder, but through individual game board setups. The players may compete, who will complete the same game board (same images) in fewer turns. The challenges may be interesting especially against best ranking players. Although the game scoring is independent of the quality of tags extracted from annotations (which we pursued intentionally), a secondary "tag-quality-based" scoring could be imposed later "offline", when multiple players put their tags for a particular image. The rewards and achievements could be delivered to players later, reminding them about the game and inviting to play more. Another scoring mechanics may reward players for novel tags (which is useable especially if there is one dominant concept related to an image, which every player uses). This could be done via an approach similar to the one devised by Mandel and Ellis (2007), where players were rewarded most for tags which were entered before exactly once.

The game could also use already existing image tags or annotations from past games to "auto-tag" some of the images for the player (i.e. the annotations to some images would be available for review upon the game start). This way, the player would be confronted with a less uniform game board, bringing some diversity into the game. Instead of relying on one's own annotations, the player would be tempted to go directly for a second card that has a promising annotation (even if he has not seen it flipped before). Moreover, this would also have a positive effect on tag acquisition: if the player selects a second card only according to its description, than this description may be somehow relevant to the first card (and thus, can be counted as a "vote" for its image during tag extraction).

3.4 Cheating prevention strategies

We have identified certain vulnerabilities of our GWAP which we needed to address. The most prominent cheating threat in PexAce is an automated user interface wrapper. In fact we experienced one case of this type of hack during early stages of game deployment. A user deliberately devised a wrapper application that first uncovered all cards sequentially by executing click events and taking screenshots. Then, it automatically compared images, found pairs and finished the game, without errors, annotations and virtually in no time leading to a nearly perfect score.

To address this threat we devised two anti-cheating strategies:

- The game displays bitmaps of images randomly modified by watermarks, slight rotation, scaling, blurring and color mixing, to make it harder to compare the images automatically. Visually, users hardly spot a difference between images of the same pair, but for machines, the comparison is much harder.
- We used a posterior heuristics (Šimko et al., 2011) to identify dishonest players and suggest them to game administrators based on satisfying some of these automatically evaluated criteria:
    - extremely short time of gameplay,
    - high rank in the ladder,
    - no errors during gameplay (no redundant card flips),
    - no meaningful annotations for images.

A support application constantly scans new game logs to detect dishonest player behavior in the game. Suspicious cases are identified and subject to a game administrators' review, who makes a decision about possible player disqualification. In our experience, all players with extremely good scores, without any contribution to valid tags or even no suggested tags, were cheating the game.

4 Validation of metadata extraction: general domain

To validate our approach for images in the general domain, we conducted precision evaluation against a reference set of pre-tagged images and independently on that, a posteriori expert evaluation of the acquired tags.

As a gold standard, we used the Corel 5k image dataset, a popular dataset in image annotation, in order to be able to compare our approach with others. The Corel 5k dataset consists of 5,000 images, where each photo is assigned 1 to 4 tags (65% of images had 4 tags) that can be considered entirely correct (Ke et al., 2011). Since the default tags in this dataset come in small numbers and do not cover all major image features, we extended the dataset by additional tags acquired through expert work, which was done by three independent experts who tagged the images featured in the experiment. The experts were asked to provide between 10 and 15 English tags for each photo (they had no knowledge about already existing tags assigned to these images). Afterwards, a voting procedure was performed and tags that were suggested by two or more experts were accepted. This way, 2 to 12 additional tags were assigned to each image (with a mean of 6 added tags). The extended Corel 5k dataset then had a tag count range of 3 to 15 tags per image with 8- and 9-times tagged images being the most numerous.

Even after this extension we assumed that there were more tags correctly assigned to images by the game than experts could identify themselves in the gold standard dataset since an image can be described in many different ways (albeit with the same semantics) and different people see the same image differently, focus on different aspects and the context resulting in a broader set of tags. Despite the extensions, we did not have satisfyingly broad sets of tags in the gold standard. Therefore we decided to perform an additional experiment that would directly validate the correctness of the game-produced tags: A group of experts, instead of creating reference tags, validated the existing game-produced tags.

4.1 Game Deployment: Experimental dataset acquisition and its properties

Our experimental dataset of raw annotations from game log data was acquired by publicly deploying PexAce on the Web, featuring the images of the Corel 5k dataset. We propagated it amongst possible users via social networks, word of mouth and organized a local tournament at our university. In total, we have recorded:

- 107 unique players who played a total of 814 games;
- 2,792 images annotated using 22,176 annotations;
- 5,723 tags that passed the voting procedure.
- 1,373 images received enough feedback (either received 15 annotations or at least 5 output tags, i.e. ones that could be inferred by processing the annotations and voting). Out of these 1,373 images, we randomly selected 400 for the purpose of further evaluation.

The structure of the collected data set was not uniform. Due to the greedy algorithm for selecting images: "use a single image until it have 15 annotations from different players or at least 5 output tags", the acquired tags were not distributed uniformly over the entire 5,000 images, but covered a smaller group of images. We selected these thresholds for the sake of experiments; in practice any constants or strategies could be employed for suspending an image from further gameplay. Figure 3 shows how the average number of tags extracted for an image slows with constantly increasing number of annotations for that image.

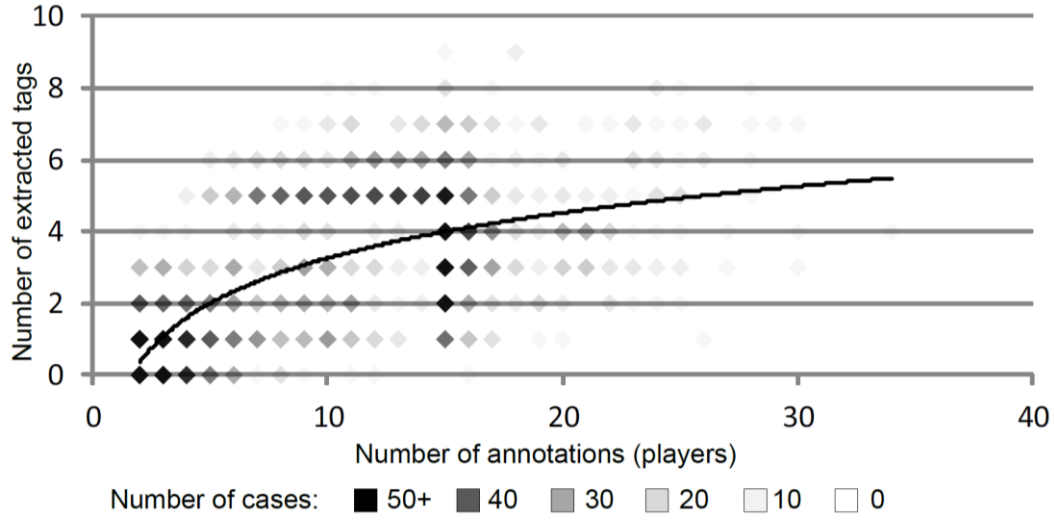

Figure 3 Dependency between the number of in-game image annotations (X-axis) and the number of tags extracted from them (Y-axis). Darker points correspond to more images (black corresponds to more than 50 images). The approximation indicates that the speed of growth of the number of extracted tags slows with the increasing number of annotations.

Another interesting phenomenon we observed in the collected data, were the differences in tag counts for individual images. We believe this was caused by the heterogeneous distribution and varying dominance of certain concepts for particular images. For instance, players found some images simple and distinct enough and assigned them only one tag, especially if there were no other dominant features in the image (e.g. a photo of a horse on a meadow, which has been annotated 15 times still resulted with only one tag: "horse"). Good examples of the opposite cases were images of crowded beaches, where many features were present and players entered a wide variety of words. The differences can be well observed in the Figure 4 which shows the distribution of identified tag counts for images that have been annotated exactly 15 times.

By these observations we see, that there are several cases when the plain strategy of greedy picking images to the game (and stopping their occurrence by thresholds) is causing inefficient use of human work, e.g.. the players enter the same term too many times, or create too diverse set of words. By dynamically detecting the tag suggestions diversity and output tag number increase in the future, the problematic cases could be detected and measures could be made: either the image would be excluded from use in the game (to allow more effective annotation of other images) or special game mechanics be imposed, like the mentioned "similar image feature" (provoking more diverse tags) or "auto-tag feature" (engaging players to validate the existing suggestions in a too diverse tag set).

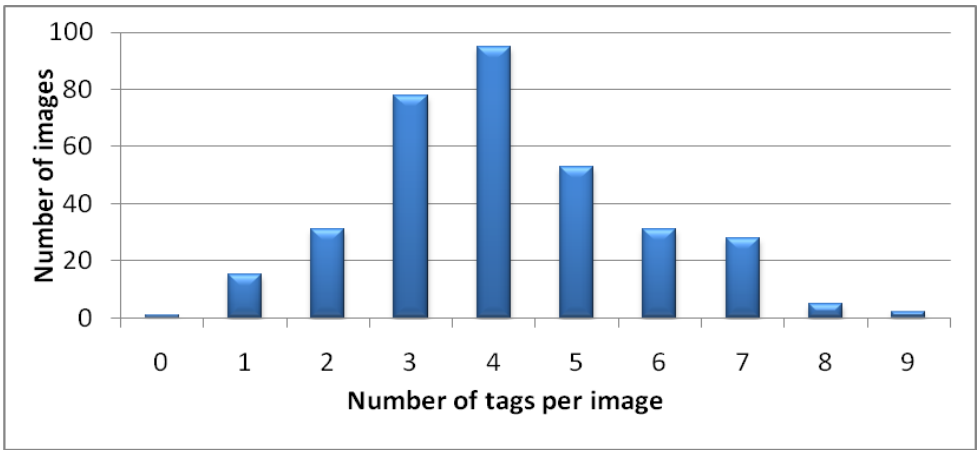

Figure 4 Distribution of identified tag counts for images that were annotated by 15 players.

4.2 Experiment: Validation of annotation capabilities

To validate the correctness of the tags assigned to images by our approach, we measured precision of the extracted tags. In total we performed four experiments:

1. Gold standard precision evaluation over the Corel 5K dataset with original 1-4 tags per image.
2. Gold standard precision evaluation over the Corel 5K dataset containing only tags assigned by experts with an average of 2-12 tags per image.
3. Gold standard precision evaluation over the Corel 5K dataset containing both the original tags and the tags additionally assigned by experts (3-15 tags/image).
4. A posteriori precision evaluation of extracted tags from PexAce, where a different group of experts evaluated the correctness of the produced image tags and rejected invalid tags. This experiment involved direct evaluation of the game-produced tags, which differs from the previous experiments where the actual game assigned tags were not known to experts.

Hypotheses. *The tags acquired through our game are correct*. We expected the majority of tags assigned to images to be correct in experiments 1-3. We expected higher precision for the expanded reference set than with the original reference set (as it contains more tags that can match with game-tags). We also expected that if synonyms are accepted as correct answers (for which we used the WordNet synsets) the precision would further increase. We expected the highest precision in experiment 4.

Participants. All 107 players of the game were adults and regular users of the Web. All of them were Slovak native speakers and majority of them played the game in the Slovak language (the rest played in English). No further knowledge about them was considered. The 6 experts, who participated in validation, were academics familiar with the concept of multimedia metadata.

Data. For evaluation, we used a set of 400 images, annotated by the game. Each of these images had its set of game-produced tags and a set of reference set tags from gold standard dataset.

Process. Precision was computed separately for all three reference sets in experiments 1-3. The total precision values were computed as averages of partial precision values of individual images. The average precisions were also computed for images grouped by the number of tags they had assigned in reference sets (so we can observe the changes of precision as seen in the Figure 5). We wanted to observe the effect of acceptance of synonyms as correct tags, therefore, we computed everything twice, first without and then with the use of WordNet synsets.

For experiment 4, three independently working judges were consecutively shown images with game assigned tags. They were asked to explicitly reject tags they considered incorrectly assigned. Only tags not rejected by *any* of the judges were considered correct. Precision was computed as average precision for each image (number of accepted tags divided by the total number of tags).

Results. Figure 5 shows the average precision for images grouped by the number of reference tags in experiment 3. Table 1 shows an overview of the aggregated precision values for experiments 1-3. The best precision of 73% was achieved with the extended reference set with accepting of synsets; without the use of synsets, precision was also promising at 68%. The resulting precision in experiment 4 reached a very promising 94%.

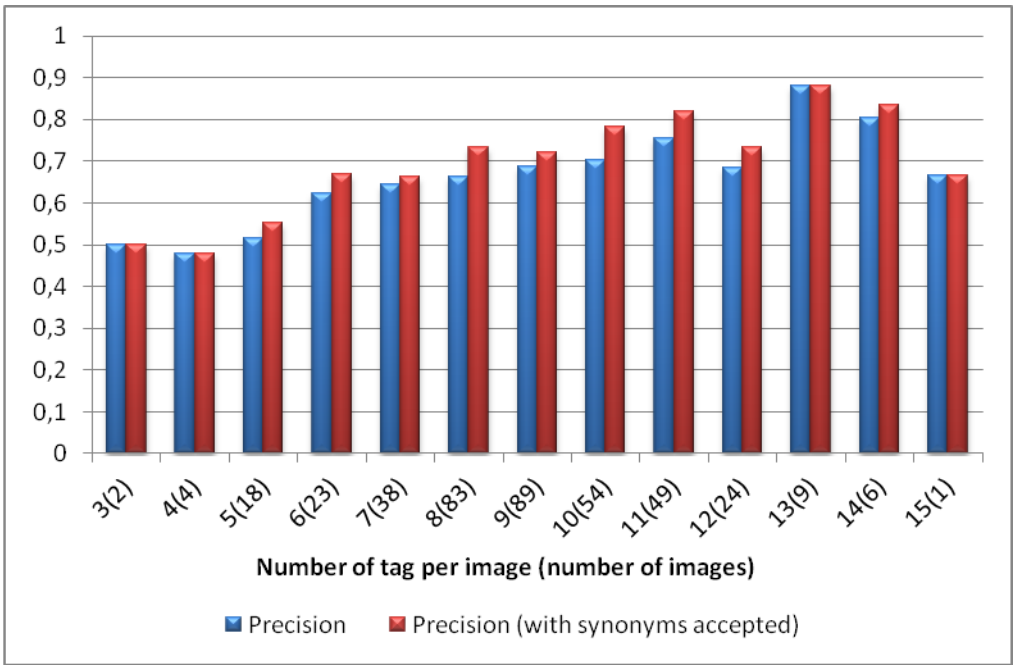

Figure 5 Results of experiments over the joint (Corel + manual) image tag dataset. The graph shows the precision values for images with different numbers of assigned tags with and without the use of WordNet synsets.

Table 1. Results of precision measurement for three reference tag sets and Wordnet synsets. The best achieved precision was 73.03%.

	Precision	Precision+synsets
Gold standard: original Corel only	37.42%	39.66%
Gold standard: Extension only	65.15%	69.96%
Gold standard: original Corel + Extension	68.03%	73.03%
A posteriori tag evaluation	94,0%	N/A

Experiment conclusion. Referring to results above, we claim that PexAce can provide *highly precise tag annotations* for images as the fourth experiment has shown (see Figure 6 for assigned tag examples). The first 3 experiments have shown that not all but the majority of the assigned tags were included in the "first choice" set of the gold standard. We believe these relatively worse results were caused by insufficient broadness of the reference tag sets, which have not contained all possibly correct tags. Furthermore, considering our informal observations of players during the experiment, we can report several other assumptions and future extensions.

We believe that many of the incorrectly assigned tags were introduced by *bad translation* into English. Since players had no language restrictions they mostly used non-English (Slovak) annotations (sometimes without accents, which feature important role in distinguishing of word semantics in the Slovak language). This resulted into translation bias. We expect that restricting annotations to English words or sentences would therefore lead to better results. As a future extension of our approach auto-complete functionality could be considered, which would proactively suggest English words or even concepts (if the auto-complete box was backed by a domain model) and reduce the bias.

The role of PexAce as an *image presentation tool* was also apparent. We have observed players who besides playing, enjoyed the images and took longer pauses between card flipping and watched the images just for their enjoyment. Later this led us to use PexAce as a *presentation and image annotation tool for personal image collections*, especially for families or groups of friends, where it can ease the unpopular task of image organization.

Figure 6 Tag examples for four random images from the Corel 5k dataset assigned by PexAce.

While we have been unable to completely eliminate the threat of automatic user interface wrappers (i.e., more sophisticated image similarity algorithms can be used), we were able to introduce sufficient posterior cheating detection heuristics. We consider the combination of active anti-cheating measures and a posterior cheating detection to be a good general solution for games with a purpose. Since GWAPs aim to create virtual artifacts for solving human intelligence tasks, it is unlikely that an automated cheating method would produce them in sufficient quality (if so, there would be no need for the game). Therefore, player actions earning highest scores in the game can always be examined whether they lead to the creation of useful artifacts. If not, they may be subject to disqualification. Problem and/or game specific artifact quality evaluation methods – player agreement (von Ahn and Dabbish, 2008; Ho et al., 2009) or an automated way to validate artifacts (Peck et al., 2007)) can help in pre-filtering of the suspected players.

5 PexAce for personal images

Apart from the general domain, we see potential for the use of our approach for images in a specific domain such as a personal multimedia repository.

The currently prevailing practices of the majority of users as multimedia creators are described well by a qualitative study of Vainio et al. (2009). When asked, users admitted that they like to interact with their content (watching, but also editing), but not with its metadata. They recognize the value of metadata, especially in larger repositories (which many of them possess), but are generally not willing to systematically create them. The most common practice in organizing a personal image repository is just labeling the whole collections (albums) with names or sorting images chronologically. Therefore, users must rely on time-consuming sequential browsing when trying to find a particular image. An important finding was, that users would primary welcome metadata about persons in the images, person who took the pictures and situational context of pictures (e.g. information about places and events) (Vainio, 2009). The latter strongly implied the directions of our research: in our study, we examine the performance of our approach with these types of metadata.

In the second use case scenario, we deployed *PexAce* to work with personal images. The key differences to the general domain scenario were:

- The desired metadata are much more specific than in the general domain.

- Unlike the general domain, there is only a very narrow group of potential metadata contributors (players) that are capable to provide these metadata.

These are clear disadvantages to any crowdsourcing process (including GWAP). The key to overcome these lies in player motivation:

- Players are challenged by the game (a motivation which is present also in the original version). They wish to reach the highest possible score to best themselves or members of their social circle.
- Players interact with familiar content, i.e. personal images (which is an enjoyable experience). Therefore they know specific information about them and also their context.
- Players annotate their images. Players who know about the true purpose of the game - annotating their own albums - provide annotations not just for the sake of the game, but also with respect to the future use of tags in image search and organization.

The idea is to let users play an image annotation game, but with their own images and for their sake (metadata, which can be later used e.g. for better image search). Through this, the average quality of annotations would be higher (they would be more accurate and specific) so the need of redundant task solving (for cross-player artifact evaluation) would decrease. Ultimately even, *a single player would be capable to annotate his image album alone by enjoying the game*. The game also aims to exploit a possible scenario when a player is sharing the game with his social circle (to share the images and acquire more metadata for his use).

For this scenario, we extended the original game by adding to the original cross-player tag validation (two player agreement) several less restrictive heuristics enabling the game to accept tags entered only by a single player. We have performed a combined quantitative-qualitative study to evaluate this approach. We show that the game is able to provide *valid* and also *image-owner-specific* metadata to images used in the game, while retaining a key feature of *low number of participants needed to play*. We further examine the types of tags extracted as well as performance of the game over different types of images. Also, with respect to the possible scenario of sharing the game within one's social circle, we measured how the game performs, when players are unaware of its purpose (but still play over "their" images).

5.1 The tag extraction modifications

The result of the game phase of our extended approach is the same as with the original game: a set of raw player-made textual annotations assigned to images. The only difference is, that the images may (and for best performance should) be assigned to albums. After basic NLP - tokenization, lemmatization and stop-word removal – they are used to create tag suggestions. A tag suggestion is a hypothetical "vote" of a player for a certain term to be connected with an image. The follow-up processing then tries to assume, how strong that relatedness is.

The heuristics that estimate if the individual terms should be assigned to the images were defined as follows. Given a *tag assumption* which is a quartet of *player*, *term*, *image* and *album* identifiers, an image is decorated by a certain tag:

- If a certain number (in our experiments we used 2) of players agree on the same suggestion for the same image. This was the validation method used also in PexAce, yet as a traditional crowdsourcing heuristics was very restrictive. It relies on a sufficiently large number of players annotating the same image. With the low number of players in case of personal image annotation, the method is prone to accepting only few tags, rendering the approach inefficient. Other heuristics were therefore devised.
- If the same user repeats the same suggestion multiple times (we used 2) on the same image. Sometimes, players encounter the same image in different game sessions and provide annotations containing common terms. Repetition might indicate (considering also the player's motivation) that the term is relevant to the image, even if it was not confirmed by another player.
- If there is a suggestion used by the same player over multiple images (we used 4) in the same album. Here, the heuristics counts on a widespread practice of image owners to organize their images into album directories. It assumes that a concept or term might recur within multiple images in an album or even in the whole image set. In particular, this might be the case for some owner-specific information like person names.

5.2 Experiments

The purpose of our experiments and the associated study was to explore the capabilities of our method for personal images metadata acquisition. We conducted it via a combination of qualitative and quantitative methods. Quantitatively, we measured correctness, specificity and understandability of tags acquired through our method. The qualitative study examined further characteristics of gameplay and of the produced metadata and was conducted as an interview, using a similar methodology as used by Vainio et al. (2009).

Hypotheses and questions. Assuming gameplay over personal images and a low number of players:

- Do tags generally describe the images involved in the game?
- Do tags address the specific context of the social group they belong to?
- How is the tagging performance affected, if players are aware of the true purpose of the game (e.g., annotation of their own images)?
- What types of group-specific metadata players express through their annotations (e.g., person names, places, events)?
- Over which types of images does the approach work (e.g., portraits, groups, situational images, no-persons images)?

Participants. The study was conducted with 8 participants split into two groups. In each group, participants belonged to the same social circle, having common interests and also images they were familiar with. In each group, two participants were designated as players, one as a judge for tag evaluation and one helped with preparation of the image data set.

Data. For each group, a set of 48 images was created. Images were drawn from three albums belonging to the group with the help of a group member. Each image was categorized either as *portrait*, *group*, *situational* (may involve persons, but also have strong dynamics) and *other/non-person* (e.g. architecture, landscapes, objects). The albums and images were selected in a manner that from each album, an equal number of images were drawn from each category.

Methodology. The game experiment and study were controlled and were performed as individual interviews. We strictly followed prepared scenarios for each type of participant. The players were explained the rules and features of the game. Prior to that, in one group, they were also introduced (as in Vainio's study) to the concept of image metadata and also about the true purpose of the game -- creating annotations for themselves. This was done, so we could measure the impact of this knowledge with reference to the other group. All the players were free to speak at any time during gameplay and after it, not just to answer our questions. These expressions were evaluated qualitatively later.

As for the judges, they were introduced to the concept of (personal) image metadata, but not about the game. We did this to keep judges as objective in the validation process as possible.

Process. The player interviews with gameplay were performed and Slovak was used as the language of choice. During the interviews, each player played three games (with board sizes 6x6, 8x8 and 10x10), thus annotating each image twice (in 10x10, the two pairs remaining for 48 prepared pairs have been drawn randomly and were not considered in the experiment). After all games have been played, the tag extraction procedures were run. Then, the judges evaluated each tag assignment for their group, answering questions:

- Is the tag *correctly describing* the image?
- If it is a correct tag, is it also *specific* for the group (e.g. probably not discoverable by non-group player)?
- If it is a correct tag, is it *understandable* for a non-group member (e.g. for a portrait image an assigned name is a self-explanatory and understandable tag)?
- If this is a specific tag, of which *type* is it (choose one of the followings: a person name, a place name, an event where a picture was taken, other)?

Results. Throughout the experiment, a total number of 366 tags were extracted using the three heuristics mentioned above. The “traditional” *cross-player-voting* heuristics (used also in the original PexAce) yielded (as expected) just one third of this number (exactly 122 tags). The rest was identified by the second (196) and third (48) heuristics (in that order, i.e. already identified

tags were skipped by latter methods). This shows a *major increase in tag output quantity* for our extension of the original PexAce game - the traditional validation heuristic was weak, but the use of less restrictive heuristics paid off and increased the tag gain.

The quantitative results of our experiment are summarized in Table 2. For both groups together, the correctness of tags is about 90% (96% resp. 84%) and therefore we consider our approach as able to acquire valid tags. On average, about 38.5% (44% resp. 33%) of correct tags were social-circle-specific, so our approach was also able to produce tags valuable for personal archives.

Table 2: Table shows the summary results of image tag evaluation (correctness, specificity for the social group and understandability by group non-members) for image sets of both social groups (aware or unaware of the game's purpose).

	Aware (253 tags)			*Unaware (108 tags)*		
	Corr.	*Spec.*	*Und.*	*Corr.*	*Spec.*	*Und.*
Portraits	0.98	0.61	0.71	0.77	0.53	0.87
Groups	0.97	0.57	0.74	0.76	0.45	1.00
Situations	0.92	0.41	0.77	0.93	0.19	1.00
Other	0.98	0.18	0.82	0.88	0.15	1.00
Average	**0.96**	**0.44**	**0.76**	**0.84**	**0.33**	**0.97**

Considering the total number of tags produced by each group and the relative correctness, and the specificity of tags, we can see that both groups provided some value. The group where players were aware of the purpose of the game produced significantly better tags in absolute quantity (total number of tags produced by the aware group was 2.5 times larger) and relative quality. Only the understandability factor is reverse, which we explain to be a consequence of higher absolute number of specific tags passing through and having exclusive meaning (for instance, there was a tag carrying the name of event related to images, not known to people who did not participate). We conclude that awareness about the purpose significantly improves the method's results, however as our qualitative study has shown, players were also less enthusiastic about gameplay when they were aware of the purpose.

According to judges, the specific tags mostly involved person names (53%). In lesser counts events (21%) or places (15%) were present too. The rest (11%) was categorized as “other” and involved mostly features originating from humor related annotations that players were using to entertain themselves. The humor is also a possible difference between the performance of social groups in terms of specificity of tags for the *situational* type of images. In the “unaware group”, the humor was present within the player annotations and resulted in term matching that was not so successful for specific tags (with the far more disciplined “aware group” on the other hand). Both groups (as was expected) performed well with *portrait* and *groups* types and were quite unsuccessful with *non-person (other)* type of images.

As a part of the qualitative study, we observed that both groups enjoyed playing over their own images, particularly with those, they have not seen for longer time. A possible negative feature of the game was also detected: the players were confused and skeptical about having the same image in more than one game in short succession (which happened because of the design of our experiment). They reported, that they tended to use exactly the same text for annotation in both cases (which could potentially harm the term extraction process) and that they were not sure, whether they had seen the image in the current or a previous game. A future work image picking algorithm would have to take into account the last date-of-use for each image and player.

6 Conclusions

We presented our original *image metadata authoring approach* consisting of the interactive game PexAce – a single-player modification of the popular board game Concentration, the purpose of which is the acquisition (authoring) of textual metadata for images.

Our primary contribution lies in devising an *approach for image metadata authoring* via an interactive image annotation game. It motivates humans by non-monetary incentives to provide image metadata qualitatively comparable with expert annotations for resources (images) that would otherwise remain untouched. We also claim success in devising PexAce as an *image*

presentation tool and as a means of providing players with *entertainment*. The main benefits of our approach include:

- *Valid metadata acquisition* via processing of freeform text annotations supplied by players (even in multiple languages) into correct tags with 94% precision as judged by human experts.
- *Effective use of human labor* within the game. The method prevents unnecessary redundancy of human work by using thresholds to measure the need for further annotation effort for a particular image.. Nevertheless, the heuristics we used were very basic and more sophisticated tag-diversity measures might be used.
- Our approach *does not suffer from the cold start problem* and only needs a set of images to work, which is generally hard to achieve in GWAP design.
- *Implicit scalability* of our approach due to the possibility of massive parallelization. The only limit is imposed by the number of active players motivated to play, similarly to other existing GWAPs. The motivation aspect of our approach is further supported by the ability to work with players' own image collections strengthening the interest and motivation to participate via tailoring the presented content to specific user preferences which has not yet been addressed by other GWAPs.
- *Anti-cheating heuristics* for automatic detection of dishonest player behavior and the corresponding design tips for game designers.

A modification of our approach is also *capable of delivering metadata suitable for personal image archives*. Its major advantage is that it can work with a limited number of participating players. This makes the game *suitable to deploy within small social-circles* or ultimately also *for individual users*. The small number of players is substituted by a unique combination of player motivation to provide *valid annotations* - the game and competition experience as well as promise of working on one's own metadata (resp. image retrieval capabilities). We have further shown that:

- The quantity of tags gained by our method was three times higher, than it would have been if we had only used cross-player-voting (the original PexAce approach). Our method is therefore capable of delivering tags in higher quantities while retaining a comparable quality of tags (even if we assume, that cross-player-voting contributes the 100% quality, the latter extraction heuristics would have at least 85% output correctness over the rest of tags).
- The difference caused by increased motivation and awareness of the game's purpose have significant positive impact on the overall approach performance in terms of tag quality and quantity (in contrast with other GWAP-related works stating the exact opposite (Krause, 2010)). A minor drawback was the lesser enthusiasm for playing in the purpose aware group.
- The majority of specific tags assigned corresponded to person names and to a lesser extent the names of places or events.
- The game delivers correct tags for all image categories, but in case of images without any persons, the number of social-circle-specific tags is low, which limits the usefulness of our approach for this type of image.

What are the lessons learned for GWAP design in general? In this work, we a priori considered several design aspects of our game which influenced its outcome. We have given focus on cheating vulnerability and motivation to play (demonstrating the potential of non-monetary, yet personal gain a GWAP can deliver to its players). The major focus however was on the artifact validation scheme of the game, which resulted into the "helper artifact" principle. We regard this principle as useful for future GWAP designs, especially those that are derived from existing "non-purposeful" games (such as PexAce, which is derived from the Concentration game). While with other artifact validation schemes, the game rules must usually be designed "from scratch", with "helper artifacts", the designer may take advantage of the original game mechanics (and aesthetics) and simply inject the "helper" mechanics into the game. Naturally, one has to beware of damage to the original balance of game rules and the "original" game must already contain a proper mental challenge that synergizes with its new purpose, but it still may be the easier way.

Yet a posteriori, we have also explored the potential for a more proper task selection (image picking) scheme in our game. This includes the selection of when to start and end the solving of a particular task (image) and also the use of player skill and expertise models to improve the game outcome, which in general is applicable to all GWAPs. Currently to the best of our knowledge, no game with a purpose takes into account, what degree of expertise the player has for a particular

task instance. Knowing this might help to effectively assign the right tasks to players and exploit their expertise in a particular domain. The “player model” could be built implicitly by measuring how many tags which a player had contributed actually became valid (accepted by the voting procedure). This information would indicate the expertise for a particular set of images or a domain. Regarding this, the preliminary examinations of the recorded logs show differences between individual players concerning their participation on the consensus with other players and also their real “usefulness” (i.e., whether they were true suggestions). Exploring these differences would be a perspective direction for further research.

Our approach can also be used in other domains as it is not limited to images. The game can be modified for the annotation of other multimedia such as sound or music streams (the description of music is an interesting intellectual challenge for players). We also consider our approach to be applicable to video streams (e.g., by selecting suitable screenshots).

Acknowledgements. This work was supported by grants No. VG1/0675/11, APVV 0208-10 and it is a partial result of the Research and Development Operational Program for the project Support of Center of Excellence for Smart Technologies, Systems and Services, ITMS 26240120005 co-funded by ERDF.

The authors wish to thank colleagues from the Institute of Informatics and Software Engineering and all students (members of the PeWe group, pewe.fiit.stuba.sk) for their invaluable help in experimental evaluation of the work presented in this paper. A special thanks belongs to Balász Nagy for his contribution to the PexAce game idea and implementation.

References

1. Ahn, L.v., Dabbish, L., 2008. Designing games with a purpose. Communications of the ACM 51, 58-67. doi:10.1145/1378704.1378719
2. Bolettieri, P., Falchi, F., Gennaro, C., Rabitti, F., 2007. Automatic metadata extraction and indexing for reusing e-learning multimedia objects, in: Workshop on multimedia information retrieval on The many faces of multimedia semantics - MS ’07. ACM, New York, NY, USA, pp. 21-28. doi:10.1145/1290067.1290072
3. Chamberlain, J., Poesio, M., Kruschwitz, U., 2009. A demonstration of human computation using the Phrase Detectives annotation game. Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP ’09. ACM, New York, NY, USA, pp. 23-24. doi:10.1145/1600150.1600156
4. Chang, E., Goh, K., Sychay, G., Wu, G., 2003. CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. In: IEEE Transactions on Circuits and Systems for Video Technology, vol.13, no.1, pp.26-38. doi 10.1109/TCSVT.2002.808079
5. Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, AC., Beenen, M., Salesin, D., Baker, D., Popović, Z., 2010. The challenge of designing scientific discovery games. Proceedings of the Fifth International Conference on the Foundations of Digital Games - FDG ’10, ACM, New York, NY, USA, pp. 40-47.. doi 10.1145/1822348.1822354
6. Cusano, C., Ciocca, G., Schettini, R., 2004. Image annotation using SVM. In: Proceedings of internet imaging V. SPIE, Bellingham, pp. 330-338. doi:10.1117/12.526746
7. Duygulu, P., Barnard, K., 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: 7th European Conference on Computer vision. Springer-Verlag, London, UK, pp 97-112.
8. Feng, S.L., Manmatha, R., Lavrenko, V., 2004. Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the International Conference on Computer vision and pattern recognition – CVPR '04. IEEE Computer Society, Washington DC, pp 1002-1009.
9. Foss, J.G.K., Cristea, A.I., 2010. The next generation authoring adaptive hypermedia: sing and evaluating the MOT3.0 and PEAL tools. Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10. ACM, New York, NY, USA, pp. 83-92. doi:10.1145/1810617.1810633
10. Guy, I., Perer, A., Daniel, T., Greenshpan, O., Turbahn, I., 2011. Guess who?: enriching the social graph through a crowdsourcing game. Proceedings of the 2011 annual

conference on Human factors in computing systems - CHI '11. ACM Press, New York, New York, USA. 1373-1382. doi 10.1145/1978942.1979145
11. Hladka, B., Mirovsky, J., Schlesinger, P., 2009. Designing a language game for collecting coreference annotation. Proceedings of the Third Linguistic Annotation Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, 52-55.
12. Ho, C.J., Chang, T.H., Lee, J.C., Hsu, J.Y., Chen, K.T., 2009. KissKissBan: A competitive human computation game for image annotation. Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '09. ACM, New York, NY, USA, pp. 11-14. doi:10.1145/1600150.1600153
13. Ke, X., Li, S., Cao, D. 2011. A two-level model for automatic image annotation. Multimedia Tools and Applications. Vol. 61, Issue 1, pp 195-212. doi:10.1007/s11042-010-0706-9
14. Krause, M., Takhtamysheva, A., Wittstock, M., Malaka, R., 2010. Frontiers of a paradigm - exploring human computation with digital games. Proceedings of the ACM SIGKDD Workshop on Human Computation – HCOMP '10. ACM, New York, NY, USA, pp. 22-25. doi:10.1145/1837885.1837893
15. Kuric, E., Bielikova, M., 2011. Automatic image annotation using global and local features. In Proceedings of the International Workshop on Semantic Media Adaptation and Personalization – SMAP '11, IEEE Computer Society, Los Alamitos, pp 33-38. doi 10.1109/SMAP.2011.14
16. Lavrenko, V., Manmatha, R., Jeon, J., 2003. A model for learning the semantics of pictures. In: Proceedings of the 16th Conference on Advances in neural information processing systems (NIPS '03). The MIT Press, pp 553–560.
17. Markotschi, T., Völker, J., 2010. GuessWhat?!–Human intelligence for mining Linked Data. Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data (KIELD) at the International Conference on Knowledge Engineering and Knowledge Management (EKAW), RWTH pp. 28-39.
18. Mandel, M.I., Ellis, D.P.W., 2008. A web-based game for collecting music metadata. Journal of New Music Research, 37; 2; 151–165.
19. Papadopoulos, G.T., Mylonas, P., Mezaris, V., Avrithis, Y., Kompatsiaris, I., 2006. Knowledge-assisted image analysis based on context and spatial optimization. International Journal on Semantic Web and Information Systems.Vol. 2, Issue 3, 17-36. doi:10.4018/jswis.2006070102
20. Peck, E., Riolo, M., Cusack, C., 2007 Wildfire Wally: Volunteer computing using casual games. Proceedings of the 2007 Conference on Future play - Future play. ACM, New York, NY, USA, 241-242. doi:10.1145/1328202.1328250
21. Roman, D., 2009. Crowdsourcing and the question of expertise. Communications of the ACM 52, 12-12. doi:10.1145/1610252.1610258
22. Seneviratne, L., Izquierdo, E., 2010. An interactive framework for image annotation through gaming. Proceedings of the international conference on Multimedia information retrieval. ACM, New York, NY, USA, 517-526. doi:10.1145/1743384.1743473
23. Simko, J., Bielikova, M., 2012. Personal Image Tagging: a Game-based Approach. In: Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12). ACM, New York, NY, USA, 88-93. doi 10.1145/2362499.2362512
24. Simko, J., Tvarozek, M., Bielikova, M., 2011. Semantics Discovery via Human Computation Games; International Journal on Semantic Web and Information Systems (IJSWIS); vol. 7; issue 3; 23-45. doi 10.4018/jswis.2011070102
25. Siorpaes, K., Hepp, M., 2008. Games with a purpose for the semantic web. IEEE Intelligent Systems 23, vol 3, 50-60. doi:10.1109/MIS.2008.45
26. Terry, L., Roitch, V., Tufail, S., Singh, K., Taraq, O., Luk, W., Jamieson, P., 2009. Harnessing human computation cycles for the FPGA Placement Problem. ERSA. 188–194.
27. Vainio, T., Väänänen-Vainio-Mattila, K., Kaakinen, A., Kärkkäinen, T., Lehikoinen, J., 2009. User needs for metadata management in mobile multimedia content services. In Proceedings of the 6th International Conference on Mobile Technology, Application & Systems (Mobility '09). ACM, New York, NY, USA, Article 51, 8 pages. doi 10.1145/1710035.1710086
28. Verborgh, R., Deursen, D., Mannens, E., Poppe, C., Walle, R., 2011. Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. Multimedia Tools and Applications. Vol 61, Issue 1, pp 105-129 doi:10.1007/s11042-010-0709-6

29. Wang, L., Khan, L., 2006. Automatic image annotation and retrieval using weighted feature selection. Multimedia Tools and Applications. 29, 1, 55-71. doi:10.1007/s11042-006-7813-7
30. Wang, Y., Mei, T., Gong, S., Hua, X., 2009. Combining global, regional and contextual features for automatic image annotation. Pattern Recognition. vol. 42, issue 2, 259-266. doi:10.1007/s11042-006-813-7
31. Zichermann, G., Cunningham, C., 2010. Gamification by Design. O'Reilly Media.
32. Zsombori, V., Frantzis, M., Guimaraes, R.L., Ursu, M.F., Cesar, P., Kegel, I., Craigie, R., Bulterman, D.C.A., 2011. Automatic generation of video narratives from shared UGC. In Proceedings of the 22nd ACM conference on Hypertext and hypermedia - HT '11. ACM, New York, NY, USA, 325-334. doi 10.1145/1995966.1996009

Author biographies

Jakub Šimko received his Master degree (with cum laude) in 2010 from the Slovak University of Technology in Bratislava. Since then, he has been working as a PhD. student and researcher at the Institute of Informatics and Software Engineering at the Faculty of Informatics and Information Technologies at the same university. His research interests are in the area of domain modeling, exploratory information retrieval, semantics acquisition, human computation and games with a purpose. He has co-authored over 15 papers published in scientific journals and proceedings, and presented his work at several conferences including some supported by ACM and IEEE. He is a member of the Slovak Society for Computer Science and ACM SIGWEB.

Michal Tvarožek received his Master degree (with magna cum laude) in 2007 and his PhD. degree, awarded with the Rector's award, in 2011, both from the Slovak University of Technology in Bratislava. Since then, he has been working as an assistant professor at the Faculty of informatics and information technologies at the same university. His research interests are in the areas of exploratory information retrieval, adaptive user interfaces, personalized web-based systems and user modeling. He has co-authored over 40 papers published in scientific journals and proceedings, and presented his work at several conferences including some supported by ACM, IEEE and IFIP. He ranked as the 2nd place winner in the ACM SRC Grand Finals 2010 with his work on "Personalized Semantic Web exploration based on adaptive faceted browsing". He is a member of the Slovak Society for Computer Science and ACM.

Maria Bieliková received her Master degree (with summa cum laude) in 1989 and her PhD. degree in 1995, both from the Slovak University of Technology in Bratislava. Since 2005, she has been a full professor, presently at the Institute of Informatics and Software Engineering, Slovak University of Technology in Bratislava. Her research interests are in the areas of software web-based information systems, especially personalized context-aware web-based systems including domain and user modeling and social networks. She co-authored over 70 papers in international scientific journals and she is editor of more than 30 proceedings, 7 of them published in Lecture Notes in Computer Science series of Springer. She is a senior member of IEEE and ACM, a member of the IEEE Computer Society and the International Society for Web Engineering.