Human Computation and Serious Games: Papers from the 2012 AIIDE Joint Workshop
AAAI Technical Report WS-12-17

Finding Image Regions with Human Computation and Games with a Purpose

M. Lux, A. Müller, and M. Guggenberger
Alpen-Adria Universität Klagenfurt
Universitätsstraße 65-67
9020 Klagenfurt, Austria

Abstract

Manual image annotation is a tedious and time-consuming task, while automated methods are error prone and limited in their results. Human computation, and especially games with a purpose, have shown potential to create high quality annotations by "hiding the complexity" of the actual annotation task and employing the "wisdom of the crowds". In this demo paper we present two games with a single purpose: finding regions in images that correspond to given terms. We discuss approach, implementation, and preliminary results of our work and give an outlook to immediate future work.

Automatic image understanding is limited. In image and visual information retrieval the main problem is defined as the *semantic gap*, which is the gap between low level features, like color, texture and shapes, and high level understanding like image semantics. Not only the image as a whole cannot be interpreted solely by algorithms, but also relation between image regions and semantics cannot be found in a fully automated way. Many applications would profit from an algorithm determining which pixels of an image correspond to which semantic concepts. Figure 1 gives an example photo on the left side. The photo (taken from flickr.com, from one of the authors' photo stream) is tagged *squirrel*. However, the semantic concept *squirrel* is covered by the small region on the right hand side, a fraction of the original image. With this information algorithms for saliency maps on images, e.g. (Itti, Koch, and Niebur 1998), image re-targeting, e.g. (Avidan and Shamir 2007), etc. could be leveraged. Note at this point that the region heavily depends on the concept given by the terms. Annotations like *squirrel on a handrail* or *squirrel in a park* need significantly larger regions in our example in figure 1.

Automated localization of concepts has limited success. (Lampert, Blaschko, and Hofmann 2008) for instance report average precision values of 0.223 and 0.148 for two selected concepts on the PASCAL VOC 2006 data set. So besides the limitation in precision and recall there is also a limitation in the number of concepts to be localized. (Liu et al. 2011) report higher precision and recall values for finding salient regions in images, but while this is a generic approach for a broad range of objects and concepts, it's based on salient objects and does not assign labels to salient regions found. Therefore, finding regions corresponding to terms with images human computation (von Ahn 2007) is a valid alternative to automated approaches. Especially games with a purpose (von Ahn 2006), which are fun, simple and entertaining games that hide the actual task and motivate users on a playful level to do things they otherwise wouldn't do, have a huge potential. Image annotation & games with a purpose are quite a prominent combination. Labeling images in a competitive multiplayer game has been one of the very first applications called games with a purpose, cp. *The ESP Game* (von Ahn and Dabbish 2004).

An effort very similar to our approach is *PeekaBoom* (von Ahn, Liu, and Blum 2006). In this game two players simultaneously play a cooperative 2-player game, where one player reveals regions of an image according to given terms, while the other guesses terms that apply to the uncovered region. Score points earned depend on how few pixels had to be revealed for the correct term to be guessed. Motivation for the users is to enter a global highscore. Another related effort is *LabelMe* (Russell et al. 2008), a web based annotation tool, where users can annotate images by drawing regions and assigning text to them. However, LabelMe is not a game, but a collaborative annotation tool.

Figure 1: Example photo with the original version on the left hand side and the region covering the semantic concept *squirrel* on the right hand side.

In this publication we present two games with a purpose, called *RecognizePicture* and *rpMobile*, trying to detect regions that cover the semantics of the given tags. These are not necessarily objects in the images, but also concepts like *car race*, *summer* or *beautiful*.

Architecture & Implementation

Given an image I and a label or tag t reflecting a concept and describing (a part of) the image, we are searching for a significantly smaller sub-image $I_t \subseteq I$ that reflects the concept of t "well enough", meaning that people can infer t from I_t as well as from I. In the human computation approach we assume that N people define sub-images, which are not necessarily the same. So for N different sub-images I_t^n with $n \in \{1, 2, ..., N\}$ we further define the minimum sub-image necessary to infer the concept t as $I_t^{min} = \cap_{n \in N} I_t^n$. We define the maximum sub-image necessary to infer the concept of t the same way as $I_t^{max} = \cup_{n \in N} I_t^n$.

The goal of the games presented is to motivate people to deliver as many I_t^n for as many pictures as possible. We further aim to keep I_t^{max} significantly smaller that I, while the difference between I_t^{max} and I_t^{min} should be as small as possible.

Game Design

Both games, *RecognizePicture* and *rpMobile*, have similar game mechanics. The games are played in several rounds, whereas each round features one image and four possible answers. One of the answers is the right one, all other three are considered wrong. Both games operate on the same data structure and data set, but results are analyzed separately. The game screens shows (i) the four answers, (ii) the current score, (iii) the game time remaining, and (iv) the animation revealing an image. Based on the revealed image parts, players guess which one of the four answers applies. The faster they guess right, the more points they get. Image revealing animations are chosen randomly and uncover images (i) from the center (cp. figure 2 A), (ii) from a corner (cp. figure 2 B), or (iii) from a side (cp. figure 2 C). As there are four corners and four different sides for the revealing animations to start, there are all in all nine animation styles possible. Each of them results in a rectangular area of pixels revealed at the time of user input.

Figure 2: Reveal animations for *RecognizePicture* and *rpMobile*. Animations start from the center (A), a corner (B), or a side(C).

RecognizePicture RecognizePicture is a competitive, 2-player game. It's a desktop application, written in C# based on Microsoft XNA Game Studio[1]. It's played on a shared screen by two players with Xbox360 gamepads. The answers given correspond to the controller buttons (cp. figure 3) to allow for a low entry barrier to the game. If one player hits a button giving the answer, the other controller starts to vibrate and the image revealing animation stops. Then the second player, who has not given her answer yet, has three seconds left to guess. For a correct answer one point is awarded, for a wrong answer there is a one point penalty. For skipping the answer no points are given or taken. Our test have shown that most people give an answer within these three seconds. Our hypothesis is that the vibration puts people under pressure. Also the social dynamics between the two players sharing one screen may add to the pressure on the player being late with the answer. A game is typically a sequence of five rounds. The game itself is meant as a fixed installation, so there is no highscore list as a global reward system, just the social interaction between the two players.

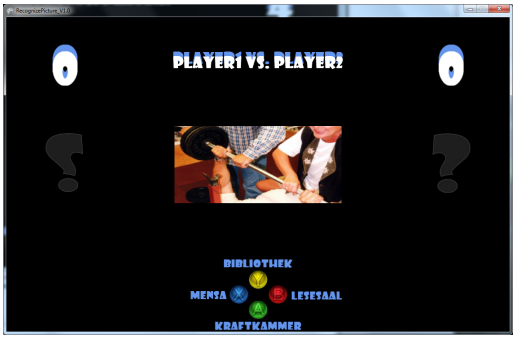

Figure 3: Screenshot of *RecognizePicture* showing a partially revealed image and four answers, which translate to *library*, *reading room*, *exercise room*, and *cafeteria*, clock wise from top.

rpMobile rpMobile on the other hand is a single player mobile game. It's currently under development using libGDX[2], a game library for Android and Java desktop applications. rpMobile uses the same revealing animations (cp. figure 2), but builds on a play-against-time paradigm and a score system with a global highscore. The ultimate goal for the player is to get as many right answers as possible in a limited time. So if you are very fast, you get more pictures to play with and can earn more score points. This of course leads to game dynamics different to RecognizePicture. We assume rpMobile can be played more often, as it is – in contrast to RecognizePicture – decentralized and a single player experience, which qualifies rpMobile for a *casual game* (being a game that can be played for fun for a short amount of time, i.e. to kill time).

Figure 4 shows a concept drawing of rpMobile. Instead of buttons, rpMobile features the four answers in the corners of the touchscreen. Besides the image, a progress bar indicating the remaining time on the right and the current score on the left are shown.

Backend

Both games use the same backend. The backup allows for (i) upload of new data, (ii) management of images and answers, and (iii) analysis of the user input. All administrative tasks are done using a web interface. Considering only the

[1] http://msdn.microsoft.com/en-us/centrum-xna.aspx

[2] http://libgdx.badlogicgames.com/

Figure 4: Concept drawing of *rpMobile* showing a partially revealed image and four answers.

correct answers, the backend then renders the results and visualizes I_t^{max} and I_t^{min}, the maximum and minimum sub-image needed to determine the correct answer (cp. figure 5).

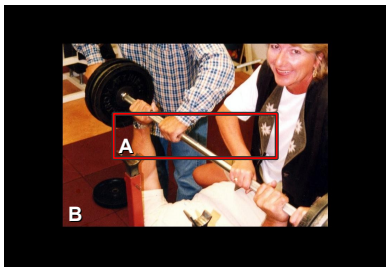

Figure 5: Result visualization of the backend. Rectangle A denotes I_t^{min}, and rectangle B denotes I_t^{max}.

Results

While rpMobile is under development, RecognizePicture has been deployed at an open house event of the Alpen-Adria Universität Klagenfurt, where people visit the university presenting a kind of "research fair". The game was installed in a small exhibition room, which represents the physical counterpart of a digital museum of the university called *AAU HiStories*[3]. This digital museum features more than 300 original images collected from alumni and retirees with the goal to preserve stories as addition to historical facts. 14 of those images were selected and each of the images was assigned four answers for the game. All in all we collected 486 sub-images I_n cumulative for all 14 images. 257 of those sub-images were submitted based on the right answer. Due to the small answer set we could not yet do statistically significant evaluations.

The evaluation strategy involves the employment of human judges, who annotate the images by selecting rectangles of interest for a given term. With multiple judges we can (i) compute inter-rater agreements to see how well experts agree on a region, and (ii) compare the regions from the expert judges to I_t^{max} and I_t^{min} resulting from a field deployment of the game, to see how well the "wisdom of the crowds" correlates to the judgment of experts.

[3]http://www.aau.at/histories

Conclusion

In this paper we presented two games with a purpose. One has already been deployed, the other is still under development, but both have the same goal: finding regions in images that correspond to given terms. User feedback was encouraging and we will do field tests and an evaluation of the results in near future with the deployment of the mobile client. While the game concept and the reception of the first game are promising, there are further obstacles to overcome. Image selection along with the selection of proper terms for the games proves to be difficult. Random selection strategies for wrong answers did not yield satisfying results as nearly right answers get mixed in and players tend to get confused about them. We also did not discuss cheating and the influence of guessing on the system due to the fact that we did not yet have a large user base.

Acknowledgments

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria, and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant 20214/22573/33955.

References

Avidan, S., and Shamir, A. 2007. Seam carving for content-aware image resizing. *ACM Trans. Graph.* 26(3):10.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(11):1254–1259.

Lampert, C.; Blaschko, M.; and Hofmann, T. 2008. Beyond sliding windows: Object localization by efficient subwindow search. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1 –8.

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H.-Y. 2011. Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33(2):353 –367.

Russell, B.; Torralba, A.; Murphy, K.; and Freeman, W. 2008. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77:157–173. 10.1007/s11263-007-0090-8.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326. New York, NY, USA: ACM.

von Ahn, L.; Liu, R.; and Blum, M. 2006. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, 55–64. New York, NY, USA: ACM.

von Ahn, L. 2006. Games with a purpose. *Computer* 39(6):92–94.

von Ahn, L. 2007. Human computation. In *K-CAP '07: Proceedings of the 4th international conference on Knowledge capture*, 5–6. New York, NY, USA: ACM.