Pushing the Envelope: Solving Hard Multimedia Problems with Crowdsourcing

Wei-Tsang Ooi
National University of Singapore
ooiwt@comp.nus.edu.sg

Oge Marques
Florida Atlantic University, USA
omarques@fau.edu

Vincent Charvillat Axel Carlier
University of Toulouse, France
vincent.charvillat@enseeiht.fr
axel.carlier@enseeiht.fr

1. Introduction

The solution to many contemporary problems in multimedia research involves discovering ways to bridge – or at least narrow – the *semantic gap* (the difference between the data that can be captured from raw pixels or sound samples and the high-level interpretation assigned by humans to the associated images, videos, or music clips). The difficulty in bridging the gap has led the multimedia research community to break the problem down into smaller sub-problems, such as image segmentation, image tagging, speech-to-text conversion, and natural language processing. Alas, in almost every field of multimedia research, the performance achieved by state-of-the-art algorithms is far inferior to humans performing comparable tasks. Moreover, most of the research focuses on specific domain or applications. Finding a general solution remains an open problem.

More recently, *crowdsourcing*, in which inputs from large numbers of human participants are pooled to serve as a basis for statistical analysis and inference, has arisen as a promising approach to address the sub-problems, in an effort to bridge the semantic gap. In this paper, we argue that – despite the success achieved by several of these early works – crowdsourcing is a much more powerful weapon and we should use it to directly solve the significantly harder *primary* problems, instead of its sub-problems. The rest of this article provides a few examples to support this argument, including two from our own research work.

2. Example 1: Identifying Interesting Regions in a Video

Our first example concerns the following primary problem: *Given a video clip, which region in the video would a user be interested to view at a given time?* This problem is extremely challenging if we attempt to tackle it using content analysis techniques. The notion of "interesting" region is not even computationally well defined.

To answer the question "what the user would like to look at?" one would have to at least model the content of the video, which is extremely hard. For the sake of exposition, let's consider the problem in the specific domain of lecture videos — typically consisting of a person lecturing in front of a lecture hall, with a blackboard or projected slide in the background. Understanding the content of such video can be broken down into several sub-problems: (i) extracting the text from the slides or the blackboard, (ii) separating the voice of the lecturer from the background noise, (iii) transcript the speech into text, (iv) perform natural language understanding to understand the context of the lecture, (v) separating the lecturer from the background in the video, (vi) identifying the gesture of the lecturer. Once the context of the lecture and the relationship between the speech and the text are known, one can then make a guess about the region of video that is of interest to the user. For instance, when the lecture says "According to the Fermat's Theorem" (from (ii), (iii) and (iv)) and pointing to the direction on the board (from (v) and (vi)), one can guess that at that moment in the video, the scribbled Fermat's Theorem on the board (from (i)) would be the region of interest to the users.

The computer vision and multimedia research community, to a varied degree of success, has extensively studied each of the sub-problems above. Some problems, such as transcribing a professor's handwriting on the board to text, remain a challenge. Even if each of the sub-problems above is satisfactorily solved, the solution to infer the interesting regions is restricted to a specific domain.

One could be tempted to tackle these sub-problems via crowdsourcing. While we acknowledge that solving some of these sub-problems would be useful in some applications (e.g., speech to text is useful for indexing and retrieval), we argue that we can skip the sub-problems, and use crowdsourcing to directly address the original question, "which region would the user like to view in the video."

We presented an approach to identify the interesting regions that user would like to view in a video through crowdsourcing [1][2]. Our idea is to utilize a novel interface for watching video, that supports zoom and pan operations. The interface allows user to explicitly zoom into regions they are interested in to view in more details. The zoom action provides explicit feedback to the system that a particular region is of

interest to a user. By aggregating the feedback from multiple users, we are able to determine a set of regions that user are likely to be interested in naturally.

3. Example 2: Describing an Image

The second example concerns the following primary problem: *describe a given image in text.* This problem is extremely challenging to solve, since it requires understanding the content and context of the image in order to generate the appropriate text.

The term "description" could be broad, and similar to the first example, is not computationally well defined. Here, we narrow down the problem into describing the objects in the image, their actions, and relationships to each other. One can break the problem down into several sub-problems: (i) segment the image into objects, (ii) identify *what* is each object in the image, (iii) identify the *role* of each object, and (iv) identify the *relationships* among the objects. It is reasonable to assume that additional meta-data from the camera – containing information about *when* and *where* the photo was taken – is available. Even with this assumption, being able to automatically generate a description such as "Obama hugs Michelle" (from the most-liked image of all time from Facebook at the time of writing) is extremely hard.

Researchers have had great success with image segmentation techniques (Step (i)) and are actively pursuing the sub-problem of annotating images (Step (ii)-(iv)), including many recent efforts that are based on crowdsourcing.

Again, we argue that one could attempt to solve the primary problem with crowdsourcing, instead of solving the sub-problems individually. Unlike our first example, where pooling information about what existing users look at naturally tells us what other users should look at, however, this second example is non-trivial to solve with crowdsourcing. A naïve way would be to ask the crowd to describe an image in natural language, or ask them to explicitly label the objects, actions, and relations in the image. Each of this approach has its own drawback: the former would require natural language understanding, which itself is a hard problem, while the latter requires incentives for users to perform the labor-intensive tasks of labeling.

We recently presented an approach towards answering the problem, focusing on identifying the relevant objects in an image and their spatial relationship through a game with a purpose (GWAP). The game, Ask'nSeek, is a two-player Web-based guessing game that asks users to guess the location of a hidden region within an image with the help of semantic and topological clues [3]. The information collected from game logs is combined with results from content analysis algorithms and used to feed a machine learning algorithm that outputs the outline of the most relevant regions within the image and their names (Figure 1). The approach solves two computer vision problems – object detection and labeling – in a single game and, as a bonus, allows learning of spatial relations (e.g., sky is above the man) within the image.

Figure 1. Examples of object detection and labeling results obtained with the game-based approach described in [3]: (left) four objects /regions were detected and their bounding boxes were labeled as 'woman', 'sky', 'motorbikes', and 'man'; (right) two objects ('cat' and 'dog') were detected and labeled.

4. Other Examples

Besides the two examples taken from our own research work, there are many other examples of primary multimedia problems that require bridging of the semantic gaps and are well known to be hard. We sample two such problems in this section that remains open but we believe could benefit from crowdsourcing.

Lyrics Transcription. Lyrics transcription involves transcribing the lyrics of a given songs to text, and is well recognized as a hard problem. Traditional approaches typically address three sub-problems: (i) separating the vocal from the instrument, (ii) segmenting the vocal into words, and (iii) transcribing each word.

Video-Song Matching. The problem of automatic soundtrack generation involves finding the most appropriate song from a collection to use as a soundtrack of a video clip (e.g., a home video), such that the song's content fits the scene of the video. One could view this problem as a combination of several sub-problems: (i) find the interesting regions of the video, (ii) transcript the scene, (iii) find the song with lyrics that best fit the description of the scene. Even if we assume the lyrics of the songs are available, the problem is still extremely difficult.

There is no known work that uses crowdsourcing to address the two problems above. We, however, believe that the technique is powerful enough to solve such

hard problems, with a well thought out system or GWAP. A system similar to reCAPTCHA could perhaps be useful for lyrics transcription. A song-guessing game may help with matching between video clips and songs.

5. Discussion

In the above, we argued that crowdsourcing is a powerful tool and we should push its envelope and use it to directly address hard multimedia problems that are challenging to solve using traditional content analysis approach, instead of using crowdsourcing on the sub-problems. Doing crowdsourcing correctly and effectively, however, is non-trivial.

Our experience with crowdsourcing leads to the following insights to designing useful crowdsourcing systems. The tasks or games should be designed such that *the input collected from the users is as simple as possible, but carries as much meaningful information as possible*. One key aspect in crowdsourcing is to cope with outlying data, and it is easier to filter out diverging contributions if the input is simple (for example, a click on a video to zoom in, a click on an image to play the game of Ask'nSeek). These inputs, however, should contain meaningful information, i.e., information hard to gather computationally, but easy to get by a human, to establishing relations between (different) media, semantics, and user interest.

We also would like to highlight that, despite the power that crowdsourcing yields, we should not ignore the traditional approach of content analysis. We argue that *content analysis should be used to augment crowdsourcing,* to reduce the number of participants needed to obtain meaningful results. To obtain meaningful relations from a limited number of user inputs in our work [2][3], we tried to plug the results from a limited number of contributions with content analysis outputs. In order for this combination to make sense, it is interesting to visualize the contributions from users as constraints that can help improve content analysis algorithms. For example, the relations obtained from Ask'nSeek can help filtering false positives from the classical approaches of object detection, as well as usage-based attention maps can complement saliency maps from the content analysis. In that sense we think that semi-supervising the content analysis by usage analysis could be a promising way of research.

6. Conclusion

In this paper we have postulated that it is time for the multimedia research community to take early successful attempts to apply crowdsourcing (including micro-tasks and games) to a new level, moving from bite-sized building blocks (e.g., image tagging or object detection) to grander, more encompassing primary problems, e.g., video-song matching, scene understanding and context-aware object recognition. We shared how crowdsourcing can be used to address two examples of such problems, drawing from our own research work. We also highlighted insights from our work on how to make crowdsourcing effective when addressing such hard problems.

References

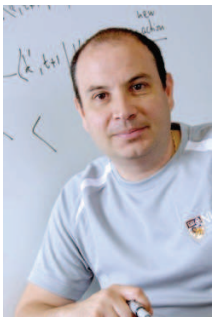
[1] Axel Carlier, Vincent Charvillat, Wei Tsang Ooi, Romulus Grigoras, and Geraldine Morin. 2010. Crowdsourced automatic zoom and scroll for video retargeting. In *Proceedings of the international conference on Multimedia* (MM '10). ACM, New York, NY, USA, 201-210

[2] Axel Carlier, Guntur Ravindra, Vincent Charvillat, and Wei Tsang Ooi. 2011. Combining content-based analysis and crowdsourcing to improve user interaction with zoomable video. In *Proceedings of the 19th ACM international conference on Multimedia* (MM '11). ACM, New York, NY, USA, 43-52.

[3] Axel Carlier, Oge Marques, Vincent Charvillat: Ask'nSeek: A New Game for Object Detection and Labeling. ECCV Workshop on Web-Scale Vision and Social Media, Florence, Italy, 249-258.

Wei Tsang Ooi received his B. Sc. (Hon.) degree from National University of Singapore in 1996, and Ph. D. in Computer Science from Cornell University in 2001. He spent a year as postdoc at Berkeley Multimedia Research Center in U.C. Berkeley, before re-joining NUS in 2002, where he is currently an Associate Professor in the Department of Computer Science. Wei Tsang's research focuses on interactive multimedia systems, including zoomable videos and networked graphics.

Oge Marques is Associate Professor of Engineering and Computer Science at Florida Atlantic University (FAU). He received his Ph.D. in Computer Engineering from FAU in 2001. His current research interests include the use of serious games and crowdsourcing to advance human and computer vision. He is a senior member of both the ACM and the IEEE.



Vincent CHARVILLAT received the Eng. degree in Computer Science and Applied Mathematics from ENSEEIHT, Toulouse France and the M.Sc. in Computer Science from the National Polytechnic Institute of Toulouse, both in 1994. He received the Ph.D. degree in Computer Science from the National Polytechnic Institute of Toulouse in 1997. He joined the Computer Science and Applied Mathematics department of ENSEEIHT in 1998 as an assistant professor. He obtained the habilitation degree in Computer Science in 2008 and is currently a full professor at the University of Toulouse, IRIT research lab, ENSEEIHT Eng. School. Vincent CHARVILLAT is the head of VORTEX research team at ENSEEIHT (Visual Objects: from Reality To EXpression). His main research interests are visual processing and multimedia applications.

Axel Carlier received his Master's Degree from INPT in 2011 and is currently a Ph. D. student, working under the supervision of Vincent Charvillat on the combination of content analysis with crowdsourcing.