Turning Wisdom and Effort of Crowds into Complex Media Annotation

ABSTRACT

Media annotation consists of supplementing media objects, such as videos, images, and audios, by adding metadata about their content and context, also to describing media characteristics such as quality, encoding, among other features. Complex media annotation involves annotating different aspects of media objects as well as relating them. This kind of annotation usually is associated with a demanding process that requires experts and elaborated annotation system. This paper presents a method to achieve complex media annotation without requiring complex tools, experts nor trained workers. In this method, the complex annotation process is divided into a set of simple annotation microtasks, and based on them is defined a process workflow for generating complex annotation. To demonstrate the operation of our approach, we developed a video enrichment system and carried out an experiment in which a crowd was responsible for executing a set of simple microtasks which are aggregated to produce enriched video content.

CCS CONCEPTS

• **Information systems → Multimedia information systems**; **Crowdsourcing**; • **Human-centered computing → Web-based interaction**; **Computer supported cooperative work**; • **Applied computing → Annotation**;

KEYWORDS

Crowdsourcing, Media Annotation, Video Annotation, Human Computation, Microtasks, Multimedia Systems, Video Enrichment

ACM Reference Format:

. 2017. Turning Wisdom and Effort of Crowds into Complex Media Annotation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Media annotation consists of supplementing media objects, such as videos, images, and audios, by adding metadata about their content and context, it is also used for describing media characteristics such as quality, encoding, among other features [25]. This supplementary information can be used to make easier the work of users and systems that can handle annotated items [19].

Annotations can be used to highlight key points and add information to contents presented [4], facilitating the creation of media applications for content-based distribution, indexing, summarization, navigation, composition and more, through automatic and manual means [15, 24].

Automatic approaches for media annotation, such as rule-based and deep learning, usually present satisfactory results in generating media annotation, although they require well-structured media objects and extensive examples database [14]. Thus, spontaneous scenarios involving unplanned and non-standard videos, images, and audios may not provide the requirements to apply these automatic techniques for media annotation [18].

Manual media annotation is suitable for these scenarios because it uses human intelligence to handle the tasks. However, this approach can be high-costly because of the potentially high-density of annotation points in the time-based media, as well as the complex nature of some annotation tasks.

Considering the amount of information, the number of interactions, and the expertise needed to generate an annotation, this annotation is classified in this paper as simple and complex. While simple annotations can be acquired with a simple interaction of annotators in a microtask, a complex annotation requires them to perform a more tedious, difficult, or time-consuming task in which he needs to perform multiple interactions.

Distributed approaches, whether cooperative or collaborative, are an alternative to bypass the high effort required for individual manual annotation. In a collaborative approach the annotators work together to solve the main problem. In a cooperative approach each contributor solves a part of the main problem to produce a final result in a divide-and-conquer strategy [16].

Cooperative approaches are efficient for media annotation because they allow the distribution of items to be annotated among the annotators. In this context, crowdsourcing emerges as an interesting strategy for cooperative processes, because it allows the execution of a large-scale cooperative process for media annotation, using a large number of contributors efficiently. [23]. Following the crowdsourcing principles, the tasks distributed to the crowd members are modeled to be done independently, maximizing the parallelism [10]. Moreover, each task can be sent to many crowd members, allowing to compare, check and aggregate the contributions, reducing the chance of producing a biased result [6].

There are frequent problems in using a crowdsourcing approach for media annotation, such as balancing the relationship between the complexity of the task and the cost. This cost refers to the profile and qualification required from the crowd members as to the complexity and difficulty of the task delegated to them.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Complex annotation usually requires more complex tasks, demanding some expertise from annotators and are harder and time-consuming to achieve. In opposite, simple annotation tasks can be performed easily and quickly by less skilled people [5]. Therefore, this paper aims to use untrained and unskilled people, to execute simple annotation tasks that, when tied together, are capable to generate a complex outcome.

Hence, the proposed method presents some characteristics to get around some problems faced in achieving a complex media annotation, including:

- Manual annotation is used to dispense examples bases and restricted conditions.
- The annotations are provided by ordinary contributors rather than experts or trained ones.
- It is based on simple and quick microtasks rather than difficult, time-consuming and demotivating tasks.
- Simple annotation tools are used instead of complex systems.

The objective of this work is to present and validate a method for video enrichment using annotation provided by the crowd. The proposed method share some common features with the proposal of [19], however some important improvements have been implemented in this enhanced version:

- The method can be now applied to any media object and not only to videos.
- It supports the use of complex workflows, that specify the progression of steps (tasks, events, interactions) that comprise a complex annotation process.
- A microtask can generate multiple outputs.
- The aggregation of microtask outputs can be automatic, manual and supervised, instead of just automatic ones.

An experiment involving a crowd composed by unskilled people was carried to evaluate the proposed approach for video annotation. The result of this experiment was evaluated by three criteria:

(1) Comparing the points of interest identified by the crowd with those highlighted by experts.
(2) Identifying whether the crowd is able to enhance a multimedia content by providing suitable extra content.
(3) Verifying if the position determined for the extra content on the video causes occlusion of important scene objects.

This paper is organized as follows. Section 2 presents the concepts used and how they are employed in the proposed approach. Section 3 presents related works. Section 4 presents the method. Section 5 presents the conducted experiment. Finally, section 6 concludes the paper presenting final considerations and future prospects.

2 RELATED CONCEPTS

The presented method is based on the human computation paradigm [23], using a crowdsourcing approach [10] to annotate media objects. To do so, this method employs some fundamental concepts commonly associated with human computation. These concepts are presented in this section.

2.1 Human Computation

There are jobs that are trivial to all people but they are extremely difficult even for the most powerful and sophisticated software and machines. This type of task has as its characteristics the need for creativity, emotional sensibility, empathy, sarcasm and level of abstraction which, in the present moment, is inherited in the minds of human beings.

There are still cases where automatic methods can accomplish the task, but the necessary resources do not exist as training datasets or a well-defined set of rules. Some examples of this type of task are the image recognition ones, especially when there is occlusion or involves subjective analysis, content authorship, analysis of emotions and so many other activities that inherently require human intelligence to be performed. Luis von Ahn introduced in his dissertation [23] a paradigm named Human Computation that allows approaching the problems from this point of view, identifying in it what tasks can be automated and which ones require human treatment. Additionally, Human Computation can improve performance by the division of labor because it helps to define tasks that can be executed in parallel [21].

One of the benefits of modeling a system according to the Human Computation paradigm is to focus the effort of human collaborators only on tasks that really require their attention, this is done by identifying the tasks that inherently require human intelligence. These tasks are called HITs (Human Intelligence Tasks) and correspond to something that humans can easily solve while a machine presents extreme difficulty in trying to solve [13].

A HIT may involve creative activities such as authorship, simple observation as in identifying events in videos, or empathy such as the detection of emotion in facial expression images. In terms of complexity, a HIT can range from a simple microtask to a more complex macrotask.

Macrotasks require more effort from workers, often requiring them to be experts in the field or to have training on subjects related to the task. In the context of media annotation, this kind of task is suitable for complex annotations because it assumes that the worker is qualified and will devote the time and effort required to complete it. Also, macrotasks often require sophisticated annotation systems and limit the group of workers who are able to execute them [8].

On the other hand, microtasks are usually modeled in a way that can be accomplished quickly and easily by less skilled workers. For media annotation, this kind of task is usually used to generate simple annotations, which refer to one or a few items to be annotated in each task. Often a Microtask can be performed using a simple annotation tool. Microtasks are widely used in crowdsourcing projects and this kind of task is supported by well-established commercial platforms such as Amazon Mechanical Turk[1], CrowdFlower[2], and Microworkers[3]. Hence, a microtasks should be:

- **Small:** a worker must complete a task by means of few interactions.
- **Quick:** it should be possible to complete a task in a very short time.
- **Easy:** the easier the task, the less skilled the workers can be.

[1] https://www.mturk.com
[2] https://www.crowdflower.com
[3] https://ttv.microworkers.com

2.2 Crowdsourcing

To support the human computation paradigm, crowdsourcing has emerged as a proposal to annotate media objects using a large number of contributors efficiently [23].

This approach generally delivers good quality results using contributions from the crowd and can distribute, collect, validate and combine large amounts of tasks results [17]. Since this approach is designed to handle a huge number of cooperators and contributions for tasks that require human intelligence [10], Crowdsourcing is appropriate to allow the Human Computing paradigm to be applied in a massive-scale online cooperation [1]. Crowdsourcing is supported by four pillars: The Crowdsourced Task, The Crowdsourcer, The Crowd, and The Crowdsourcing Plataform [9].

- **The Crowdsourced Task** is the HIT designed, according to the Human Computing paradigm, to acquire workers' contributions. Task instances are presented to the workers as jobs that must be performed [5].
- **The Crowdsourcer** is the owner of a project, it may be an individual or institution that wishes to have a completed task. The owner is responsible for starting the crowdsourcing process, defining what task must be completed and how it should be presented to the workers as jobs [9].
- **The Crowd** is the workforce that moves the process once it is composed of all the workers who perform the jobs needed to generate the outcome. Each worker carries out his work independently so that the works can be executed in parallel, according to the paradigm of Human Computation [21].
- **The Crowdsourcing Platform** is a computational system responsible to manage the whole process, serves as an entry point for both the owner making the tasks available and for the workers to execute them. This kind of environment can be something sophisticated such as CrowdFlower, Microworkers, and AmazonMechanical Turk [5], or really simple systems with screens and forms for data collection such as the mobile application used by the Google Crowdsource project [11]. A crowdsourcing environment is necessary, as the tasks must be made available to a potentially large number of workers. The crowdsourcing platform is a key element of support to massive-scale cooperation.

The use of a commercial crowdsourcing platform brings benefits such as not having to worry about management's issues of workers, jobs, and contributions such as employee recruitment and collection of contributions as well as facilitating employee payments. Also, payouts are a good way to motivate and keep the crowd, although there are other coping factors such as personal accomplishment and gamification.

2.3 The Wisdom of Crowds

Francis Galton wrote in 1907 an article in which he reported an experiment that a crowd of people at an agricultural fair tried to guess the weight of a particular ox. Galton verified that the average of the assumed weights converged to a value very close to the actual weight of the ox and analyzing the distribution of values grounded the Wisdom of the Crowd concept, according to which a heterogeneous crowd large enough tends to provide such a good expert result [6]. Additionally, James Surowiecki listed in his book "The Wisdom of the Crowds" [22] four requirements for a crowd to deliver good results.

- **Diversity:** each person adds private information or bias.
- **Independence:** people form their opinions independently.
- **Decentralization:** people draw on their own knowledge.
- **Aggregation:** a mechanism exists to turn private judgments into a collective decision.

Crowdsourcing platforms are environments suited to this theory, especially because it is possible to reach workers from different parts of the world and contexts with different backgrounds, which favors diversity. A common scenario in these environments involves workers receiving tasks and executing them without interacting with others, which helps create conditions for Independence and Decentralization.

For each task, an appropriate aggregation method should be applied. These methods may involve the convergence of responses, geometric means, contribution merging, as well as other types of processing. In this work, three categories of aggregation methods are also considered: automatic, supervised and manual.

- **Automatic:** an automatic aggregation method employs algorithms to process the contributions in order to generate the desired outcome. This kind of method involves convergence analysis, geometric mean, instance prevalence, and ranking.
- **Supervised:** in some cases its possible to apply automatic convergence methods, although it is required additional human verification or some human interaction in the contributions' processing.
- **Manual:** manual aggregation is used when it's not possible to define the rules or algorithm that should be followed to generate the outcome. These situations are usually related to subjectivity and emotional aspects that must be analyzed in the contributions to generate the result.

3 RELATED WORK

Crowdsourcing media annotation approaches are used in various applications and are used to gather information of various types, such as temporal synchronization[26], events[12], scene objects[3], emotions[2], actions[20], and geo-tagging[7]. Among these works, [12, 20, 26] used microtasking approaches and unskilled workers, while [2, 3, 7] required skilled or trained workers to perform more complex tasks.

The work conducted by Kim [12], he used a crowd to jointly summarizing large sets of Flickr[4] images and YouTube[5] videos in order to create novel structural summaries of online images as storyline graphs. The workers received a set of frames extracted from a video segment, and had to select some of them to create a summary. Once the set of convergent images were obtained, they were used to find similar images on a dataset extracted from Flickr.

Riek used a game approach [20] to annotate a video dataset with tags related to facial expressions, posture, gestures and more. The crowd used a very simple annotation tool in which they could insert tags by clicking buttons.

[4] https://www.flickr.com
[5] https://youtube.com

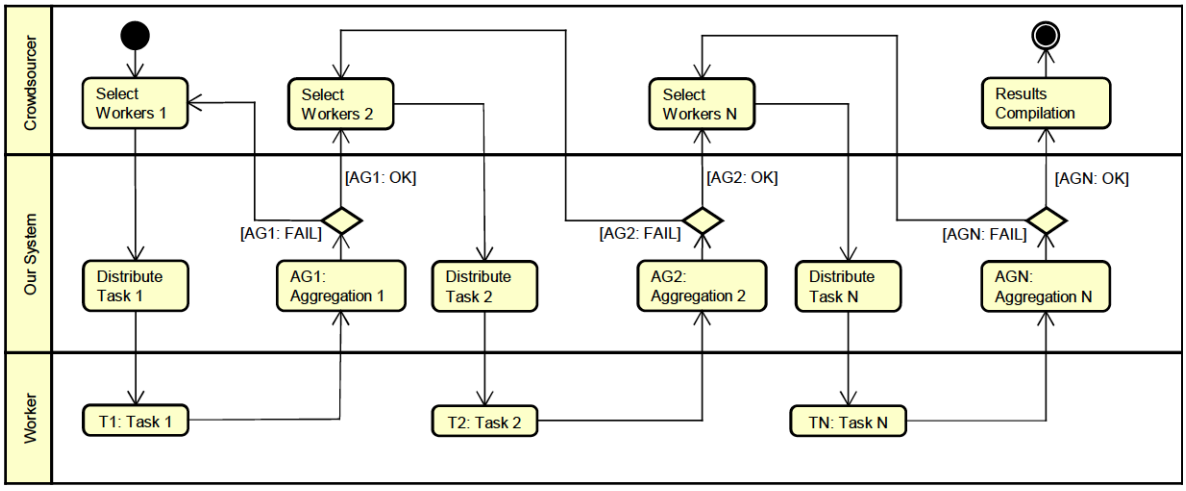

Figure 1: Process workflow. T (Task); AG (Aggregation).

VidWiki[3] is a complex system to improve video lesson by annotating them, which provides a complex annotation tool that allows the worker to edit video scenes by adding various types of annotations, including LaTex equations. This system requires some skills, including knowledge about Latex[6].

An important observation of the works related to crowdsourcing media annotation is that, in general, microtasks are performed by unskilled workers using simple annotation tools, to obtain simple annotations. On the other hand, jobs that aim at complex annotation often use larger tasks that require skilled or trained workers as well as more elaborate tools.

4 THE PROPOSED METHOD

The method presented in this article aims to achieve a complex annotation of media objects such as audios, videos, images and texts. The differential of this method is that it allows us to reach these annotations using accessible resources such as simple annotation tools and the work of unskilled and untrained workers.

For this, a crowdsourcing approach was adopted, following a production model based on the cascade of microtasks, as can be seen in Figure 1. In this way, the final result is built as the tasks are performed, with a task complementing and refining the result of the previous one.

This method follows three phases: Planning, Production, and Delivery. In the Planning phase, the process workflow is defined, in the Production phase the contributions are collected and processed and, in the Delivery phase, the outcome is available for consumption.

4.1 Planning Phase

All activities involved in this step are performed by the crowdsourcer, who started the media annotation process. The objective of this phase is to determine what should be produced and how the production process should be.

Once the desired complex media annotation has been defined, it is analyzed and all information required for its construction is identified. Then a plan is defined for the construction of the outcome, determining the order in which the partial results must be produced, refined and improved. Each partial result corresponds to the output from the aggregation activity related to one of the tasks.

Therefore, it is necessary to design the necessary microtask to obtain each partial result and the order of dependency between them. For each of these tasks, it is necessary to define:

- What kinds of point of interest should be annotated.
  e.g. events, objects, subjects, issues.
- What annotation type will be used for each of these kinds.
  e.g. free write, item selection, button click, image upload.
- What data type will be collected for each annotation type.
  e.g. plain text, location, image, video.

Once the microtasks are defined, a workflow can be constructed to determine how they fit into the complex annotation production process. The built workflow should follow the general model shown in Figure 1. This workflow will guide the production phase.

Another important point to be defined in the planning phase is how the jobs should be distributed among the workers in each task. This concerns the distribution method and the stopping criteria. For example, in a given project, can be prioritized the annotation of items that have not yet received contributions, and set a maximum number of contributions per item as stop criteria.

Also at this stage should be defined the item that should be sent to the worker for each job. In this way, it is necessary to define the segmentation or selection strategy of the media objects to be annotated. In the case of image data sets, it is necessary to determine the set to be sent in each job, and if it is audio and video, determine the segmentation strategy. Segmentation can be done, for example, by duration (ex: send a 5 seconds segment to each worker), or using contextual criteria such as to send to each user a segment that contains a single dialog.

Finally, the annotation tools that will be used to collect the contributions in each of the microtasks are chosen.

[6]https://www.latex-project.org

4.2 Production Phase

In the production phase, the workflow (Figure 1) defined in the planning phase must be executed. This workflow illustrates a cascade of sub-processes where the output of one of them is used as input to the next. Each of these sub-processes, which is represents to a microtask, is composed of four activities: (i) select workers, (ii) distribute task, (iii) task, and (iv) aggregation.

The proposed method recommends using a own system to perform the tasks of this phase, using from the crowdsourcing platforms only the workforce of your crowd. However, you can follow the method using features provided by crowdsourcing commercial environments such as Microworkers, CrowdFlower, and Amazon Mechanical Turk.

The system developed for the experiment carried out offers all the resources to carry out the production phase. This system can be obtained, used and extended freely [7]. It support the following activities of the production phase:

(1) **Select Workers:** recruitment and selection of workers can be done in different ways. When using a crowdsourcing platform, this activity is outsourced to it simply by defining the desired profile and how much will be paid for work. In this method, this activity can also be done by an open call, or by using a closed group of selected workers. For validation experiments, it is still possible to use a group of experts.
(2) **Distribute Task:** the distribution of tasks can be done by both the crowdsourcing platform and a own system, even when using an commercial crowdsourcing platform. To ensure better control of the distribution process as well as the stop criteria, the presented method recommends that a own system be used to distribute tasks even when you use the crowdsourcing platform. This system is responsible for determining the information about the next job to be delegated to a worker, such as the item to be noted.
(3) **Performe the Task:** the execution of the microtask consists of to present for the worker the job received by the distribution process and collecting its contribution. Each task must be presented to the worker through the correct annotation tool, giving him the necessary instructions so he can perform the job.
(4) **Realize the Aggregation:** the aggregation is responsible for verify, filter, group, and process the collected annotations of the crowd according to the rules defined for each task. In this approach, the aggregation process can be fully automatic, supervised or manual. Manual aggregation is useful when is desired to evaluate each contribution as in authoring tasks. Supervise aggregation can be a good option when is possible to apply automatic methods but is required human verification of the result. Automatic aggregation is the default choice, this class of methods includes grouping, comparing, counting, calculating and other operations. The aggregation activity can generate different outputs, which can feed the next task and generate other artifacts. For example, the output of a microtask in which workers annotate events occurring in a video may, in addition to fueling a next task, generate indexes and summaries for this video.

[7] https://github.com/[REMOVED FOR BLIND REVIEW]

4.3 Delivery Phase

The result generated by the aggregation activity of the last performed task in the workflow is the outcome of the production process. The delivery phase is responsible for make this outcome available to consumers, displayed by a player or exported in different formats because the complex annotation generated is stored as metadata and offers flexibility in this regard.

5 CASE STUDY

This experiment used the presented method to annotate and enrich videos, adding semantic information about personalities, theories, and technologies in an explanatory video about the history of the computer, produced for this experiment.

In order to operate this experiment, a crowdsourcing system was built capable of executing the cascade microtasks workflow according to the presented method. This system can be used for free and serves as a basis for other projects that wish to use this method to obtain complex media annotations.

This system consists of a simple annotation tool to execute each microtask, as well as the aggregation methods applied at the end of each sub-process. Also present in the system is a player capable of reproducing the final outcome of the process and the modules responsible for the distribution, management, and storage of contributions.

The crowdsourcing process used in the experiment followed a workflow of four annotation microtasks. Each of these simple tasks was modeled that could be performed by a worker through a simple annotation tool.

5.1 Our Crowd

To reach a heterogeneous crowd we used the Microworkers platform to recruit and pay workers, and the whole process was supported by our system.

Microworkers proposes different models for a crowdsourcing process. Initially, you can choose between starting a basic campaign or using contracted groups. In a basic campaign, all registered workers in the system can see the task on the job wall and accept to run it. A campaign that uses hired groups allows you to select the crowd by choosing groups of workers with a certain profile. In addition, it is possible to create lists with good workers who have made good contributions to recruit them for other tasks.

The purpose of this paper is to use unskilled workers, although it was necessary for them to know enough of the English language to understand the video used in the experiment. For that reason, the tasks were launched as campaigns that used contracted groups, to increase the chance of the workers who contributed in a task also participate in others, was chosen a group of moderate size and with workers relatively assiduous so that the contributions were made quickly. The group chosen was Data Services, with 1153 potential workers to accept the jobs.

Some groups are made up of workers who only accept tasks that offer slightly larger payments, but considering the group chosen, it was feasible to offer a payment of 0,02 USD per task. Also, each task was active for 24 hours to reach all time zones in the same way.

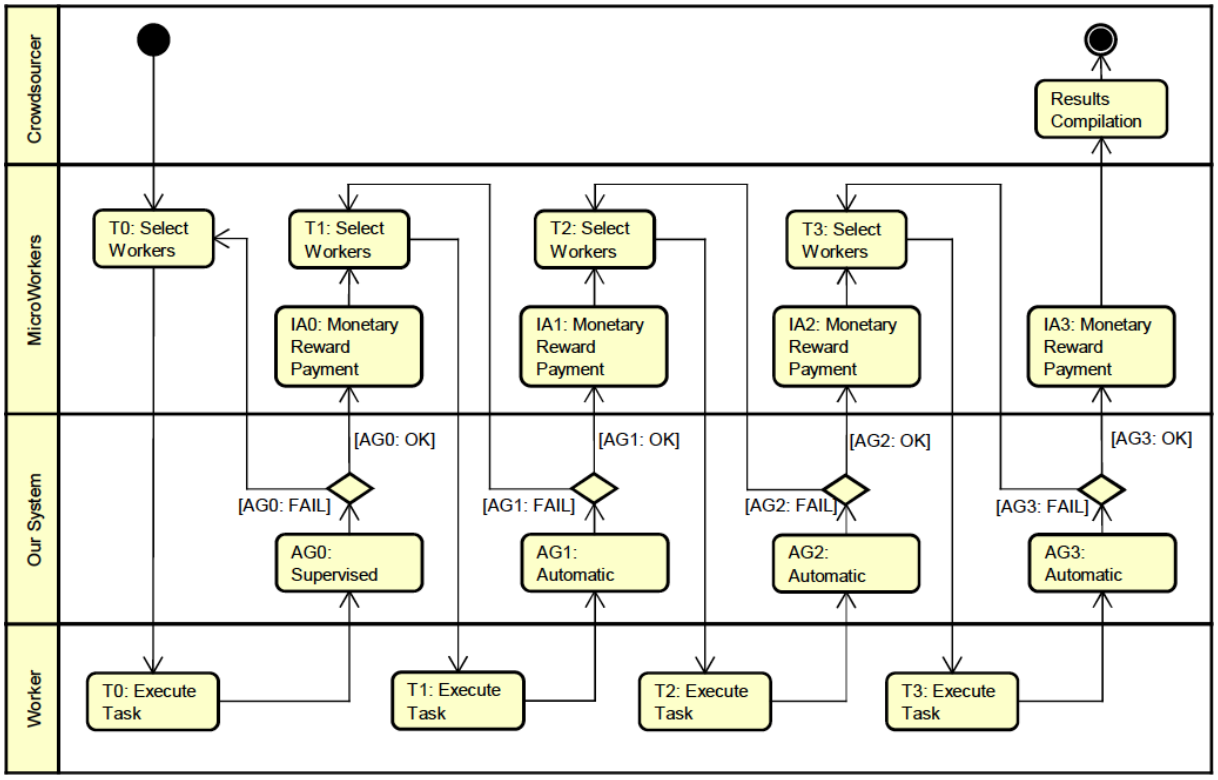

Figure 2: Video Enrichment Workflow. T (Task); AG (Aggregation); IA (Incentive Application).

5.2 Planning

The planning step has started with the identification of the micro-tasks needed to achieve the expected result. Since the goal was to generate enriched videos through supplementary content at points of interest, it was determined that they were needed. Thus, it was determined that four microtasks would be performed by the crowd:

- Identify the points of interest in the video;
- Gather content suggestions for each point of interest;
- Select the best extra content to each point of interest;
- Position the trigger items at the best spot over the video.

Related to the items distributed to be annotated in each task, a circular list policy was adopted. In this way, there was a greater chance of each task, each item being annotated received the same number of contributions.

It was also necessary to determine how the video should be segmented to generate the input of the first task. The strategy chosen was to target the video based on the automatic captions generated by Youtube[8]. In this way, it was possible to generate short segments, but they tend to contain complete sentences. This method divided the video into 13 segments.

For budget and time issues, it was determined that the minimum amount of contributions each item could receive was five, so *Task 0* should receive at least 75 contributions. The number five was chosen because it is odd, which avoids drawings, and because it is an amount that already allows seeing a tendency of convergence of opinion in a heterogeneous multitude.

5.3 Production

In this section will be described the 4 annotation tasks, as well the annotation tools, aggregation methods and results for each of them. Because there is a dependency order between these tasks, it's not possible run them in parallel. In this way, an execution workflow in which the output of one task is used for the next one, as can be seen in Figure 2.

5.3.1 *Task 0.*

- **Title:** Identify Points of Interest.
- **Description:** A video segment is displayed to the worker and he must identify the moment at which a point of interest appears or is mentioned. This point of interest can be a person, a technology or a theory.
- **Objective:** Identify points of interest in a video segment.
- **Input:** A dataset with 13 video segments of 4.5 to 7 sec.
- **Output:** A set of points of interest and the instant when they happen. In addition to serving as input to Task 1, this result also generated a summary based on points of interest.
- **Instructions:**

(1) Identify in the video something that you found interesting.
(2) Pause the video the moment it appears.
(3) Select its type: Person, Technology or Theory.
(4) Write what you have identified.

[8] https://youtube.com

- **Annotation Tool:** This tool (Figure 3) receives information about the video segment to be displayed to the worker. The video is hosted on YouTube for the player to be created using the official API[9]. A timeline control has been implemented that ensures that the worker can use the features of the YouTube player, but only the interval related to the segment to be annotated is reproduced. The worker also has fine tuning buttons to position the video at the exact moment you identify the point of interest, a list with the category options (Person, Theory, and Technology) and the buttons to view the instructions and send the contribution.

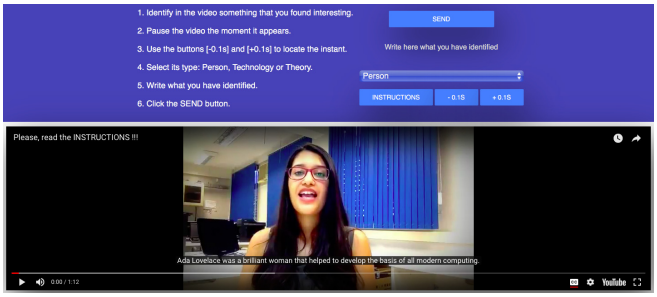

Figure 3: Annotation Tool for Task 0

- **Aggregation:** For this task, a supervised aggregation method was chosen in which the human interaction occurs at the end of the process. The contributions are filtered discarding the useless contributions (empty, nonsense, bad works), and the valid contributions are candidates to Points of Interest. The candidates are grouped in two-seconds interval ranges. In each group, a similarity analysis is done and the most frequent occurrence is selected. Finally, the human supervisor edits the selected label so that the text is visually pleasing. The aggregation method extracted 11 Points of Interest from the 75 contributions.

5.3.2 *Task 1.*

- **Title:** Provide Suggestions for Extra Content.
- **Description:** In this task, the worker receives a point of interest and the video synchronized at the moment it occurs. The worker should suggest extra content to complement the video, which may be a short text, an image or a link to a YouTube video or a Wikipedia page.
- **Objective:** Obtain from the crowd extra content to each point of interest.
- **Input:** The set of points of interest collected in *Task 0*.
- **Output:** A set of extra content associated with each point of interest. Each extra content may be an explanatory text, an image, even a link to a Wikipedia page or a Youtube video. Besides feed the *Task 2*, this output is also used to generate a content-based index to the video.
- **Instructions:**

(1) Select the type of content to send.
(2) You can upload an image, write a short text.
(3) You also can paste a link to Youtube or Wikipedia.

[9]https://developers.google.com/youtube/v3

- **Annotation Tool:** This tool (Figure 4) receives information about the video segment, a point of interest present and the instant the point of interest occurs. The worker can use the player to play the video segment and understand the context of the point of interest, which is displayed in a text area at the top of the tool. He can select an option related to the kind of extra content they will provide, and according to this is displayed in a text field to paste a link or type a text, or a button is displayed to send an image. There is also a button to view the job instructions and a button to send the contribution.

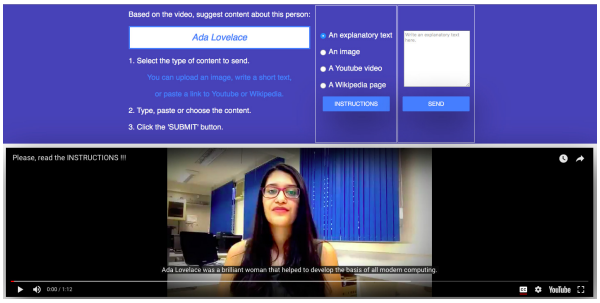

Figure 4: Annotation Tool for Task 1

- **Aggregation:** For this task was chosen an automatic aggregation method. The filtering step discarded the corrupted files and broken links. The suggested contents were grouped by point of interest, and the duplicated items were merged. The aggregation method extracted 34 different suggestions of extra content from the 55 contributions received. All 11 points of interest received at least two content suggestions, some of which received four or five.

5.3.3 *Task 2.*

- **Title:** Ranking Suggestions.
- **Description:** The worker receives a point of interest and the video positioned at the moment it occurs. It also gets the list of suggested extra content to complement this point of interest. The job is to choose the extra content that best complements the content related to the point of interest.
- **Objective:** Determine the best extra content to each point of interest, according to the crowd.
- **Input:** A set of points of interest, with the suggested contents associated with each one.
- **Output:** An updated set of points of interest with the extra contented associated with each one. This output also can be used to generate a ranked list of alternative extra contents to supplement the points of interest.
- **Instructions:**

(1) Use the buttons bar to navigate through the contents.
(2) You can use the zoom button to visualize each content.
(3) Vote for the content you liked best.

- **Annotation Tool:** This tool (Figure 5) receives as input the information relating to the video segment, the point of interest and a set of suggestions for extra content for it. The worker can browse content suggestions through the navigation buttons and can better visualize the content suggestions by clicking the zoom button. He can also play the video segment to understand the context of the point of interest. There is also the button to see the instructions and also the button to vote for the content you have chosen.

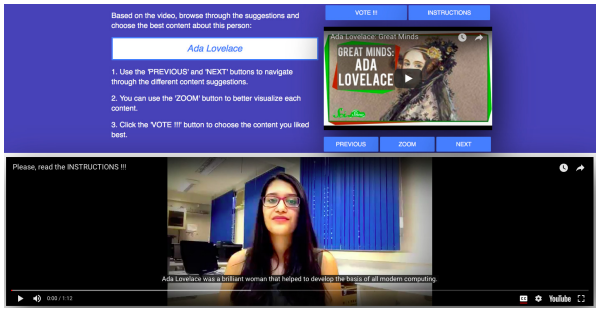

Figure 5: Annotation Tool for Task 2

- **Aggregation:** This simple automatic aggregation method determines the most popular extra content suggested to each point of interest. As the contents were associated with 11 points of interest, were collected 55 contributions. In the filtering step the suggestions without votes was discarded. Finally, was generated an output with all 11 points of interest and the extra content associated with each one.

5.3.4 *Task 3.*

- **Title:** Determine the Position of the Trigger Items.
- **Description:** This task consists of placing a trigger item in a video scene. This trigger item is represented as a rectangle containing a text or an image and should be positioned so as to minimize occlusions of important scene objects.
- **Objective:** Determine the best position to display each trigger item over the video.
- **Input:** A set of points of interest with the extra content associated with each one of them.
- **Output:** Aset of points of interest with the extra content associated with each one, including now the position where each item should be displayed over the video. These positions are represented as coordinates (X,Y). The output of this task is the outcome of the video enrichment process, and can be executed in the Player. In addition, the metadata can be used to generate alternative outputs such as NCL to reproduce them in digital TV environments.
- **Instructions:**

(1) Drag the item by the video until finding the best position.
(2) When you have decided on the best position, click send.

- **Annotation Tool:** This tool (Figure 6) receives the information related to a point of interest and the video segment in which it occurs. This tool has a drag-and-drop feature that lets you drag the item through the video until you find the appropriate position. Using this feature, the worker can move the item over the video to find the position that looks best, avoiding occlusions of important scene objects.

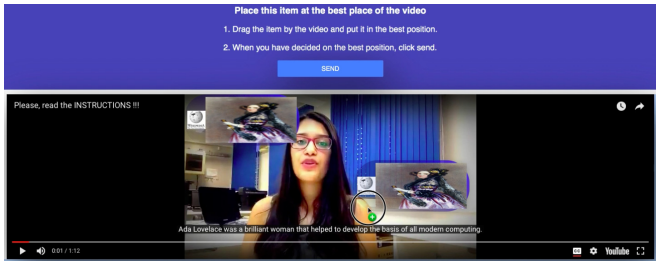

Figure 6: Annotation Tool for Task 3

- **Aggregation:** The automatic aggregation method applied at the final of this task determines the position at each item should be displayed over the video. The contributions for each item were grouped and the very discrepant positions of the others were discarded. Then, for each item, the average geographic position was calculated based on the X and Y coordinates predicted in the contributions. Was collected 55 contributions, of which 38 were after filtering. Of these 38 contributions were extracted the positions of the 11 items related to the points of intersection.

5.4 Delivery

Since this experiment aimed to generate enriched videos, the delivery of the outcome was done through a presentation system. This presentation system, shown in Figure 7, receives the video, extra content, and necessary metadata from the Player Provider. This system is capable of reproducing the original video synchronized with the extra content, that is displayed every time a point of interest happens in the video. Is important to remind that all extra content displayed with the video was provided, selected and positioned by the crowd.

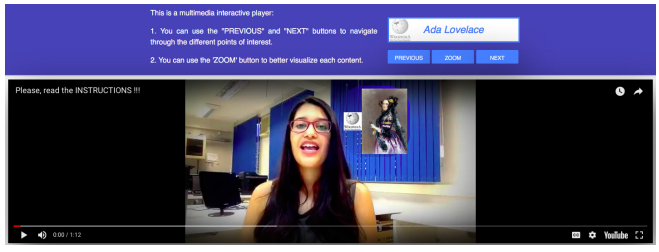

Figure 7: Displaying an extra content item over the video

This tool has a control bar with 3 buttons: Previous, Next and Zoom. These buttons are used to control two useful features, the content zoom and the navigation by point of interest.

The navigation by point of interest uses the list of points as a content index, allowing the user to navigate between points of interest by clicking the Previous and Next buttons. The Player automatically syncs the video at the time the current point of interest occurs.

When the user clicks the trigger item or the Zoom button, the extra content associated with the current point of interest is displayed on an upper layer as the video is paused. When you close the extra content, the video resumes its execution from where it was. The zoom view can be seen in Figure 8.

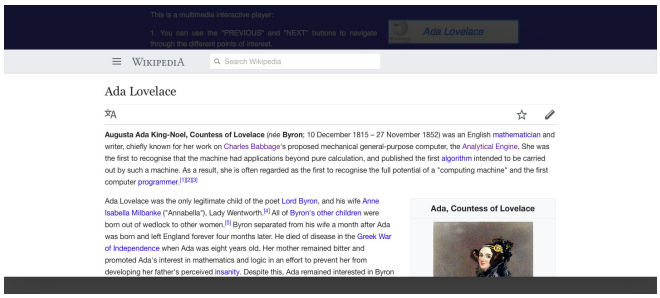

Figure 8: Player Zoom

5.5 Runtime Observations

During some tasks, we observed some interesting facts about the behavior of workers. Keeping in mind that the motivation of the workers in the experiment was the payment and that the amount paid for each job was small, the workers tended to do the bare minimum. This aspect proved to be correct that the decision was made to use microtasks that needed only a simple interaction to be completed.

The reflection of this led to some decisions during the experiment, as in Task 1. In this task, the worker were asked to suggest extra content to supplement the points of interest and for this, he could write a short text, paste a link to Wikipedia or Youtube, or upload an image. Most workers provided links and only two uploaded images. We thought that the reason was because sending an image was more laborious than pasting a link into the text box, as it was necessary to download the image and upload it.

However, this occurrence was used positively, because few image was obtained in the contributions, so, decided to add an automatic mechanisms to retrieve them automatically during aggregation. In cases where the selected point of interest was a Wikipedia link mechanism, it retrieved the main page image and, in cases where it was linked to YouTube, the thumbnail was retrieved from the video.

Although this operation is simple, it shows that it is possible to improve the aggregation methods by associating more automatic processing with human contributions. In this way, aggregation methods may be possible integration points with other systems, even with secondary human tasks to create a formal model of supervised aggregation activity.

5.6 Result Evaluation

The rich video generated is a self-contained multimedia document, so the user doesn't need to access supplementary content from other sources.

Each supplemental content can be accessed through trigger items that are images and labels that appear in scenes where the point of interest is mentioned. When the trigger items are clicked, the video pauses the extra content is displayed so they can be viewed, and when closed, video playback resumes normally.

This outcome was evaluated according three criteria: points of interest identified, suitability of the associated content, and occlusion of scene objects.

According to the author of the video used in the experiment, there were 21 different points of interest, which after aggregation would result in 14, because at each interval of two seconds, only the most relevant point of interest would be selected. The crowd managed to identify 19 points of interest, which after the aggregates generated 11 relevant points. However, the crowd highlighted the term "Formalism" that was not predicted by the author, who after checking it reported that it was valid in the video's context.

The extra content selected for each point of interest was also verified by the author. Of these, only one content did not meet the author's expectations. The content selected for "Logical relations and properties" corresponded to what the author wanted to convey. However, he realized that there was a flaw in the video script and the correct text would be "logical relational and its properties". This way this error can not be attributed to the crowd.

In relation to the occlusion of the objects of the scene, the criterion to calculate items that were positioned in the actress who narrated the text. Of the 11 items, only 2 slightly obscured the outline of the actress, as shown in Figure 9, but no item occluded her image significantly.

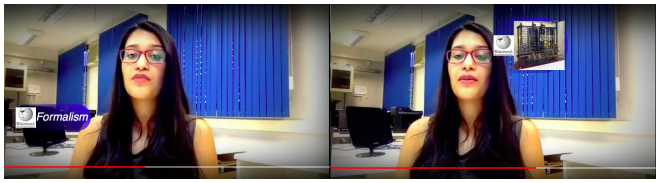

Figure 9: Occlusion Issues

In general, the result generated by the crowd was well adapted to convey the content intended by the author, having enriched virtually all the points that he had predicted. Finally, the result of our experiment demonstrate that the crowd is capable of generating dependable additional content to improve videos. The result of our case study can be seen in the player of our system[10].

[10] http://159.203.171.150:83

6 CONCLUSION

This paper presented a method for achieving a complex media annotation by running a process workflow composed of simple annotation microtasks in a cascading arrangement.

Although more studies are needed, the experiment showed in the first analysis that the cascading method based on microtasks can produce an annotated content that is coherent and comparable to the one produced by an experienced annotator. The worker's contributions were obtained from a commercial crowdsourcing environment, but with a differentiated approach in which only the resources related to the workers and none of the platform resources were used. This approach ensured that both the data set used and the data collected in the contributions were stored only in our database, not the crowdsourcing platform.

It was also observed that the concept of supervised aggregation introduced in this paper obtained positive results. This aggregation approach has proven to be interesting for improving the results of annotation tasks that receive open answers that need to be adjusted manually. A direct conclusion is that this approach can also be used to insert a human verification step at the end of certain aggregation activities.

The video enrichment process has been able to produce an interactive multimedia presentation from a simple raw video through a crowdsourcing approach. This leads to the conclusion that the updated method presented is appropriate to guide this type of project that aims to generate complex media annotations.

6.1 Issues

Some issues were detected in relation to the collection process. About 20% of the contributions cannot be used because they did not meet the desired specifications or because they had invalid or malicious content. This showed that it is necessary to insert in the annotation tools some resources that induce workers to provide valid contributions. It also alerted us to the importance of giving workers clearer instructions on how to properly perform tasks.

In fact, it drew attention to the need to define criteria for the evaluation of instructions given to workers. Also it brings the discussion about the need for a research on whether textual instructions are really sufficient to perform any simple annotation microtasking.

6.2 Crowds Comparison Experiment

Another relevant discussion from the observations of the results concerns the extent to which the composition of the crowd can influence the outcome.

To better understand, the experiment conducted will be replicated in different scenarios:

- Increasing the salary of workers;
- Choosing different contracted groups;
- Choosing only the best workers;
- Using a closed group with volunteers;
- Using a group familiar with the subject treated in the video;

6.3 Next Steps

This work is constantly improving on three different fronts: improvement of the method, improvement of the system, use of the method in different scenarios and applications.

Regarding the method, the next steps include the formalization of the aggregation process, defining specific flows for automatic, supervised and manual models.

New annotation tools are being designed for the structure, which can be adapted for more tasks. In addition, a wizard is being designed to guide the creation of complex media annotation projects based on the method presented.

Some experiments are being prepared to be carried out, some of which seem especially promising:

- Multisensorial video annotation for mulsemedia applications.
- Gesture segmentation in signal language video datasets.
- Semantic approximation of idiomatic expressions in automatic translations.
- Crowdsourcing creation of learning object.

ACKNOWLEDGMENTS

[REMOVERD FOR BLIND REVIEW]

REFERENCES

[1] Luis von Ahn. 2011. Massive-scale online collaboration. (Apr 2011). http://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration

[2] Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on* 15, 1 (2013), 41–55.

[3] Andrew Cross, Mydhili Bayyapunedi, Dilip Ravindran, Edward Cutrell, and William Thies. 2014. VidWiki: enabling the crowd to improve the legibility of online educational videos. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1167–1175.

[4] Bruna C.R. Cunha, Rodolfo Dias Correia, and Maria da Graça Campos Pimentel. 2015. Mobile Video Annotations: A Case Study on Supporting Rehabilitation Exercises. In *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web (WebMedia '15)*. ACM, 245–252. https://doi.org/10.1145/2820426.2820449

[5] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. ACM, New York, NY, USA, 238–247. https://doi.org/10.1145/2736277.2741685

[6] FRANCIS GALTON. 1907. Vox Populi (The Wisdom of Crowds). *Nature* 75, 1949 (1907), 450–451. https://doi.org/10.1038/075509f0

[7] Luke Gottlieb, Jaeyoung Choi, Pascal Kelm, Thomas Sikora, and Gerald Friedland. 2012. Pushing the limits of mechanical turk: qualifying the crowd for video geolocation. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*. ACM, 23–28.

[8] Daniel Haas, Jason Ansel, Lydia Gu, and Adam Marcus. 2015. Argonaut: Macrotask Crowdsourcing for Complex Data Processing. *Proc. VLDB Endow.* 8, 12 (Aug. 2015), 1642–1653. https://doi.org/10.14778/2824032.2824062

[9] M. Hosseini, K. Phalp, J. Taylor, and R. Ali. 2014. The four pillars of crowdsourcing: A reference model. In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*. 1–12. https://doi.org/10.1109/RCIS.2014.6861072

[10] Jeff Howe. 2006. The Rise of Crowdsourcing. *Wired Magazine* 14, 6 (06 2006). http://www.wired.com/wired/archive/14.06/crowds.html

[11] Google Inc. 2017. Google Crowdsource project. (2017). https://play.google.com/store/apps/details?id=com.google.android.apps.village.boond

[12] Gunhee Kim, Leonid Sigal, and Eric P. Xing. 2014. Joint Summarization of Large-Scale Collections of Web Images and Videos for Storyline Reconstruction. In *Proceedings of the 2014 IEEE CCVPR (CVPR '14)*. IEEE Computer Society, Washington, DC, USA, 4225–4232. https://doi.org/10.1109/CVPR.2014.538

[13] Edith Law and Luis von Ahn. 2011. Human Computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 3 (2011), 1–121. https://doi.org/10.2200/S00371ED1V01Y201107AIM013 arXiv:http://dx.doi.org/10.2200/S00371ED1V01Y201107AIM013

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[15] Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. ACM, New York, NY, USA, 233–242. https://doi.org/10.1145/1321440.1321475

[16] Melanie Misanchuk and Tiffany Anderson. 2001. Building Community in an Online Learning Environment: Communication, Cooperation and Collaboration. (2001).

[17] Luyi Mo, Reynold Cheng, Ben Kao, Xuan S. Yang, Chenghui Ren, Siyu Lei, David W. Cheung, and Eric Lo. 2013. Optimizing Plurality for Human Intelligence Tasks. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 1929–1938. https://doi.org/10.1145/2505515.2505755

[18] Venkatesh N Murthy, Subhransu Maji, and R Manmatha. 2015. Automatic image annotation using deep learning representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 603–606.

[19] REMOVED FOR BLIND REVIEW. [n. d.]. REMOVED FOR BLIND REVIEW.

[20] Laurel D Riek, Maria F OâĂŹconnor, and Peter Robinson. 2011. Guess what? a game for affective annotation of video using crowd sourcing. In *Affective computing and intelligent interaction*. Springer, 277–285.

[21] Paul Rohwer. 2010. A Note on Human Computation Limits. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 38–40. https://doi.org/10.1145/1837885.1837897

[22] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

[23] Luis Von Ahn. 2005. *Human Computation*. Ph.D. Dissertation. Pittsburgh, PA, USA. Advisor(s) Blum, Manuel. AAI3205378.

[24] Meng Wang and Xian-Sheng Hua. 2011. Active Learning in Multimedia Annotation and Retrieval: A Survey. *ACM Trans. Intell. Syst. Technol.* 2, 2, Article 10 (Feb. 2011), 21 pages. https://doi.org/10.1145/1899412.1899414

[25] Meng Wang, Xian-Sheng Hua, Jinhui Tang, and Richang Hong. 2009. Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *Trans. Multi.* 11, 3 (April 2009), 465–476. https://doi.org/10.1109/TMM.2009.2012919

[26] Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang. 2014. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 721–730.