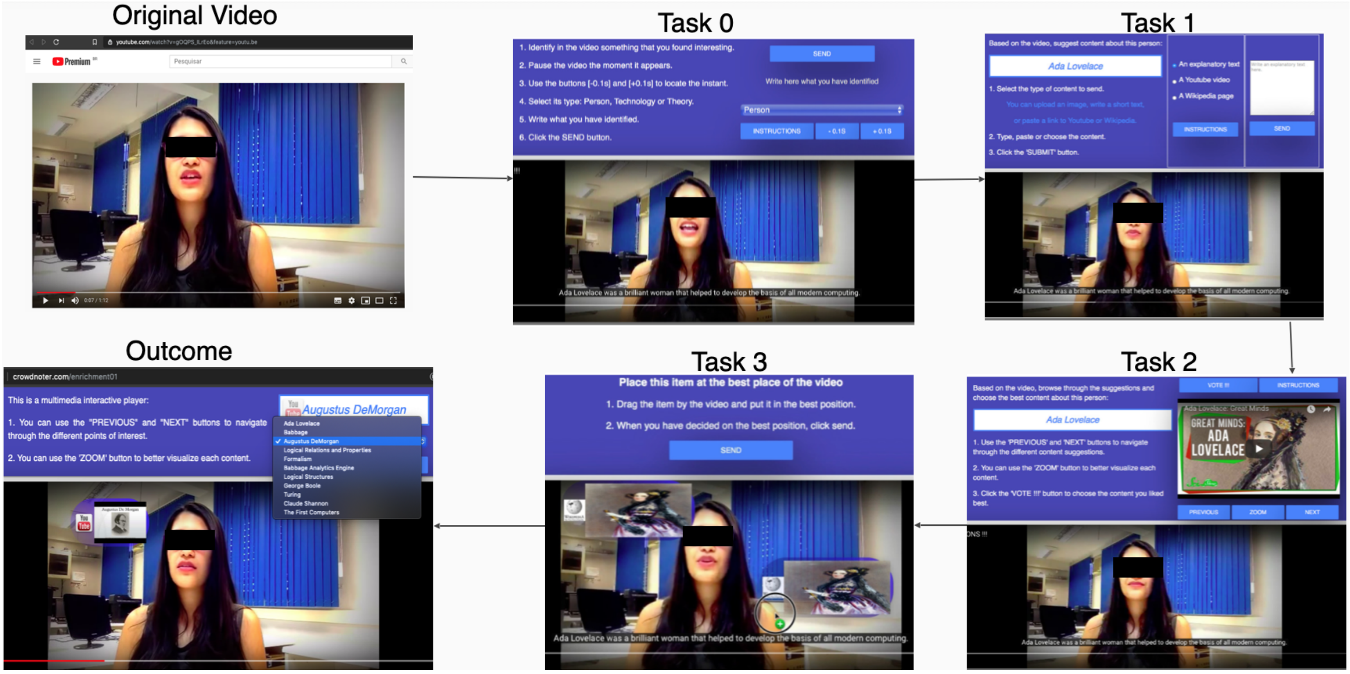
# Integrating Crowdsourcing and Human Computation for Complex Video Annotation Tasks

Marcello N. Amorim
novaes@inf.ufes.br
PPGI / UFES

Celso A. S. Santos
saibel@inf.ufes.br
PPGI / UFES

Orivaldo L. Tavares
tavares@inf.ufes.br
UFES



Task 0:

1. Identify in the video something that you found interesting.
2. Pause the video the moment it appears.
3. Use the buttons [-0.1s] and [+0.1s] to locate the instant.
4. Select its type: Person, Technology or Theory.
5. Write what you have identified.
6. Click the SEND button.

Ada Lovelace was a brilliant woman that helped to develop the basis of all modern computing.

Task 1:

Based on the video, suggest content about this person:

1. Select the type of content to send.
   You can upload an image, write a short text, or paste a link to Youtube or Wikipedia.
2. Type, paste or choose the content.
3. Click the 'SUBMIT' button.

Task 2:

Based on the video, browse through the suggestions and choose the best content about this person:

1. Use the 'PREVIOUS' and 'NEXT' buttons to navigate through the different content suggestions.
2. You can use the 'ZOOM' button to better visualize each content.
3. Click the 'VOTE !!!' button to choose the content you liked best.

Task 3:

Place this item at the best place of the video

1. Drag the item by the video and put it in the best position.
2. When you have decided on the best position, click send.

Outcome:

This is a multimedia interactive player:

1. You can use the "PREVIOUS" and "NEXT" buttons to navigate through the different points of interest.
2. You can use the 'ZOOM' button to better visualize each content.

**Figure 1: Video enrichment pipeline.**

## ABSTRACT

Video annotation is an activity that aims to supplement this type of multimedia object with additional content or information about its context, nature, content, quality and other aspects. These annotations are the basis for building a variety of multimedia applications for various purposes ranging from entertainment to security. Manual annotation is a strategy that uses the intelligence and workforce of people in the annotation process and is an alternative to cases where automatic methods cannot be applied. However, manual video annotation can be a costly process because as the content to be annotated increases, so does the workload for annotating. Crowdsourcing appears as a viable solution strategy in this context because it relies on outsourcing the tasks to a multitude of workers, who perform specific parts of the work in a distributed way. However, as the complexity of required media annoyances increases, it becomes necessary to employ skilled labor, or willing to perform larger, more complicated, and more time-consuming tasks. This makes it challenging to use crowdsourcing, as experts demand higher pay, and recruiting tends to be a difficult activity. In order to overcome this problem, strategies based on the decomposition of the main problem into a set of simpler subtasks suitable for crowdsourcing processes have emerged. These smaller tasks are organized in a workflow so that the execution process can be formalized and controlled. In this sense, this thesis aims to present a new framework that allows the use of crowdsourcing to create applications that require complex video annotation tasks. The developed framework considers the whole process from the definition of the problem and the decomposition of the tasks, until the construction, execution, and management of the workflow. This framework, called CrowdWaterfall, contemplates the strengths of current proposals, incorporating new concepts, techniques, and resources to overcome some of its limitations.

## KEYWORDS

crowdsourcing. human computation. video annotation.

In: II Concurso de Teses e Dissertações (CTD 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

# 1 INTRODUCTION

This article presents an overview of the results obtained in the thesis defended at the Informatics Postgraduate Program of the Federal University of Espírito Santo (PPGI / UFES) in October 2, 2019. The work was developed by Marcello N. Amorim, over 48 months, under the guidance of professor Celso A. S. Santos in collaboration with professor Orivaldo L. Tavares.

Video annotation is a process that aims to supplement this type of multimedia object with metadata that describes its content, context, characteristics, quality, user evaluation, that classifies them, or that associates videos with other media artifacts[11]. This supplementary information can be used to facilitate the work of users and systems that handle annotated items. The annotation serves to highlight points of interest and add information to presented content[5, 10], facilitating the distribution, indexing, summary, navigation and composition of multimedia content.

The systems that require annotated videos have different degrees of complexity and sophistication, ranging from simple search bases to sophisticated object tracking systems throughout videos. This is natural, since media annotation is a resource used to provide information to the system, which can use this information for different types of processing. This observation is essential to delimit the scope of the project, which is the process of generating the annotations, and not their later use.

However, video is a very expressive type of media, capable of carrying a high density semantic load since short videos are capable of transmitting large amounts of information. This characteristic makes the activities necessary for the annotation of videos tend to be costly, both concerning the time needed and the nature of the tasks necessary to perform the annotation. This cost is not so noticeable in cases where the volume of video to be recorded is not significant, but it becomes evident as this volume increases. The demand for annotation of large video datasets motivated the search for automatic annotation methods.

The automation of activities of this nature brings different benefits, ranging from the possibility of automating the entire annotation process to reducing costs with people trained to carry out the necessary tasks. As the complexity of the annotation tasks increases, the degree of training of the people needed to perform them also tends to increase, which also increases the cost of video annotation as well as the difficulty in being able to find people with the requirements.

Automatic methods are an interesting option to solve the problem, however they cannot be applied in all scenarios. Pal et al. presented a case study that exemplifies how to use automatic methods for video annotation [7]. The presented method identifies segmentation points in a video based on the similarity between the frames. In cases where it is possible to describe objectively what should be noted, it is possible to use automatic methods based on similarity or pattern recognition, but in the problems addressed in this thesis, the notes are related to subjective criteria that cannot be clearly described.

There are cases in which it is not possible to describe exactly the sequence of steps required, or what should be observed in the video during the annotation activities, making it impossible to create algorithms to automate the process. Even Machine Learning (ML) systems, which are known for their ability to learn by example, face difficulties in applying in some scenarios since there is no training data for all cases. Besides, when there is a change of context, it is necessary to update the ML models so that they can work correctly, and this restricts the application of this type of system to some domains for which training data exist.

The study presented by Abbas et al. pointed out challenges that still need to be overcome by the deep learning algorithms (from the English Deep Learning (DL) in the context of the analysis of scenes in videos [1]. One of the challenges described is the difficulty to computationally model the way the human being can deal with real-world situations such as occlusion of objects, changes in scenarios, and even geometric transformations. These are examples of situations that make some activities trivial for humans, but difficult for machines. Another problem presented was the difficulty in obtaining appropriate training dataset for certain application scenarios The scarcity of these example bases is a limiting factor for ML The limiting factors identified in the study are present in the human computing problems addressed in this thesis, so that the methods based on DL find it difficult to resolve them.

The study of the literature revealed that, so far, video annotation tasks that require human characteristics such as empathy, sensitivity, subjectivity and intuition still require work from people, and depending on the complexity of the task, or the degree of refinement of the annotations increases , it is necessary to employ increasingly specialized people. It is worth mentioning that the work of specialists tends to be expensive, and it is difficult to find qualified personnel in various activities, such as in the case of interpreters of LIBRAS (Brazilian Sign Language).

Crowdsourcing is a strategy that stands out in this scenario since it consists of outsourcing the execution of tasks to a crowd of workers, who can even be anonymous and non-specialized. Workers' contributions are processed and aggregated using different statistical-based algorithms to generate similar, equivalent, complementary results, or with acceptance similar to those that would be generated by specialists. This can be seen in the article [4], in which crowdsourcing was used to generate multimedia presentations based on video annotation.

The use of crowdsourcing in this type of application is based on two other concepts: human computing and crowd wisdom. Human Computation (HC) is a paradigm introduced by Von Ahn and consists of identifying in a process which are the tasks that require human intelligence to be performed. In turn, the concept of Wisdom of Crowds (WoC), introduced by defines how workers' contributions can be aggregated to produce results that represent common sense, similar to what was approached, on a more conceptual level, by writing about the development of knowledge in collaborative systems.

# 2 MOTIVATION

Video annotation is an activity that, in a centralized approach, takes considerable time and effort from an annotator. Even workgroups face problems annotating longer video content. In this scenario, distributed approaches are an alternative to circumvent the high effort required for manual media annotations[8].

Regarding the work dynamics, the distributed processes for video annotation can be cooperative or collaborative. In collaborative dynamics, note-takers work (work) together to solve the main problem while, in a cooperative approach, each individual solves part of the main problem, operating together to produce a final result, in a Divide and Conquer[6]. Both work dynamics can be used for media annotation, presenting different advantages, disadvantages, and requirements to be applied.

Due to the possibility of parallelizing the annotation process, and because it demands less from the participants by allowing them to annotate smaller media segments, it was decided that the strategy adopted in this research will use a cooperative work dynamic. The approach was chosen to operationalize the crowdsourcing process, which defines an efficient strategy to outsource, for a multitude of workers, tasks that expand human intelligence, aggregating contributions in a final result, which allows recording large amounts of video objects[12].

However, the generation of complex video annotations tends to demand large or complex tasks, in which it is necessary to annotate several aspects of a media object, as well as to relate and contextualize the annotations. This type of task requires the work of specialists or people who are willing to receive training or devote more time and effort to carrying out the task. Thus, it is important to define a strategy to reshape the problems that require complex video annotation tasks, so that they can be solved through a set of simple and easy tasks, which can be distributed and performed by a multitude of collaborators, not necessarily specialized.

## 3 OBJECTIVE

This thesis aims to present a way to apply crowdsourcing and HC to complex video annotation tasks that require human intelligence. One way found to develop a framework that allows you to remodel these complex task is a workflow of small and simple tasks. These simple tasks can then be performed, quickly and easily, by unskilled people and without the need for considerable training or qualification. A workflow represents the order in which a set of tasks must be performed to automate a process.

For organizational purposes, the formalization of the process and computational resources will be listed as specific objectives: (OB1) define the activities, phases, and roles involved in the process in order to make operational the proposed framework, as well as presenting a formal workflow for it, and (OB2) develop a software environment that provides computational support for the execution of the workflow associated with the crowdsourcing process.

The specific objective OB1 aims to formalize a process that represents the roles, phases, and activities necessary to make operational the proposed framework, as well as to present a workflow model that is adequate to guide this process. Thus, it is expected to present a formalized process and a workflow model that is capable of representing crowdsourcing video applications that require complex annotation tasks, based on a composition of microtasks, enabling high-level specification and representation, control workflow, and management of the crowdsourcing process.

The specific objective of OB2 is to develop the software resources and algorithms necessary to filter and aggregate the contributions throughout the proposed process, as well as to develop the computational resources necessary to execute it. The software resources include the environment capable of supporting the execution of the crowdsourcing process, including the tools that allow the tasks to be carried out and the module responsible for monitoring and managing the workflow.

## 4 RESEARCH QUESTIONS

To help understand the extent to which the objectives can be achieved, there are two research questions that will be answered based on the analysis of the problems treated in the case studies.

The first question (RQ1) is: can the video annotations obtained by a crowdsourcing process based on a workflow of small and simple tasks, performed by people who do not need to be experts, generate results equivalent, or complementary, to those generated by large or complex tasks performed by specialists?

The second question (RQ2) is: how can a crowdsourcing process be formalized for the execution of complex video annotation tasks, based on a well-defined workflow, with a description of the activities, control points, and roles involved in the process?

## 5 CASE STUDIES

Three main case studies were carried out based on the framework proposed in this thesis. Although these are different scenarios, everyone used the framework in the same way, as well as using the same computational resources to execute each one. For each of the case studies, a problem was selected that requires complex video processing tasks, as well as requiring human intelligence for its execution. Problems were chosen exclusively that demand subjective tasks and that the final results must reflect the collective intelligence captured by the members of the crowd. Thus, the tasks necessary to solve the problems addressed in the case studies, although trivial for human beings, can be considered as difficult for automatic methods. One of the criteria for choosing case studies was that automatic solutions that could solve them were not available so far. Thus, the framework used does not present itself as a competitor to automatic methods, but as an option for solving problems for which these methods are not available or are not suitable. The case studies used different numbers of crowdsourcing tasks, as well as produced results in different formats.

The first case used four different tasks to produce videos enriched with multimedia features by adding additional content and navigation on raw videos [2]. The second case used two different tasks to analyze videos and produce MPEG-V file with metadata about sensory effects of wind and vibration [3]. The third case used two different tasks to produce, from sign language videos, accessible multimedia presentation with navigation on segmented gestures.

To objectively assess the difference between the result obtained in each case study, the Root Mean Square Error (RMSE) was used, which is suitable for comparing a vector of values obtained with a vector of reference values and provides a mean error value on the same scale as the analyzed values. This error metric is usually applied in crowdsourcing experiments [13].

## 6 RESULTS AND DISCUSSION

In this thesis, a framework called CrowdWaterfall was presented, which aims to facilitate the use of crowdsourcing and human computing in the execution of complex annotation tasks in videos. The attacked scenarios were those in which tasks require human characteristics such as subjectivity, emotion, and empathy to be performed.

The focus of the research was not to present a strategy concurrent to the work of the specialists but to provide a solution that can be used when the work of the specialists is a limiting factor either due to the associated cost or the difficulty in prospecting these professionals. Similarly, the idea was not to present an alternative to automatic methods, but a method capable of covering the scenarios in which they encounter difficulties to be applied, either because there are no annotated datasets to be used in training or because there are no clear rules about how these notes could be produced automatically.

The developed framework is based on a modular workflow in which it is possible to insert crowdsourcing tasks, as well as tasks performed by specialists or even by automatic methods, could be added, there is no factor that prevents this type of task composition. CrowdWaterfall uses the strategy of decomposing a complex video annotation task into a workflow made up of simple annotation tasks, which are suitable to be performed in crowdsourcing campaigns. A strategy was presented to perform the decomposition, as well as the steps for defining workflows, were presented.

## 7 CONTRIBUTIONS

The main contribution of this thesis is the crowdsourcing framework called CrowdWaterfall, which is based on the decomposition of a complex video annotation task into a set of simple crowdsourcing tasks, which are organized in a workflow that can be performed to produce video annotations that traditionally require the work of specialists. The simple tasks used to compose the workflow follow the concept of micro-task and can be performed from crowdsourcing campaigns by workers without specific qualifications or skills, and without the need to dedicate significant time or effort. This framework facilitates the use of crowdsourcing in a set of applications that require complex video annotation activities that require the work of specialists to be carried out.

Another important contribution is the definition of the well-structured workflow that is used in the presented process. This workflow proved to be adequate to guide video annotation processes, but nothing prevents it from being used in other crowdsourcing applications that require the use of multiple simple tasks. The techniques and methods developed during the project are also important contributions since they offer resources that can be useful even for projects that do not use CrowdWaterfall.

Among these features, it is worth mentioning the possibility of creating a separation between the crowdsourcing commercial platform and the system responsible for managing the flow of execution of the crowdsourcing process and the distribution of tasks. This allows you to use commercial crowdsourcing platforms just to recruit and reward workers. The division between the commercial platform and the process execution environment allows collected contributions and aggregated results to be stored exclusively in the place where the owner deems appropriate. This allows companies that deal with sensitive data to consider crowdsourcing as a possible approach for their projects that use sensitive data.

Another contribution produced in this work is a solution for controlling latency in crowdsourcing contributions. This functionality is especially useful when it is necessary to speed up the process of collecting contributions without giving up monitoring the convergence of results in real-time.

The generalization of temporal aggregation methods for crowdsourcing tasks of annotating moments or intervals is also an important contribution since they can be reused in a large number of very common tasks in crowdsourcing processes.

The strategy for monitoring and managing the crowdsourcing workflow is a relevant contribution of this thesis since it provides the means for the crowdsourcing process manager to monitor, in real-time, the status of the result obtained and to control the workflow, either interrupting or resuming the process, whether by arbitrarily directing the execution flow to one of the tasks.

Finally, a software environment was created, called CrowdNoter, which implements all the features mentioned in this section. This environment is free and open, so it can be modified and used by research groups or anyone who wants to implement crowdsourcing processes based on a simple task workflow. CrowdNoter can contribute to the popularization of this type of crowdsourcing process.

## REFERENCES

[1] Qaisar Abbas, Mostafa EA Ibrahim, and M Arfan Jaffar. 2017. Video scene analysis: an overview and challenges on deep learning algorithms. *Multimedia Tools and Applications* (2017), 1–39. https://doi.org/10.1007/s11042-017-5438-7

[2] M. N. AMORIM, F. R. A. NETO, and C. A. S. SANTOS. 2018. Achieving Complex Media Annotation through Collective Wisdom and Effort from the Crowd. In *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)*. 1–5. https://doi.org/10.1109/IWSSIP.2018.8439402

[3] Marcello Novaes de Amorim, Estêvão Bissoli Saleme, Fábio Ribeiro de Assis Neto, Celso A. S. Santos, and Gheorghita Ghinea. 2019. Crowdsourcing authoring of sensory effects on videos. *Multimedia Tools and Applications* (08 Feb 2019). https://doi.org/10.1007/s11042-019-7312-2

[4] Marcello N. de Amorim, Ricardo M.C. Segundo, Celso A.S. Santos, and Orivaldo de L. Tavares. 2017. Video Annotation by Cascading Microtasks: A Crowdsourcing Approach. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web* (Gramado, RS, Brazil) *(WebMedia '17)*. ACM, New York, NY, USA, 49–56. https://doi.org/10.1145/3126858.3126897

[5] 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* (Lisbon, Portugal) *(CIKM '07)*. ACM, New York, NY, USA, 233–242. https://doi.org/10.1145/1321440.1321475

[6] M. Misanchuk and T. Anderson. 2001. Building Community in an Online Learning Environment: Communication, Cooperation and Collaboration. (2001).

[7] G. Pal, S. Acharjee, D. Rudrapaul, A. S. Ashour, and N. Dey. 2015. Video segmentation using minimum ratio similarity measurement. *International journal of image mining* 1, 1 (2015), 87–110. https://doi.org/10.1504/IJIM.2015.070027

[8] CAS Santos, Alexandre SANTOS, and TA Tavares. 2007. Uma estratégia para a construção de ambientes para a descrição semântica de vídeos.

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

[10] 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

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

[12] Mengyao Zhao and André van der Hoek. 2015. A brief perspective on microtask crowdsourcing workflows for interface design. In *Proceedings of the Second International Workshop on CrowdSourcing in Software Engineering*. IEEE Press, 45–46. https://doi.org/10.1109/CSI-SE.2015.16

[13] Tingting Zhu, Joachim Behar, Tasos Papastylianou, and Gari D Clifford. 2014. CrowdLabel: A crowdsourcing platform for electrophysiology. In *Computing in Cardiology 2014*. IEEE, 789–792.