Recording and Analyzing Benchmarking Results: The Aims of the MediaEval Working Notes Proceedings

Martha Larson[1], Gareth Jones[2], Bogdan Ionescu[3], Mohammad Soleymani[4], Guillaume Gravier[5]

[1]Delft University of Technology, Netherlands [2]Dublin City University, Ireland
[3]University Politehnica of Bucharest, Romania [4]University of Geneva, Switzerland
[5]CNRS IRISA and Inria Rennes, France
m.a.larson@tudelft.nl, gareth.jones@computing.dcu.ie, bionescu@imag.pub.ro
mohammad.soleymani@unige.ch, guig@irisa.fr

ABSTRACT

We present an in-depth look at the structure and the strategy of the working notes proceedings that is published by the Benchmarking Initiative for Multimedia Evaluation (MediaEval) in conjunction with its yearly workshop. The proceedings records information on the tasks that were offered by the benchmark and the approaches that were developed by participants to address them. The proceedings is called a 'working notes' because its aim is to support work in progress. Specifically, it presents analyses of participants' algorithms for discussion at the workshop. This year, in addition to sections devoted to each of the tasks, we are piloting a new section called *MediaEval Letters*, for papers that transcend individual tasks or years of the benchmark.

1. INTRODUCTION

MediaEval is a benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval. It organizes an annual campaign, which offers challenges to the multimedia research community. The challenges take the form of tasks that invite the exploitation of multiple modalities (i.e., speech and music, visual content, textual metadata, context). MediaEval's focus on human and social aspects of multimedia sets it apart from other benchmarks.

MediaEval is a 'bottom-up benchmark'—it is not run by a central project or institution. Instead, tasks are proposed and organized by autonomous groups of task organizers. Proposed tasks are accepted into the benchmark on the basis of a community-wide survey that determines 'grassroots' interest in participating in the task. If the survey reveals a sufficient level of 'demand' for the task, the task is vetted for viability and is offered by the benchmark.

The responsibility for vetting tasks lies in the hands of the MediaEval Community Council, a group of volunteers. In 2015, the MediaEval Community Council comprised the authors of this paper. The Council also works to ensure that tasks offered are truly 'MediaEval tasks', in that they involve human and social aspects, and encourage multimodal solutions, and also complement other benchmarks.

A guiding principle for MediaEval is 'less is more'. Each task must commit itself to a single, official evaluation metric (or evaluation procedure), and each participating team may submit no more than five different algorithms (referred to as 'runs') that address a task. These limitations force the task organizers to clearly formulate the goal of a task for a given year, and force participants to commit in advance to the methods that they find most promising.

This pressure keeps the community focused on the *motivation* behind their research and development activities. Tasks are related to user needs that arise within specific use scenarios for multimedia technology. The needs must be defined clearly enough such that task solutions can be meaningfully compared using a single evaluation procedure. Task participants do not generate solutions blindly, but rather pursue only those solutions that they find to be most promising.

The yearly MediaEval benchmarking cycle concludes with a workshop bringing together task participants to report on and discuss the current tasks and plan for the future. The workshop publishes a proceedings consisting of working notes papers. This proceedings serves as a record of the output of the benchmark in any given year, including descriptions of the tasks and the solutions offered, as well as analysis of the performance and effectiveness of these solutions. The purpose of this paper is to describe the structure and the strategy of the working notes proceedings, and to introduce *MediaEval Letters*, a new section of the proceedings that is being piloted in 2015.

2. THE WORK OF WORKING NOTES

MediaEval is a benchmark that follows in the tradition of evaluation campaigns in the information retrieval community. It was established in 2008, as a track of the first generation of the CLEF campaign (then called 'Cross-Language Evaluation Forum'), which ran 2000–2009 [1]. This campaign compiled a working notes that was made available to participants at the yearly workshop—examples include [3, 4]. Later, a proceedings volume with revised selected papers from the benchmarking year was published.

MediaEval adopted the CLEF practice of a working notes. Another example of a similar practice is the 'notebook' published each year by the TRECVid campaign [6]. The current generation of the CLEF campaign (now called 'Conference and Labs of the Evaluation Forum') publishes a conference proceedings. See [2] for a complete list of CLEF proceedings.

The form of the MediaEval working notes proceedings follows directly from its intended function. The function of the working notes is to 'freeze' the participants' approaches to

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

the tasks at the moment of the run submission deadline. Working notes papers contain the information necessary to support discussion of participants' approaches at the workshop, as well as reproduction of their approaches after the workshop. They should: describe the algorithms used in the individual runs, report the scores achieved with respect to the official evaluation metric, and analyze the strengths and weaknesses of the approach. They may also discuss runs beyond the five submitted runs, or present evaluation with respect to alternate evaluation metrics. On the basis of the information in the working notes papers, it should be possible to understand what the most promising approaches are to a task, and how the task (and/or the evaluation methodology) might evolve in future years.

3. MEDIAEVAL WORKING NOTES

The MediaEval working notes proceedings consists of sections devoted to tasks. The section begins with the overview paper, and is followed by the papers of the participants. Papers are intentionally kept short (two pages), which has several functions. First, it forces authors to think carefully about the task, and convey only the most important information or insights to the readers. Second, it allows the people involved in the task to get a quick overview of what was done in the task, because the papers are short and easy to read. A team of editors is drawn from among the task organizers, and coordinates the peer review and revision process that ensures the quality of the proceedings.

The overview paper explains the objective of the task, and the task definition, which provides a specification of the challenge to be addressed. It describes the use scenario that motivates the task, and discusses related work. If the task has been offered before, the overview paper explains the relationship of the current task to previous editions in past years. In many cases, the overview paper will also offer an outlook for further development of the task in future years.

Participant papers focus on the approaches developed by the participating teams to address the tasks. They provide a description of the approach chosen by a team, and explain why this choice was made. The paper should explain succinctly the novelty of the approach, and/or the main insight on which the authors build. The participants' papers cite the overview paper for the task. In this way, participants avoid repeating the entire description and motivation of the task in their own papers, and can focus instead on their own algorithms.

In the participant paper, the related work that is relevant to the specific approach to the task that was developed by the team is covered. In 2015, we extended the length of the papers to include a third page containing only references. This was done in order to make sure that the short length of the paper did not force the authors to compromise on explaining how their approach is related to existing approaches.

The short length of MediaEval working notes papers is part of the overall 'less is more' strategy. Limited space encourages authors to focus on the most essential details. This also helps to promote their status as 'work in progress'. After the workshop, participating teams are encouraged to bring their work to fully maturity, and submit it for publication at a mainstream venue. In many cases, groups consisting of organizers and also task participants form during the workshop, and go on to author joint publications about the task. In 2015, the MediaEval Working Notes proceedings will be published with CEUR-WS.org for a fifth year. CEUR-WS.org allows the rights to the papers to remain with the authors, further encouraging follow-up publications.

4. GAINS GOING BEYOND 'WINNING'

At the workshop, the task organizers present a ranked list of the runs that were submitted to the task, and a winner is declared. However, MediaEval scrupulously avoids emphasizing 'winning' as the main goal of participation in the benchmark. Over-focus on achieving the top score discourages participants from taking risks. Without risks, the algorithms that are proposed to address a task are in danger of converging on a local optimum, as participants seek only incremental improvements to the best approaches of past years. As such, a mark of the value of the contribution of a participating team to the benchmark is the quality of their working notes paper. In order to highlight innovation and productive risk taking, MediaEval chooses a certain number of teams each year to receive an MDM (MediaEval Distinctive Mention). MDMs are used by the task organizers to point out submissions that they see as having particularly high promise, but did not achieve a top ranking scores.

The focus on working notes paper quality and not winning is also important in order to highlight task organizers' solutions to their own tasks. As an outward symbol of the level playing field that MediaEval meticulously maintains, runs submitted by organizers are excluded from the official ranking. The working notes paper is the opportunity for the task organizers to allow their own approaches to stand out.

5. MEDIAEVAL LETTERS

The MediaEval community produces results and insights that often go beyond a single working notes paper of a given task in a given year. In 2015, we are piloting a section of the MediaEval working notes called *MediaEval Letters* to provide a venue for publication of such papers. Although a MediaEval Letters paper may be on any topic, we are particularly interested in promoting several topics.

- *Reproducibility and replication* Insights gained from reimplementation of algorithms from past MediaEval working notes.
- *Best Practices* Proposals for extending the effectiveness or usefulness of the benchmark, for example, Adam Rae's 2012 talk on code sharing [5].
- *Evaluation Methodology and Metrics* New ways of evaluating tasks. Such contributions are necessary to keep pace with task innovation.
- *New Tasks* Proposals or proofs-of-concept for future MediaEval tasks.
- *'MediaEval history'* Studies devoted to tracking where we have been: Bibliographic studies, or studies devoted to measuring the impact of MediaEval.

The papers in the MediaEval Letters section are reviewed by the MediaEval Community Council, who may ask for support from other community members. A Letters paper is considered successful if it triggers productive discussion in the community during the writing and review process as well as after publication. Moving forward, we hope the Letters section will contribute to making the MediaEval working notes proceedings useful and informative.

6. REFERENCES

[1] *Cross Language Evaluation Forum*, (accessed Sept. 2015). http://www.clef-campaign.org.

[2] The CLEF Initiative. *Proceedings*, (accessed Sept. 2015). http://www.clef-initiative.eu/publication/proceedings.

[3] Cross Language Evaluation Forum. *Working Notes for the CLEF 2008 Workshop*, 2008 (accessed Sept. 2015). http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[4] Cross Language Evaluation Forum. *Working Notes for the CLEF 2009 Workshop*, 2009 (accessed Sept. 2015). http://www.clef-campaign.org/2009/working_notes/CLEF2009WN-Contents.html.

[5] A. Rae. *MediaEval Code of Conduct*, 2012 (accessed Sept. 2015). http://www.slideshare.net/adamrae/code-sharing.

[6] TRECVID. *TREC Video Retrieval Evaluation Notebook Papers and Slides*, (accessed Sept. 2015). http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.