International Workshop on Deep Video Understanding

Keith Curtis
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
keith.curtis@nist.gov

George Awad*
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
george.awad@nist.gov

Shahzad Rajput*
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
shahzad.rajput@nist.gov

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
ian.soboroff@nist.gov

ABSTRACT

This is the introduction paper to the International Workshop on Deep Video Understanding, organized at the 22nd ACM Interational Conference on Multimodal Interaction. In recent years, a growing trend towards working on understanding videos (in particular movies) to a deeper level started to motivate researchers working in multimedia and computer vision to present new approaches and datasets to tackle this problem. This is a challenging research area which aims to develop a deep understanding of the relations which exist between different individuals and entities in movies using all available modalities such as video, audio, text and metadata. The aim of this workshop is to foster innovative research in this new direction and to provide benchmarking evaluations to advance technologies in the deep video understanding community.

CCS CONCEPTS

• **Information systems** → *Multimedia content creation*; **Multimedia and multimodal retrieval**.

KEYWORDS

video understanding; multimedia; multimodal interaction; information retrieval; video ontology

ACM Reference Format:
Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. International Workshop on Deep Video Understanding. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3382507.3419746

1 WORKSHOP INTRODUCTION

Deep video understanding is a difficult task which requires systems to develop an insightful analysis and understanding of the relationships among different entities in video, to use known information to reason about other, more hidden information, and to populate a knowledge graph (KG) with all acquired information.

There has been efforts to encourage research in high level video understanding such as the "MovieQA" [6] and "The Large Scale Movie Description Challenge" [5]. However these tasks revolve around isolated visual concepts retrieval and not about testing systems for their overall understanding of entities, relations and events within the video/movie. Early visions of this kind of work [4] proposed to use visual and audio descriptors, in addition to employing semantic analysis and linking with external knowledge sources in order to populate a knowledge graph. A large-scale dataset of corresponding movie trailers, plots, posters, and metadata was developed by [1] who study the effectiveness of visual, audio, text, and metadata-based features for predicting high-level information about movies such as their genre or estimated budget.

To work on this task, a system should take into consideration all available modalities (speech, image/video, and in some cases text). The aim of this workshop is to push the limits of multimodal extraction, fusion, and analysis techniques to address the problem of analyzing long duration videos holistically and extracting useful knowledge to utilize it in solving different types of queries. The target knowledge includes both visual and non-visual elements. As videos and multimedia data are getting more and more popular and usable by users in different domains, the research, approaches and techniques we aim to be applied in this workshop will be very relevant in the coming years and near future. The call for contributions in this workshop supported long, short and abstract papers related to multimedia understanding, in addition to an optional track for researchers who are interested to apply their techniques on a new pilot creative commons movie dataset (HLVU)[3] collected by the NIST, who distributed the dataset with development data, testing queries, and finally evaluated and scored the submitted runs by participating researchers. In this optional challenge track, all participants were invited to submit a paper describing their approaches to solve the testing queries. The following sections summarize the two tracks of the workshop.

1.1 Track 1

In this track authors were invited to apply their approaches and methods on a novel High-Level Video Understanding (HLVU) dataset made available by the workshop organizers which included 10 movies with a Creative Commons (CC) license. These movies were annotated by human assessors and full ground truth, including the Knowledge Graph of all entities and relationships, was made available for 6 of these movies which made up the development set.

*Georgetown University

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3419746

Evaluation and scoring was supported for two main query types distributed with the dataset:

(1) Multiple choice question answering on Knowledge Graph for selected movies.
(2) Path analysis between persons / entities of interest in a Knowledge Graph extracted from selected movies.

Movies were annotated by human assessors who drew a Knowledge Graph of nodes. Nodes were comprised of important persons, entities, and major concepts from the storyline. Edges connecting these nodes were of the relationship type between persons, entities, and concepts. For example if *Person A* was a Student at *University*, there was an edge named *Studies At* connecting *Person A* to *University*. Annotators were given a predefined ontology of most common relationships between people (family-based, workplace-based, social, etc). Graphical examples of this are available at [3] and is available online at [2].

Using this annotation scheme, a set of multiple choice questions were generated for the test set movies on the relationships between various nodes in this KG. In addition, the full path connections between various nodes were analysed and participants were asked to submit possible paths. For example, a path question would ask how is person X connected to person Y and systems were required to return back all of the possible paths with correct relationships between the two persons. Further details on this annotation scheme, question types, metrics, and movies used for this dataset are available from [3].

1.2 Track 2

In this track authors were invited to submit contributions related, but not limited, to the following topics applied on the provided HLVU dataset or any external datasets:

- Multimodal feature extraction for movies and extended video
- Multimodal fusion of computer vision, text/language processing and audio for extended video / movie analysis
- Machine Learning methods for movie-based multimodal interaction
- Sentiment analysis and multimodal dialogue modeling for movies
- Knowledge Graph generation, analysis, and extraction for movies and extended videos

2 WORKSHOP CONTENT

2.1 Keynote Speakers

In addition to authors of accepted papers, we have invited three keynote speakers bringing different perspectives to this challenging research direction:

- **Klaus Schoeffmann (Klagenfurt University, Austria)**
  Dr. Klaus Schoeffmann is an Associate Professor at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. His research focuses on video content understanding (in particular of medical/surgery videos), multimedia retrieval, interactive multimedia, and applied deep learning.
- **Dima Damen (University of Bristol, UK)**
  Dr. Dima Damen is a Reader (Associate Professor) in Computer Vision at the University of Bristol, United Kingdom. Her research interests are in the automatic understanding of object interactions, actions and activities using static and wearable visual (and depth) sensors.
- **Koichi Shinoda (Tokyo Institute of Technology, Japan)**
  Prof. Koichi Shinoda is a professor with the Tokyo Institute of Technology, Japan. His research interests include speech recognition, video information retrieval, statistical pattern recognition, and human interfaces.

2.2 Papers

At the time of writing this introduction paper, the workshop paper submissions were still open. We had one participant in Track 1 of this workshop, and a number of interesting works submitted for Track 2 of this workshop. We hope that this will provide a good balance of interesting work related to this research area. However, due to the fact that paper reviews had yet to be completed we are unable to provide further information on accepted papers at this stage.

3 WORKSHOP ORGANIZATION

3.1 Organizing Committee

- Keith Curtis (National Institute of Standards and Technology, USA)
- George Awad (Georgetown University & National Institute of Standards and Technology, USA)
- Shahzad Rajput (Georgetown University & National Institute of Standards and Technology, USA)
- Ian Soboroff (National Institute of Standards and Technology, USA)

4 CONCLUSIONS

There is a good diversity of research presented at this workshop, and we are satisfied that we have encouraged further work in this new direction. The research presented at this workshop coupled with analysis of what has been learned during the annotation of this dataset and running the benchmark campaign has provide a good basis for the extension and continuation of this research into the next few years.

ACKNOWLEDGMENTS

The authors would like to thank the Information Technology Laboratory at NIST for sponsoring all annotation work done on the HLVU dataset and sponsoring the first three organizers of this workshop. Also we thank the ICMI workshop chairs and main conference for including this new workshop, and all paper authors for their research efforts and contributions.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for

the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST, or the U.S. Government.

REFERENCES

[1] Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. 2019. Moviescope: Large-scale Analysis of Movies using Multiple Modalities. *arXiv preprint arXiv:1908.03180* (2019).

[2] Keith Curtis and George Awad. 2020 (accessed August 26, 2020). *DVU Challenge.* https://drive.google.com/drive/folders/1q1Ca0aFJrF9tB8hsw-mrI9d4tzy5wlPZ

[3] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval.* 355–361.

[4] Jeremy Debattista, Fahim A Salim, Fasih Haider, Clare Conran, Owen Conlan, Keith Curtis, Wang Wei, Ademar Crotti Junior, and Declan O'Sullivan. 2018. Expressing Multimedia Content Using Semantics—A Vision. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC).* IEEE, 302–303.

[5] Anna Rohrbach and Jae Sung Park. 2019. Large Scale Movie Description Challenge (LSMDC) 2019. https://sites.google.com/site/describingmovies/lsmdc-2019, Last accessed on 2019-11-06.

[6] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4631–4640.