Proceedings of the workshop on People in Vision, Language, and the Mind @LREC2022 , pages 12–16
Marseille, 20 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Development of a MultiModal Annotation Framework and Dataset for Deep Video Understanding

Erika Loc†, Keith Curtis*, George Awad*‡, Shahzad Rajput*‡, Ian Soboroff*
Montgomery College†, National Institute of Standards and Technology*, Georgetown University‡
Maryland, USA
x.erika.loc@gmail.com, {keith.curtis, george.awad, shahzad.rajput, ian.soboroff}@nist.gov

Abstract

In this paper we introduce our approach and methods for collecting and annotating a new dataset for deep video understanding. The proposed dataset is composed of 3 seasons (15 episodes) of the BBC Land Girls TV Series in addition to 14 Creative Common movies with total duration of 28.5 hr. The main contribution of this paper is a novel annotation framework on the movie and scene levels to support an automatic query generation process that can capture the high-level movie features (e.g. how characters and locations are related to each other) as well as fine grained scene-level features (e.g. character interactions, natural language descriptions, and sentiments). Movie-level annotations include constructing a global static knowledge graph (KG) to capture major relationships, while the scene-level annotations include constructing a sequence of knowledge graphs (KGs) to capture fine-grained features. The annotation framework supports generating multiple query types. The objective of the framework is to provide a guide to annotating long duration videos to support tasks and challenges in the video and multimedia understanding domains. These tasks and challenges can support testing automatic systems on their ability to learn and comprehend a movie or long video in terms of actors, entities, events, interactions and their relationship to each other.

Keywords: Dataset, Multimodal, Multimedia, Annotation Framework, Video Understanding

1. Introduction

In this paper we use the term Deep Video Understanding (DVU) to refer to the ability of making sense of and understanding long duration videos with a self contained storyline such as movies and TV series. This is a difficult challenge requiring a suitable dataset which has been annotated to both the entire movie and to the individual scene level. Such a dataset must include annotations of characters & entities, as well as relationships and interactions between these, chronological ordering of such interactions, scene sentiment annotations, and natural language descriptions of individual scenes.

As this research is performed over the whole movie and individual scenes, the development of this dataset is separated into these two distinct parts to support different requirements. The whole-movie annotations support research on the movie level for the extraction of all main characters, entities, and relationships between them. Scene-level annotations support research on the scene level for the extraction of characters in each scene, interactions between characters, and the chronological order of interactions.

In this paper we describe the construction of such a corpus to support this research. Our corpus consists of all 15 episodes from the BBC TV series *Land Girls*[1], and 14 Creative Commons (CC) licensed movies.

The remainder of this paper is structured as follows: Related work is discussed in Section 2. Section 3 describes the dataset in detail. Full descriptions of the annotation framework are provided in section 4, while supported query types are explained in section 5. Finally we discuss how this annotation efforts were utilized in public multimedia grand challenges in section 6.

[1]https://www.bbc.co.uk/programmes/b00xxnhv/episodes/guide

2. Related Work

MovieQA (Tapaswi et al., 2016) is a dataset which aims to evaluate automatic story comprehension from video and text. It consists of 14,944 multiple choice questions, each with 5 multiple-choice answers, with one of these being the correct answer, from about 408 movies with high semantic diversity. Movies were segmented into video clips with a maximum duration of 200 seconds where participants have to answer a question related to the clip. The dataset itself comes with multiple answering sources for questions such as plot synopses, scripts, subtitles, and audio descriptions. The plot synopses was used by annotators to come up with questions and answers rather than watching the whole movie.

The *MovieGraphs* dataset (Vicol et al., 2018) provides detailed graph-based annotations of social situations depicted in movie clips. Annotations are provided for characters in each clip, their emotional and physical attributes, and relationships and interactions between characters.

In (Lei et al., 2020) work, the authors collected 108,965 queries on 21,793 videos from 6 TV shows where queries can target the visual or subtitle modalities. Queries are textual and only target specific moments in the TV show.

Early visions of video understanding (Debattista et al., 2018) explored the usage of visual and audio descriptors, in addition to employing semantic analysis and linking with external knowledge sources in order to populate a knowledge graph.

High-level Video Understanding *(HLVU)* (Curtis et al., 2020a) describes a vision for video understanding over the whole movie level. Knowledge Graph annotations were used to describe the overall storyline of movies and characters contained within. A challenge was run testing systems on their ability to understand movies at a high-level over the whole movie. The first workshop on HLVU (Cur-

tis et al., 2020b) challenged participant systems to extract, understand, and answer queries over the full movie.
In this work our contribution is the development of an annotation framework for the specific task of Deep Video Understanding - making sense of movies, the characters there within, and the relationships and interactions between such. The work presented in this paper extends the HLVU work deeper to the scene-level, thereby requiring the development of a suitable dataset, segmented to the scene-level, and annotated over the whole movie and the scene-level.

3. Dataset

In order to undertake this new research area, there was a critical need to identify a representative dataset to work with and be able to distribute it to researchers as most of the available datasets in the computer vision and video analysis domains are not suitable due to various reasons such as lack of properly licensed free open movies, most available video datasets are either from social media user uploads, or covering specific application domains such as surveillance, action and activity detection, etc. To tackle this problem, the authors applied two approaches to recruit datasets: a) searching for Creative Common (CC) (Creative Commons, 2019) movies publicly available, b) reaching out to big broadcasting companies to license TV Series. The following sections explain these two efforts and their datasets characteristics.

3.1. Creative Common Movies

The most important criteria in selecting the movies of the dataset were reasonable video quality, duration of more than 15 min at least, and self contained story lines with clear actors, relations, events and entities. In total, a dataset of 14 movies (17.5 hrs) has been collected from public websites such as Vimeo[2], the Internet Archive[3] and YouTube[4]. Table 1 shows the current set of collected movies, their genre and durations. All movies have been deemed by the authors to be suitable for this research.

3.2. Licensed TV Series

The authors have also been in deliberations with the BBC regarding the licensing of the TV show *Land Girls* for use in this dataset. This is a 3-season / 15-episode series set in World War 2 about the lives of a group of women doing their part for Britain in the *Women's Land Army* during the war. Each episode is about 45 mins long and the whole 3-season set is about 11 hrs. Automatic audio transcripts were also provided by the BBC with the series. This paper presents our efforts annotating the first 2 seasons of Land Girls series.

3.3. Dataset characteristics

In order to highlight some content characteristics for both types of data we collected, table 2 shows the total number of scenes, entities (key characters and locations), unique relationships between either characters and each other or characters and locations, and finally interactions between characters. We differentiate between actions and interactions in this work by restricting interactions to be between people (e.g. talking with), while actions can be done solely by individual character (e.g. running). In the presented annotations framework we focused more on interactions.

Figure 1: Node Shapes in Movie-level KG. In this context the word person and character are used interchangeably

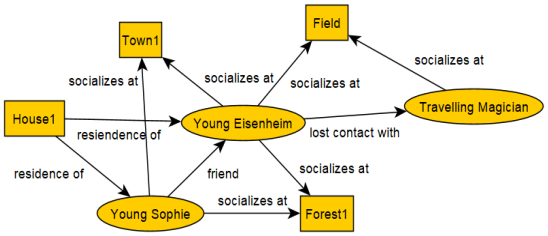

Figure 2: Movie-level KG

4. Annotation Framework

Our annotators created datasets for each film at the movie-level and the scene-level, both focusing to capture their own details from the films. Annotation at either the movie or scene level requires first that the annotator watch the film all of the way through to gain a general understanding of the story. During this stage our annotators make mental note of all of the locations and which characters and entities are relevant to the overall plot of the film as not every character or entity that appears in the film is documented in the datasets at the movie-level. When annotating at the movie-level, our annotators utilized yEd Graph Editor[5], a general-purpose diagramming tool to exhibit the relationships between locations, characters, and concepts. For scene-level annotations, we created an internal annotation tool to be employed for the process. This annotation tool was written in HTML/CSS and JavaScript, and is a combination of two pre-existing tools. Necessary components and features from both sources were integrated into the final tool. Such features include a snapshot saver to capture images of key characters and locations directly from the films, a canvas tool to create knowledge graphs (KGs), and a scene selection area to navigate between scenes. Some newly integrated components consist of a right-click menu for labeling nodes within the knowledge graph (KG), a text area to add natural language descriptions, as well as save buttons to save the knowledge graph and text area's contents locally. The vocabulary[6] that is used in the knowledge graphs at both the movie and scene level, aside from

[2]https://vimeo.com/
[3]https://archive.org/
[4]https://www.youtube.com/
[5]https://www.yworks.com/
[6]https://www-nlpir.nist.gov/projects/trecvid/dvu/training/vocab.dvu.json

Movie	Genre	Duration
Honey	Romance	86 min
Let's Bring Back Sophie	Drama	50 min
Nuclear Family	Drama	28 min
Shooters	Drama	41 min
Spiritual Contact The Movie	Fantasy	66 min
Super Hero	Fantasy	18 min
The Adventures of Huckleberry Finn	Adventure	106 min
The Big Something	Comedy	101 min
Time Expired	Comedy / Drama	92 min
Valkaama	Adventure	93 min
Bagman	Drama / Thriller	107 min
Manos	Horror	73 min
Road to Bali	Comedy / Musical	90 min
The Illusionist	Adventure / Drama	109 min

Table 1: The DVU Dataset of 14 open source movies

entity names, are derived from a predetermined ontology in order to prevent disparity within the data. This vocabulary included classes of relationships (social, family, work-related, person-place), locations, sentiments, interactions, and emotions. Overall, each movie was annotated by only 1 annotator while Land Girls TV series was all annotated by a summer student for a duration of about 5 month. In total we hired 6 annotators and they were all given 1-2 hrs training sessions to describe the purpose and usage of the tool. The following two subsections provide more details about the movie-level and scene-level annotations.

4.1. Movie-level Annotations

The yEd graph editor is used to document the film at the movie-level or as a whole. Nodes in different shapes as illustrated in figure 1 are used to distinguish between various movie elements with rectangles representing locations, ellipses for characters, and triangle for concepts. Concepts are used to highlight dominant ideas that play a major role in any movie and usually one or more key characters are involved into engaging with such ideas (e.g. bad dream, imaginary figure, etc). Figure 2 exhibits an example of the movie level knowledge graphs (KGs). Rays connecting the nodes depict the relationship between each entity the characters interact with and the locations they appear in. The final output knowledge graph are saved by annotators as xgml file storing all data structure needed to reconstruct the graph in the future if needed to do any updates.

4.2. Scene-level Annotations

Once the whole-movie annotations have been recorded, our annotators move onto documenting the film at the scene-level with the annotation tool created internally. Figure 4 shows the interface of the web tool used for scene annotation. The films are segmented into scenes each lasting roughly 20 seconds to 2 minutes long. Using the scene selector on the tool to navigate from scene to scene, the film is re-watched to observe more in-depth details not included in the knowledge graphs (KGs) created at the movie-level. Snapshots of each location, relevant character, and entity are captured throughout the scenes. Ideally 5 or more images of each, captured at various angles to ensure variety amongst the snapshots are saved across the entire film. Cataloging of relationships and interactions between each of the characters & entities within each scene is done in the canvas of the annotation tool. Similar to the whole-movie annotations, different shapes represent individual aspects as shown in Figure 3. A sample knowledge graph (KG) is shown in Figure 5 illustrating all interactions taking place in chronological order. The text description for the same scene is shown in figure 6. It can be shown how both the text description and scene graph both complements each other. Each scene knowledge graph is finally saved as a json file to store all node information and links between nodes and each other.

Dataset	Scenes	Entities	Relations	Interactions
Movies	621	1572	650	2491
TV Series	422	390	711	1622

Table 2: Dataset content of scenes, entities (characters, locations, & concepts), relationships between entities, and interactions between characters

5. Query Design and Generation

A set of queries were designed to test participating systems on their understanding of the test movies at the movie-level and the scene-level. Movie-level queries asked three main sets of questions: Multiple choice questions on the part of Knowledge Graph for selected movies, possible path analysis between selected persons / entities of interest in a movie, and Fill in the Blank Space, in which systems were asked to fill in the graph space for a partial Knowledge Graph of movies. Scene-level queries asked five main sets of questions: Find the next / previous interaction, Find the unique scene, Match selected scenes with natural language descriptions, Fill in the graph space, and Match scenes with scene sentiment labels.

The majority of queries on both the movie-level and the scene-level were generated automatically. Additional queries which required human generation were: Path analysis questions on the movie-level, Match scenes with description on the scene-level, and Match scenes with sentiment labels on the scene-level.

For generating path analysis questions, two character nodes on the movie-level KG were chosen for each question which had an indirect connecting path between them. For match scene-description questions, scene descriptions and KG's were analysed and scenes and related descriptions considered to be sufficiently different were chosen. Similarly for matching scenes with sentiments, scene KG's and sentiment labels sufficiently different were chosen.

Figure 3: Web tool interface for scene-level annotations

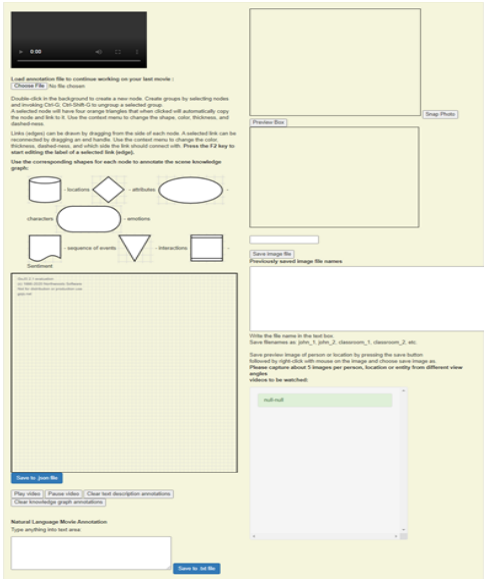

Figure 4: Web tool interface for scene-level annotations

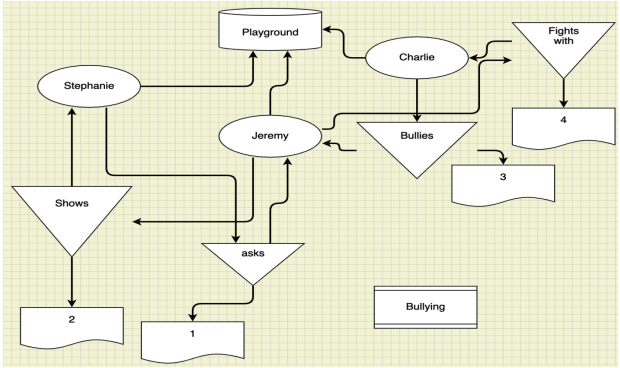

Figure 5: Scene-level KG

The bell rings for recess and Jeremy is sitting on the bench by the playground reading a comic book. Stephanie runs by but comes back to talk to Jeremy. She sits down on the bench next to him, and they discuss comics while he shows her his comic book. Charlie Luther walks up and grabs it from Jeremy's hands.and rips it up. Jeremy gathers the ripped pages from the ground and Stephanie stands up for him. Jeremy imagines the Great Celestial shows up to save the day. But in reality, he shoves Charlie to the ground out of anger then faints. Stephanie calls for Ms. Johnson.

Figure 6: Scene text description sample

6. Discussion and Conclusion

The described annotation framework was followed and used to generate queries to support the deep video understanding ACM Multimedia Grand Challenge in 2020 and 2021[7], as well as the ACM Multimedia Asia Grand Challenge in 2021[8]. In these challenges the participants were given the original whole movies, snapshot images for key characters and location entities, the ontology of relationships, sentiments, interactions, locations and character emotional status. The annotated dataset was divided into training and testing sets. In 2021 the training set consisted of 10 movies, while participants were tested on 4 movies. The provided training set additionally contained the movie-level and scene-level knowledge graphs and scene text descriptions. We should note here that unfortunately the Land Girls TV series videos couldn't be distributed due to lack of time in securing the hosting agreement between the BBC and the hosting university. However, all annotations are now public and available for researchers [9].

In total, 6 systems (Yu et al., 2020), (Baumgartner et al., 2020), (Anand et al., 2020),(Zhang et al., 2021b),(Anand et al., 2021),(Zhang et al., 2021a) submitted solutions in the two years combined. Based on these two grand challenge results, we observed that systems tend to perform better on scene-level queries compared to movie-level. This could be due to the scene specific queries such as interactions between two specific characters or the sentiment of a given scene. On the other hand the hardest movie-level query is the path analysis between two characters or in other words how is character X related to character Y which requires correctly representing the movie relationships and understanding in higher level how the whole movie storyline unravels.

To conclude, in this paper we introduced our new dataset of movies and TV series and explained how we developed a novel annotation framework to describe each movie or episode at two levels. First, a global level using a static knowledge graph to represent how each entity is related to each other, and second at a more fine-grained level per scene to capture interactions, sentiments and other scene characteristics. The framework supports automatic query generation to test systems on various visual and non-visual facets and their ability to comprehend a visual storyline with many characters, relationships and locations. As this domain is gaining attention and more research groups are looking into how to apply multimodal integration techniques to process visual, audio and textual information channels, we anticipate the need for similar annotation frameworks and datasets to support these research efforts.

7. Acknowledgements

All the work presented in this paper is supported by the National Institute of Standards and Technology, Information Technology Laboratory, Information Access Division.

[7]https://sites.google.com/view/dvuchallenge2021/home/

[8]https://sites.google.com/view/dvu-asia-challenge-2021

[9]https://ir.nist.gov/Landgirls.Challenge/landgirls.html

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.

8. Bibliographical References

Anand, V., Ramesh, R., Wang, Z., Feng, Y., Feng, J., Lyu, W., Zhu, T., Yuan, S., and Lin, C.-Y., (2020). *Story Semantic Relationships from Multimodal Cognitions*, page 4650–4654. Association for Computing Machinery, New York, NY, USA.

Anand, V., Ramesh, R., Jin, B., Wang, Z., Lei, X., and Lin, C.-Y., (2021). *MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding*. Association for Computing Machinery, New York, NY, USA.

Baumgartner, M., Rossetto, L., and Bernstein, A., (2020). *Towards Using Semantic-Web Technologies for Multi-Modal Knowledge Graph Construction*, page 4645–4649. Association for Computing Machinery, New York, NY, USA.

Creative Commons. (2019). About the licenses. `https://creativecommons.org/licenses/`, Last accessed on 2019-11-06.

Curtis, K., Awad, G., Rajput, S., and Soboroff, I. (2020a). Hlvu: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 355–361.

Curtis, K., Awad, G., Rajput, S., and Soboroff, I. (2020b). International workshop on deep video understanding. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 871–873.

Debattista, J., Salim, F. A., Haider, F., Conran, C., Conlan, O., Curtis, K., Wei, W., Junior, A. C., and O'Sullivan, D. (2018). Expressing multimedia content using semantics—a vision. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 302–303. IEEE.

Lei, J., Yu, L., Berg, T. L., and Bansal, M. (2020). Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Vicol, P., Tapaswi, M., Castrejon, L., and Fidler, S. (2018). Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8581–8590.

Yu, F., Wang, D., Zhang, B., and Ren, T., (2020). *Deep Relationship Analysis in Video with Multimodal Feature Fusion*, page 4640–4644. Association for Computing Machinery, New York, NY, USA.

Zhang, B., Yu, F., Fang, Y., Ren, T., and Wu, G. (2021a). Hybrid improvements in multimodal analysis for deep video understanding.

Zhang, B., Yu, F., Gao, Y., Ren, T., and Wu, G., (2021b). *Joint Learning for Relationship and Interaction Analysis in Video with Multimodal Feature Fusion*, page 4848–4852. Association for Computing Machinery, New York, NY, USA.