Flexible and Scalable Annotation Tool to Develop Scene Understanding Datasets

Md Fazle Elahi
melahi@iu.edu
Indiana University Purdue University Indianapolis
Indianapolis, IN, USA

Renran Tian
rtian@iupui.edu
Indiana University Purdue University Indianapolis
Indianapolis, IN, USA

Xiao Luo
luo25@iupui.edu
Indiana University Purdue University Indianapolis
Indianapolis, IN, USA

ABSTRACT

Recent progress in data-driven vision and language-based tasks demands developing training datasets enriched with multiple modalities representing human intelligence. The link between text and image data is one of the crucial modalities for developing AI models. The development process of such datasets in the video domain requires much effort from researchers and annotators (experts and non-experts). Researchers re-design annotation tools to extract knowledge from annotators to answer new research questions. The whole process repeats for each new question which is time-consuming. However, since the last decade, there has been little change in how the researchers and annotators interact with the annotation process. We revisit the annotation workflow and propose a concept of an adaptable and scalable annotation tool. The concept emphasizes its users' interactivity to make annotation process design seamless and efficient. Researchers can conveniently add newer modalities to or augment the extant datasets using the tool. The annotators can efficiently link free-form text to image objects. For conducting human-subject experiments on any scale, the tool supports the data collection for attaining group ground truth. We have conducted a case study using a prototype tool between two groups with the participation of 74 non-expert people. We find that the interactive linking of free-form text to image objects feels intuitive and evokes a thought process resulting in a high-quality annotation. The new design shows ≈ 35% improvement in the data annotation quality. On UX evaluation, we receive above-average positive feedback from 25 people regarding convenience, UI assistance, usability, and satisfaction.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; • **Human-centered computing** → **Graphical user interfaces**; • **Information systems** → Users and interactive retrieval.

KEYWORDS

Vision and Language, Scene Understanding, Data Annotation

HILDA '22 , June 12, 2022, Philadelphia, PA, USA

ACM ISBN 978-1-4503-9442-0/22/06…$15.00
https://doi.org/10.1145/3546930.3547499

ACM Reference Format:
Md Fazle Elahi, Renran Tian, and Xiao Luo. 2022. Flexible and Scalable Annotation Tool to Develop Scene Understanding Datasets. In *Workshop on Human-In-the-Loop Data Analytics (HILDA '22), June 12, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3546930.3547499

1 INTRODUCTION

Scene understanding is one of the fundamental challenges in computer vision, which is vital for the machine to interact with other functional systems by processing visual information from the external world. A model's understanding of the scene can be measured by evaluating its ability to answer open-ended questions related to the scene [9, 22]. Answering visual questions may require successful object detection, modeling the relationship and the interaction among the objects, and reasoning from an external knowledge base absent in the scene [6, 23]. These visual tasks of answering questions can be ranked in the order of cognitive complexity. For example, problems like object detection, instance segmentation, and activity recognition fall under the recognition category with a lower level of complexity. On the other hand, cognitively more complex tasks like image and video captioning, coreference resolution [11, 13], visual question answering (VQA) [1], and visual commonsense reasoning [36] encompass solving both the recognition and the cognition sub-tasks. For the rest of the article, we refer to the former visual tasks as Recognition-Only (RO) tasks and the latter as Recognition+Cognition (RC) tasks.

In the past few decades, powerful computing hardware and deep learning models significantly advanced toward solving RO tasks. Especially in the field of object detection, the researchers [4, 19, 33, 34] have made significant progress with very low error margin. Obtaining a satisfactory result in any visual task is an iterative procedure. Each iteration either refines the model's architecture, enhances training data, or does both. Among these two aspects, developing large-scale image datasets requires countless hours of human effort. For example, to solve RO tasks like object detection and instance segmentation, the training data containing millions of images [7, 14, 17] are labeled with object localization instead of global association (bounding box containing a cat vs. a whole image containing a cat).

The substantial progress in RO tasks has shifted the research focus to developing AI models for solving RC tasks. Expectantly, solving RC tasks requires more modalities than the existing object detection datasets alone can offer. When the model's architecture stays unchanged, the performance of the RC models is driven by the size and quality of the training data.

A common practice for developing such multi-modal datasets for RC tasks is by crowdsourcing additional modality to existing object detection datasets. For instance, the object-detection datasets can be gradually enhanced by adding text descriptions to solve RC tasks like image captioning and VQA. Similar modalities can be added to video datasets to conduct other video-based RC tasks like video description generations. Among these practices, models developed on training data containing the link between visual region and the text-phrase has better performance than those developed with description-only data [20]. This link between text-phrase and visual object is an important modality and is referred as object-level-grounding (OLG) for the rest of the paper.

2 RELATED WORKS

Considering the importance of training data to support RC tasks, we investigate the existing image and video datasets and the common annotation practice with the following questions:

To what extent extant datasets are generalizable to solve new RC tasks?

Initially, AI models for solving RC tasks [1] are trained on data without OLG in texts. As a result, the models suffer from various performance issues like attending wrong region [18] (correct description but model focuses on incorrect visual region) and object hallucination (describing something absent in the image) [27]. OLG in training data is found to alleviate those issues and can boost the model's performance [20]. This success inspired development of datasets like visual genome [14] and Flickr30k [25].

Though there exists large image datasets with OLG, corresponding video datasets are rather limited. Most of the video datasets belong to limited number of domains like movie [21, 28, 29], TV series [15, 16, 26], gameplay and cartoon [12, 24] , surveillance footage [31], and cooking [5]. It is not surprising that mostly movies and TV series are chosen because of available modalities like subtitles, scripts, plots, synopsis, and Descriptive Video Service (DVS). The existing video datasets cannot be conveniently generalized to other domains in solving RC problems. For example, in [26] only persons' face and their referred pronouns are linked to frames of video. As a result, such datasets can be insufficient when the research question concerns some other activity or entities.

For creating new datasets, how challenging the data collection and annotation process is to the researchers?

Typically the large-scale datasets are developed from crowd-sourcing platforms like Amazon Mechanical Turk (MTurk). Generally, the researchers re-purpose the MTurk annotation interface to augment new datasets with additional modalities. In this approach, a complex task is decomposed into multiple subtasks. Though such task decomposition is intended for productivity, some human intelligence tasks like obtaining event descriptions and text to visual object mapping are more natural when completed sequentially. In the current practice, the event description and OLG are done in multiple stages in MTurk. Firstly, a group of MTurk workers adds event descriptions to the video. Secondly, another group identifies and draws bounding boxes around the entities and their mentions in the image via a guided approach. Finally, another group of MTurk workers further checks the annotation quality. While separating the task of description and annotation provides task simplification, it can be less natural to obtain a description of an event or scene that is continuous in time.

Are existing tools sufficient to support necessary data annotation process?

Many image annotation tools already exist to support the development of datasets for solving RO and RC tasks. Most existing video annotation tools mainly focus on interpolated object labeling across multiple frames [2, 3, 8, 10, 32, 35]. The object annotation supports labeling with various geometric shapes, e.g., rectangle, oval, polygon. Few tools, for instance, ViTBAT [3] and VIA [8] support behavior or activity annotation and event description with temporal reference in the video. CVAT is an advanced tool for labeling objects in video and succeeded previous annotation tools [32, 35]. Both iVAT [2] and CVAT [10] support object annotation in interpolated frames but lack the support for OLG between text descriptions and images and video frames. Besides these open-source annotation tools, some commercial annotation solutions exist. However, these closed-source tools offer limited features and modalities and are usually not customizable for different research purposes.

3 RESEARCH GAP

An increasing shift toward solving RC tasks requires changing how the text modalities in the videos are augmented with OLG. Naturally, existing video datasets with OLG cannot meet the demand of the dynamic nature of research questions requiring complex reasoning. The existing pipeline for developing these datasets can improve when the interaction of researchers and annotators (human-in-the-loop) with the annotation process is revamped considering the following criteria:

Control over data collection design: The pipeline should have an interface that enables the researchers to customize the data annotation process with minimal effort. The data collection interface can be designed to add custom modalities. They should be able to collect data from expert and non-expert people when required by the experiment design. The data collection can be conducted on public and private platforms without sacrificing scalability. Although MTurk offers an annotation interface, its crowd workers are almost anonymous and suitable for doing micro-tasks. Additionally, working with sensitive data and complex tasks requires a specific group of people. For instance, data needs to be collected from a group of medical professionals to develop an AI system capable of reading radiology reports. However, in crowd-sourced platforms, the presence of medical professionals is less obvious.

Extraction of human intelligence: The pipeline should offer an interface that can efficiently extract knowledge for the human intelligent task. In complex experiments, for example, asking annotators to explain an event in a video for a commonsense reasoning task, the data annotators can effectively share their intelligence if the annotation interface can support an environment where objects and their inter-relation in the text and image can be connected conveniently.

4 CONCEPT OF PROPOSED ANNOTATION TOOL

To address the limitations in the current practice of creating scene understanding datasets, we propose the following improvements

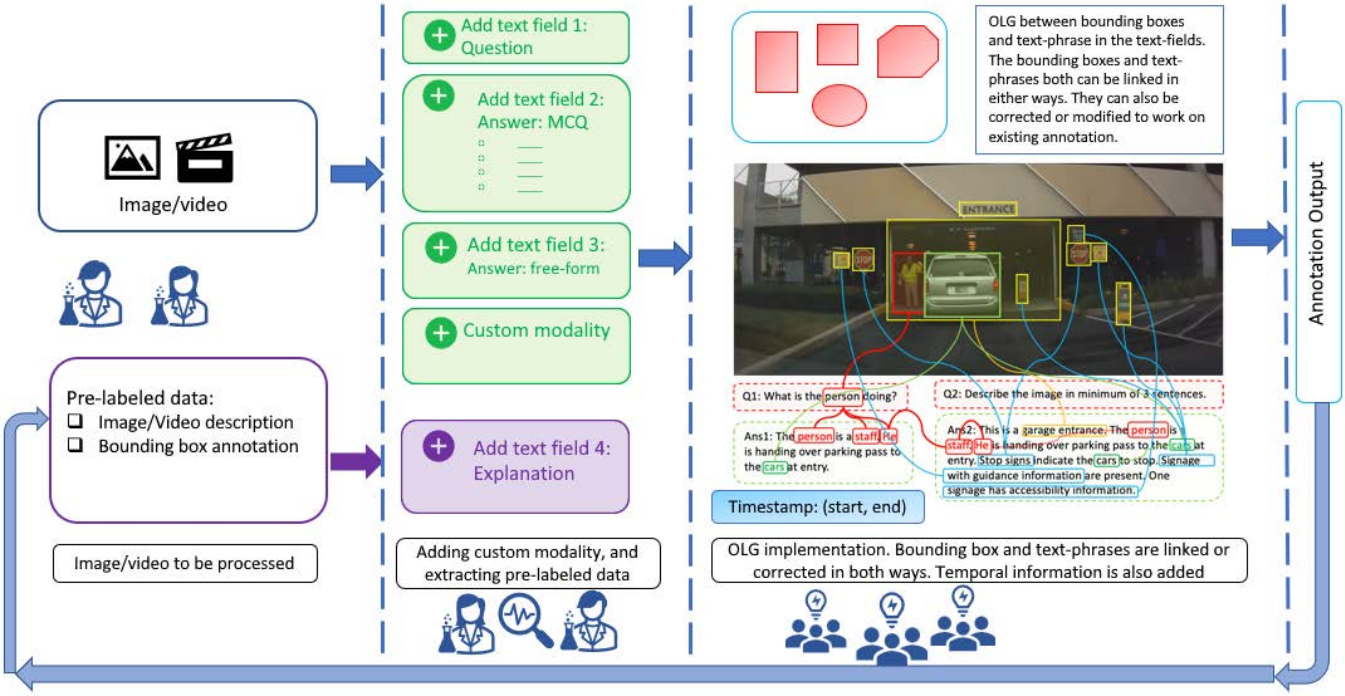

Figure 1: Conceptual workflow of processing of new and pre-labeled video and language data

in the design of the annotation process using our tool. The main contributions to the design improvement are:

- A GUI for researchers to customize Data Collection User Interface (DCUI) to label text, video, image data with spatio-temporal modalities.
- Support of OLG in the free-form description in both videos and images.
- Integration of novel interaction design for evoking human thinking while describing scenes via self-guided visual feedback (i.e., point at object in image, link, and explain them in text description.)

Fig. 1 shows an overview for annotating and processing the raw data and pre-existing annotations. In the workflow, the researchers and data annotators interact with the tool to transform the raw data into useful training data. The functionalities[1] of the tool in terms of interaction with humans are as follows:

Customizability: With this feature, researchers can exert more control over data collection in terms of DCUI design. They can design versatile research questions for the data by selecting existing or adding new modalities.

In the first stage, the researcher can select the raw data for annotation. Additionally, the tool will be capable of extracting the existing annotation linked to the input if the annotation belongs to the supported modality. The modalities include but are not limited to pre-labeled data like image or video description, subtitle, scripts, DVS, and audio. In the case of videos, temporal information of the annotation will be required for temporal alignment.

[1]demo video: https://youtu.be/h9hIYcOEuEE

Next, in stage 2, the researcher can add custom modality in the form of text or audio and define the annotation type. Ideally, for all the modalities, OLG is to be supported. Including OLG in the free-form text provides a more fluid experience for text-to-image linking. The connection of text in the image using bounding boxes provides self-guided visual feedback. Such visual feedback encourages the annotator to share more information. Section 4.1.2 follows more details on OLG implementation. In the case of adding text modality, researchers can define the question-answer (QA) format. Annotators can also fill in the QA if the research design requires it. Researchers can specify multiple formats of the answer, viz. multiple choice question (MCQ), preset check-boxes, and free-form. When the researcher finalizes the customization and deployment option, the tool generates the DCUI for collecting the data from the target people through the deployment platform like MTurk or a private server.

In stage 3, the main actors are the data annotators who label the data following the research design. The DCUI can support data correction, viz. relocation, resizing of bounding box across existing and new modalities. Additionally, the text phrase and its mentions can be linked to the bounding box and vice-versa. The DCUI can support various shapes of bounding boxes for visual region annotation. To mark the moments of interest in the video, the records of the start and end times of the moments can be saved.

Scalability: The design proposes the DCUI to be scalable and cross-platform from the deployment perspective. Such criteria ensure that people can remotely interact with it. Remote access increases the reach-ability of the DCUI to collect data from a diverse range of people. If necessary for the experiment, the DCUI can be distributed to

a special group of subjects by hosting on lab machines. The design also supports the deployment of DCUI in popular crowd-sourcing platforms like MTurk. which allows the DCUI to be accessed by non-experts. The DCUI can be conveniently used without further installation as a cross-platform and web-based app.

Compatibility: While exporting the annotation, the researchers can select the modalities of the augmented data. The export format of annotated data should be compatible and easy to parse. An evaluation tool for assessing the quality of the annotation can be integrated.

5 CASE STUDY AND PROTOTYPE OF ANNOTATION TOOL

A case study is conducted to evaluate the annotation tool's effectiveness in data collection following two criteria: (1) User attitudes towards the DCUI, and (2) the effect of integrating the OLG feature on the annotation quality. In the case study, a human subject experiment is done to study pedestrian interaction with vehicles. For this purpose, 204 15-second videos with interactions between designated pedestrians and the car driving on the road are sampled from TASI large-scale dataset [30].

A prototype DCUI based on the proposed annotation workflow was built. The DCUI's front-end is written in JavaScript, and the python flask framework supports its back-end. It was hosted on a private server in the lab. The study participants can remotely access the DCUI with an internet connection. 74 human subjects with driving experience in the USA were recruited randomly to complete the experiments. To qualitatively assess the UX performance of DCUI, a questionnaire with 11 questions (Table 1) about multiple usability types is used. This survey was optional; participants could ignore it if they wanted to.

The participants are divided into two groups to measure the effect of OLG feature on data annotation quality. One group with 50 participants used the DCUI with OLG feature to annotate 204 videos in total, and each was shown 45 videos. Keeping other conditions unchanged, the second group of 24 participants annotated from a subset of 110 videos using the DCUI with no OLG feature. Then the data quality of annotation is compared between two groups to evaluate the impact of OLG feature in UI design.

5.1 Overview of the experiment

5.1.1 Experimental Setting. After recruitment, each participant is trained by the researchers to navigate the DCUI and annotate the data through a one-one virtual meeting. For each video, at a particular timestamp, the participants are asked to answer two questions. The first question is, "would the target pedestrian cross in front of the car?" Its answer is designed in MCQ format with three options: yes, no, and "not sure." For the second question, the participant is asked to provide the explanation behind their answer to the first question. The explanation is collected in free-form text.

The participants are encouraged to follow the guidelines to think from five perspectives derived from the existing literature. However, they are free to provide any explanation outside those five categories if necessary. Fig 2 shows how the text modalities are used to obtain the answers in multiple formats for two pre-defined questions. Each free-form response is accompanied by OLG and video timestamp. At any paused moment in the video, the answer about pedestrians' crossing intention is subjective and may not be necessarily correct. However, the participant needs to justify their answer by sharing their reasoning and thought-process.

5.1.2 OLG in free-form text. To add OLG to the explanation, the participants connect any text phrase representing the entities (human, vehicles, traffic light, curb, crosswalk) they consider important to the visual region in the paused video. The participant highlights the text phrase and draws the bounding box around the entity in the paused video to link the text phrase. A unique field is created for each bounding box that can refer to multiple text phrases. If multiple text phrases refer to the same bounding box, participants can highlight another text phrase and click that field to make the connection. This way, drawing only one bounding box suffices for multiple text phrases or mentions. Fig 3 shows the implementation of linking multiple phrases to a single bounding box. Labeling this way is natural and more efficient and if required, can support datasets requiring OLG. After linking the text phrase and visual objects, the annotation is saved with the video's current timestamp. Thus, the OLG-supported explanation contains the accumulated information till that moment in the video.

After that moment, the participant keeps playing the video and observes the changes in the pedestrian's crossing intention. If there is a transition in the pedestrian's crossing intention, the participant is instructed to pause and answer the previous two questions for that moment. This explanation and answer contain the accumulated information between two pauses. The process is repeated till the participant is confident about the pedestrian's crossing intention. For each video, we collected answers and explanations from 9 such participants. A snapshot of the data from a single participant obtained using the tool is shown in Fig. 4.

5.1.3 Annotation Output. From Fig. 4, we see that the data can be used for solving a wide range of visual tasks like object-detection, VQA, VCR, visual coreference resolution, and referring expressions where OLG-supported description is required. By design, the output data also contains temporal information about the events. Additionally, the OLG-supported explanation accumulated from multiple participants can be used for ground truth which can also be used to analyze human behavior and thought processes.

5.2 Evaluation of the Prototype and Annotation

In this section, we highlight the effectiveness of the prototype annotation tool through qualitative and quantitative evaluation based on previously mentioned two criteria.

5.2.1 User Attitudes Evaluation. To assess the UX performance qualitatively, we used the voluntary responses from 25 participants collected within one week after their study completion. The 11 questions are grouped into four categories: satisfaction, usability, UI feedback, and convenience. The answers are in the Likert scale format in the range 1 through 5, while 1 being 'strongly disagree' and 5 being 'strongly agree.'

We present the result of the collected responses in multiple types of usability in Table 1. The scores are averaged across all the 25 participants for each question. For each category, 'categorical mean score' indicates the mean across average scores across all questions

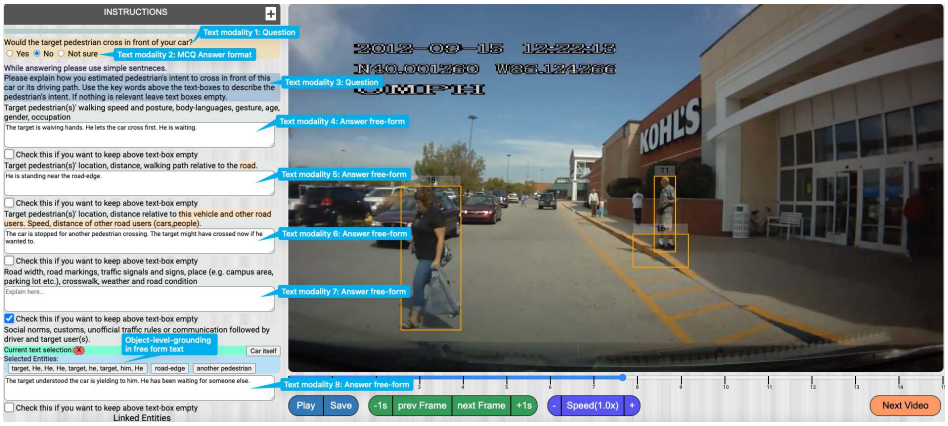

Figure 2: Additional modalities in text-form

Figure 3: Annotation of multiple text-phrases and mentions pointing to single bounding box

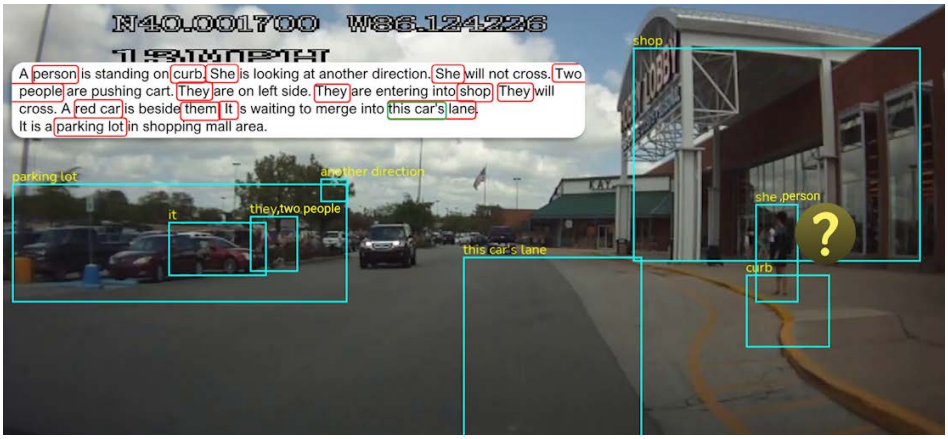

Figure 4: Output of annotated data mapped back to the video

belonging to the category. From the categorical mean scores, except in 'satisfaction,' all other categories have scored higher than 4.1, meaning the experience with the study and the prototype annotation tool is in the agree-able range. The 'satisfaction' category is 3.84, close to 4. The result suggests that there is room for improvement in the prototype design.

We also asked the participants about their suggestions to improve the user interface via an open-ended question: "Do you have any suggestions/comments that will help us make the system better?". Many of the participants mentioned several aspects of the UI to improve usability. For example, *"The data being erased after hitting save or scrubbing to a new part of the video should have a "are you sure" warning. Make a box to ignore in future, but I would have loved*

Table 1: Mean score of responses in Likert scale across 25 participants

ID	Question	Category	Mean (Std)	Category Mean (Std)
1	It was easy to learn to use this interface.	convenience	3.92 (1.17)	
2	The resources were effective in helping me complete the tasks and scenarios.	convenience	4.33 (0.88)	4.24 (0.28)
3	The training was effective in helping me complete the tasks and scenarios.	convenience	4.46 (1.04)	
4	Overall, I am satisfied with how easy it is to use this interface.	satisfaction	3.88 (1.19)	
5	I felt comfortable using this interface.	satisfaction	4.00 (1.17)	3.84 (0.19)
6	I liked using the interface of this system.	satisfaction	3.63 (0.81)	
7	The interface gave error messages that clearly told me how to fix problems.	UI feedback	3.96 (0.90)	4.13 (0.23)
8	The information (such as online help, on-screen messages, and other documentation) provided with this interface was clear.	UI feedback	4.29 (0.82)	
9	I was able to complete the tasks and scenarios quickly using this interface.	usability	4.12 (0.72)	
10	It was easy to find the information I needed.	usability	4.25 (1.09)	4.18 (0.07)
11	The organization of information on the interface screens was clear.	usability	4.17 (1.24)	

to go back and fix an error.", "It would be great if you could go back and see your annotation–because it disappears and you can't double check it". Some of them emphasized designing the DCUI to match the layout of popular social platforms. For instance, *"Maybe the interface would feel more natural if the text boxes were on the right side of the screen. It would feel better to navigate because I'm already used to that type of layout on sites like YouTube or Twitch. It's a small change but I think it would make UX better."*

5.2.2 Data Quality Evaluation. Following the experimental design, we have compared the quality of annotated data based on the volume of annotation collected from two groups of subjects. We use the average number of words disregarding punctuation and whitespace characters in annotation for each video as a comparison metric. The explanations with higher word counts are assumed to provide more information and better quality.

To compare between two groups, for each video, we calculate the average word counts across all the subjects working on that video. We have conducted a t-test for independent samples assuming unequal variance to examine the effect of the OLG feature on data annotation. We compare the average word counts across all the videos between the two groups. We find that there is a significant difference in average word counts between the group using OLG (M=77.37, SD=34.57) and the group not using OLG (M=57.08, SD=9.5); t(125)=5.93,p<.001, The result suggests that participants are likely to describe with ≈35% more words with OLG feature in UI.

Furthermore, manual inspection of the annotations reveals that with OLG, descriptions contain more references (pronouns) to the previously mentioned entities reducing repetitive mentions, e.g. "*The car is moving slowly. A woman is looking at the car. She is checking if it will yield to her.*" In the previous sentence, "*The car, car, it*" refers to the same car, and "*woman, She, her*" refers to the same person. In the descriptions without OLG, entities are described relative to another entity in the scene for localization information. For example, "*There is another pedestrian ahead. That pedestrian is on the lane the car just turned onto. That pedestrian is crossing from left to right. There are two more pedestrian standing on the sidewalk to the right.*" Such mentions are less prevalent in descriptions with OLG because entities can be linked in the image, and localization in the description is unnecessary. These differences indicate that OLG support in free-form texts makes the video annotation more natural and encourages the annotators to provide more information efficiently. Furthermore, the text-to-image linking also aids in providing more data about interaction among the objects.

6 FUTURE WORKS AND LIMITATION

The prototype annotation tool based on the proposed concepts can help the researchers augment existing data without extensively repurposing the DCUI. However, we did not discuss the assessment metric for the annotated data in the workflow. We believe these quality-check metrics are research question-oriented and can be challenging to generalize. However, they are critical regarding the development of a high-quality dataset. Nonetheless, if investigated thoroughly, that last step can also be generalized, and we include them in the future research endeavor. Additionally, the prototype design needs to be fully developed, and experiments need to be conducted to assess the efficacy of the overall workflow.

7 CONCLUSION

In this study, we present that the existing datasets in the video domain are limited and challenging to generalize for solving vision and NLP-based tasks with higher cognitive complexity. Additionally, existing annotation tools have limited support to add OLG modality from free-form description to both image and video data. To tackle this challenge, we propose a tool in the annotation workflow that enables researchers to tailor the DCUI more flexibly. Furthermore, we devise an interaction design that encourages annotators to provide better data through "indicate, connect, and explain" in both image and text modalities. Such design is conducive to providing data about objects of interest and their interaction which in turn aids in modeling human intelligence.

We conducted a case study with the participation of 74 subjects to show the feasibility of efficient data collection for solving a vison and NLP-based tasks using a prototype tool. Furthermore, we confirm the effectiveness of the UI design based on volunteered feedback collected from the participants. We also do a quantitative

analysis to measure the quality of collected annotation. We find the collected data has better quality and show how the UI design of the tool plays a vital role in extracting the thought process from human subjects.

ACKNOWLEDGMENTS

We thank Taryn Spisak and Morgan Stutzman for helping with the data collection. We thank the Toyota Collaborative Safety Research Center for funding support. This material is also based upon work supported by the National Science Foundation under Grant No.2145565

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 2425–2433. https://doi.org/10.1109/ICCV.2015.279

[2] Simone Bianco, Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2015. An Interactive Tool for Manual, Semi-Automatic and Automatic Video Annotation. *Computer Vision and Image Understanding* 131 (Feb. 2015), 88–99. https://doi.org/10.1016/j.cviu.2014.06.015

[3] Tewodros A. Biresaw, Tahir Nawaz, James Ferryman, and Anthony I. Dell. 2016. ViTBAT: Video Tracking and Behavior Annotation Tool. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, Colorado Springs, CO, USA, 295–301. https://doi.org/10.1109/AVSS.2016.7738055

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv:2004.10934 [cs, eess]* (April 2020). arXiv:2004.10934 [cs, eess] http://arxiv.org/abs/2004.10934

[5] Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Portland, OR, USA, 2634–2641. https://doi.org/10.1109/CVPR.2013.340

[6] Ernest Davis and Gary Marcus. 2015. Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Commun. ACM* 58, 9 (Aug. 2015), 92–103. https://doi.org/10.1145/2701413

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, Nice France, 2276–2279. https://doi.org/10.1145/3343031.3350535

[9] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. 2015. Visual Turing Test for Computer Vision Systems. *Proc. Natl. Acad. Sci. U.S.A.* 112, 12 (March 2015), 3618–3623. https://doi.org/10.1073/pnas.1422953112

[10] Intel. 2022. Computer Vision Annotation Tool (CVAT). https://github.com/openvinotoolkit/cvat

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 787–798. https://doi.org/10.3115/v1/D14-1086

[12] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. DeepStory: Video Story QA by Deep Embedded Memory Networks. *arXiv:1707.00836 [cs]* (July 2017). arXiv:1707.00836 [cs] http://arxiv.org/abs/1707.00836

[13] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What Are You Talking About? Text-to-Image Coreference. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 3558–3565. https://doi.org/10.1109/CVPR.2014.455

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int J Comput Vis* 123, 1 (May 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. TVQA: Localized, Compositional Video Question Answering. *arXiv:1809.01696 [cs]* (May 2019). arXiv:1809.01696 [cs] http://arxiv.org/abs/1809.01696

[16] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *arXiv:1904.11574 [cs]* (May 2020). arXiv:1904.11574 [cs] http://arxiv.org/abs/1904.11574

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. (Sept. 2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[18] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention Correctness in Neural Image Captioning. (2017), 7.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[20] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural Baby Talk. (2018). https://doi.org/10.48550/ARXIV.1803.09845

[21] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A Dataset and Exploration of Models for Understanding Video Data through Fill-in-the-Blank Question-Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 7359–7368. https://doi.org/10.1109/CVPR.2017.778

[22] Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes Based on Uncertain Input. In *Advances in Neural Information Processing Systems*, Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/hash/d516b13671a4179d9b7b458a6ebdeb92-Abstract.html

[23] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 3190–3199. https://doi.org/10.1109/CVPR.2019.00331

[24] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. MarioQA: Answering Questions by Watching Gameplay Videos. (2017), 9.

[25] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *arXiv:1505.04870 [cs]* (Sept. 2016). arXiv:1505.04870 [cs] http://arxiv.org/abs/1505.04870

[26] Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking People in Videos with "Their" Names Using Coreference Resolution. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Vol. 8689. Springer International Publishing, Cham, 95–110. https://doi.org/10.1007/978-3-319-10590-1_7

[27] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object Hallucination in Image Captioning. *arXiv:1809.02156 [cs]* (March 2019). arXiv:1809.02156 [cs] http://arxiv.org/abs/1809.02156

[28] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 3202–3212. https://doi.org/10.1109/CVPR.2015.7298940

[29] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 4631–4640. https://doi.org/10.1109/CVPR.2016.501

[30] Renran Tian, Lingxi Li, Kai Yang, Stanley Chien, Yaobin Chen, and Rini Sherony. 2014. Estimation of the Vehicle-Pedestrian Encounter/Conflict Risk on the Road Based on TASI 110-Car Naturalistic Driving Data Collection. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*. 623–629. https://doi.org/10.1109/IVS.2014.6856599

[31] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. 2014. Joint Video and Text Parsing for Understanding Events and Answering Queries. *IEEE MultiMedia* 21, 2 (April 2014), 42–70. https://doi.org/10.1109/MMUL.2014.29

[32] Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently Scaling up Crowdsourced Video Annotation: A Set of Best Practices for High Quality, Economical Video Labeling. *Int J Comput Vis* 101, 1 (Jan. 2013), 184–204. https://doi.org/10.1007/s11263-012-0564-1

[33] Chien-Yao Wang, I.-Hau Yeh, and Hong-Yuan Mark Liao. 2021. You Only Learn One Representation: Unified Network for Multiple Tasks. *arXiv:2105.04206 [cs]* (May 2021). arXiv:2105.04206 [cs] http://arxiv.org/abs/2105.04206

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2

[35] Jenny Yuen, Bryan Russell, Ce Liu, and Antonio Torralba. 2009. LabelMe Video: Building a Video Database with Human Annotations. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, Kyoto, 1451–1458. https://doi.org/10.1109/ICCV.2009.5459289

[36] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 6713–6724. https://doi.org/10.1109/CVPR.2019.00688