Multimodal Data Support in Knowledge Objects for Real-time Knowledge Sharing

Christine Kwon[1], John Stamper[1], James King[2], Joanie Lam[1], and John Carney[2]

[1] *Carnegie Mellon University, Pittsburgh, PA, USA*
[2] *MARi, Inc., Alexandria, VA, USA*

Abstract

Knowledge sharing can be made through various methods to complete a desired task, most commonly through structured manuals and more recently, through modern technology platforms. such as videos or interactive demos. Past research on knowledge sharing has focused on extracting multiple features from various sources surrounding a specific task, creating relevant tags to increase accessibility and searchability in a task. To enhance the process of knowledge sharing, we propose an advanced ontological knowledge structure which we denote as a Knowledge Object (KO) defined around a particular task of action. We present the general definition of a KO as a structural knowledge component that consists of metadata linked to multiple streams of data in examining a specific task. Furthermore, we are developing a search repository for KOs in which we implement a Human/AI approach in providing informative tags from the different data streams in each KO. Our work uses visualization tools to identify key objects and actions within a video of a certain task in combination with action recognition and object tracking modeling within Vertex AI to predict relevant tags for differing tasks.

Keywords

Knowledge Object, Video Tagging, Text Tagging

1. Introduction

The completion of complex tasks is most often broken down into a series of steps for the purpose of learning and training. It is common for these tasks to be presented to a learner in the form of a manual, which remains the most common way of disseminating information even with the improvements in technology related to multimedia learning. More recently, technology has made the recording and distribution of videos a common and viable form of knowledge transfer for complex tasks. YouTube, TikTok, and other platforms regularly share how to complete tasks such as making recipes, automotive care, home maintenance, or even complex electronics repair or computer programming. While this type of knowledge sharing has taken off among individuals, larger corporate/government organizations have yet to take maximum advantage of these modern media platforms in order to let employees quickly and effectively share expertise. We have identified several reasons for the lack of enthusiasm in large organizations including 1) the ability to validate user generated content, 2) the ability to organize the data in an easily searchable format, and 3) the ability to curate the content when changes to the underlying equipment or processes occur. In this work, we present the development of Knowledge Objects or KOs to provide ontological support for complex tasks and provide solutions to the adoption of modern technology and techniques for the purpose of real-time knowledge sharing. The proposed KOs allow for multiple streams of data to be combined effectively for the purpose of validating each data stream, organizing and generating metadata for the KO, and allowing for efficient curation when changes occur.

Proceedings of the 6th CROSSMMLA workshop "Leveraging Multimodal Data for Generating Meaningful Feedback". At the 13th International Learning Analytics & Knowledge (LAK'23). Arlington, Texas, United States.

2. The Knowledge Object

We define a KO around a specific task that is composed of one or more steps. There are several reasons why the idea of defining KOs is important for task based learning. The primary reason is to easily organize tasks for training purposes. A secondary and equally important reason is the ability to search for training and examples on demand. Scenarios where KOs excel are mechanical tasks like automotive repair (changing oil, spark plugs, tire, etc.) where finding the task currently is quite easy, but a specific instance for a particular vehicle may be much harder to find and may have details specific to that instance. Having the right metadata in the right structure is critical, especially where the metadata links multiple datastreams. It is possible to define a hierarchical structure to a set of KOs where parent-child relationships exist at multiple levels although for the purpose of broad search we are most interested in the higher level KOs.

The closest example of the meta structure for KOs can be seen in the various schemas stored by schema.org. The underlying schemas used in creative work as the base and other works act as a guide to our structure [3]. Youtube also uses the schema for creative work as the base metadata form of their videos in addition to the Clip schema [4]. The majority of our datastreams contain both text data in the form of manuals and video data which also can include transcript data. In addition, learner data may be present in the form of log data. Because of the multimodal structure of our data we add a necessary addition to the meta-data structure in the form of "anchors" that connect the data streams. Traditionally in most multimodal works, the anchors are timestamps that link various temporal data streams, for example, video and audio [10], or video and log data [5]. At the highest level of a KO, however, the anchors relate to the steps to accomplish a task to a data stream and therefore do not need to have a strictly temporal element. In the case of a manual the anchors might be page number plus paragraph, for example.

3. Background and Related Work

Previous research discusses the comparison between the usefulness and effectiveness of using video in learning versus using and reading text in learning. This specific research discusses how German secondary students interact with videos versus illustrated text to learn fairly complex content [8]. Non-interactive videos were weighted equivalent to the usage of standard text material and a significant improvement in learning was seen through interactive videos that contain micro-level features, which allowed students more control over their learning. In our context, a KO can include content in both video and text in its metastructure, but it is important to distinguish the difference between each form in order to take advantage of the unique features of each form of content.

Relying on log data to yield important insights in learning analytics research, specifically in open-ended, collaborative, or project-based learning environments can lead to significant misunderstandings in the data[7]. Multimodal data streaming can eradicate these misunderstandings by collecting, analyzing, and processing multiple types of data, however, challenges in temporal synchronization between multiple streams of data must be overcome. Tools have been created for this purpose[6]. Our proposed KOs are defined through a metastructure of tasks and information, allowing multiple streams of data to be combined to create this metastructure. This requires us to analyze these streams of data that a KO is composed of. Hence, we plan on using similar data alignment tools such as the STREAMS[5] tools and Datavyu[2] to further analyze KOs and create related tagging algorithms to search for a KO.

There is a general consensus amongst work in image and video tagging research that manual efforts in labeling and annotating data is tedious and prone to error. Hence, there is extensive work in automating this tagging process through diverse methods. One example is an interactive application for trainers, coaches, and players to receive multi-modal feedback through detection of events of interest[10]. Specifically in this research, volleyball players wore wireless sensors on their wrists in which features were extracted from these sensors to train a machine learning model that detects volleyball actions and objects within a training session or a match. These actions and information on what time they occur in the video, along with more detailed information about the videos and players, are stored and indexed in a unique repository which is directly referred to by the interactive web application. Once a user filters through actions by player or type of action, the web application will

automatically direct the user to the timestamp of the event of the action. Our work proposes a similar search repository of KOs, creating a collection of knowledge objects in which users can search through by a keyword to find a KO of their interests or search an element of interest within a KO metastructure.

Existing information retrieval sources, such as StackOverflow, use informative tags for users to assess the relevance of a post to either use that post or to find related posts[9]. Inspired by these retrieval sources, there has been research conducted on creating similar tagging approaches on videos. Due to diverse levels of expertise in software developers, gauging the relevance of usefulness of a video may be difficult. Thus, there is a high probability that users have missed useful tutorials and wasted time in watching extraneous videos. Exploring tagging algorithms and automated approaches that generate tags associated with software development video tutorials has shown that some information retrieval-based approaches were the most effective and successful in recommending developers with relevant tags for software video tutorials[9]. In relation to our research, these software development video tutorials are analogous to the defined KOs. Our research plans to use a similar tagging method that Parra et al discusses, analyzing the content of KO and KO structures to predict tags that can be searched in a KO search repository.

Further research has been conducted on how users with varying levels of mastery in a skill may view a video related to that skill in contrasting manners. For instance, Yonezawa et al discusses how recipe short videos allow people to learn how to cook in a short amount of time, but also its difficulty in mastering a recipe by watching a video only once[11]. Hence, they propose a unique method named Dynamic Cloud Tagging, which is a method that extracts cooking operations from text recipes attached to the recipe short videos and supplementary recipe information based on users' cooking levels by weighting the appearance frequencies of certain cooking operations. These short cooking video tutorials and its related information are similar to what we would include in a KO. The proposed KOs allow for multiple streams of data to generate a metastructure for a KO. Hence, we will observe and analyze these multiple streams of data, which may include generated text content and transcripts attached to a KO metastructure.

More research has been conducted on other habitual functions, such as watching television and content from other media sources. For instance, Hölbling et al addresses the need for an organized system that efficiently finds content of interest amongst the rising amount of available content in digital television[4]. Hölbling et al. presents a self-adapting and personalized recommendation system, combining filtering methods from TV guides and folksonomies, that addresses and tackles three main issues: (1) since new programs are constantly produced, there is significant lack of information and tag information related to their content. Using this content-based filtering method can infer descriptive information about these new programs and use them in automated recommendations[4]. (2) There are many instances where the same tags may have various connotations for different users[4]. The proposed filtering system uses local tagging histories of users and suggests tags for newer programs. (3) Some users may refuse sharing their tagging and search histories, so this filtering system allows users to decide on how much information they can make accessible while still producing tags and recommendations[4]. Our KO repository also aims to create a custom search and recommendation system that directs consumers to tutorials that are relevant to their needs and interests. We plan to expand upon this content-filtering system in constructing informative tags for tutorials and videos, building a tagging algorithm that tags specific manual steps, actions, and objects within the content for a more in-depth and quick search.

4. Human/AI Teaming Approach for Joining Multimodal Data Streams

For real-time knowledge sharing to be feasible within an organization, consumers of instructional content need to be able to efficiently find tutorials relevant to their needs. This will require the KOs to be tagged for relevant information. We are implementing a Human/AI teaming approach to properly tag the data streams for the KOs. There are two specific levels of search needed in our KO repository:

1) The first step of serving a consumer relevant search results is across-KO search; all KOs in the repository will be compared against the query in a standard search procedure to identify the most relevant KO.

2) The second step of identifying and serving relevant results is within-KO search; within the KO identified as most relevant, the task step most relevant to the query will be identified.

To accomplish the tagging needed for KO search within the RKS repository, we first use AI to create the initial tagging followed by human augmentation of the results for confirmation and error checking.

Automatically tagging instructional videos with labels describing task steps within the video will enable us to efficiently identify the step most relevant to a given query. Using Google Cloud Platform's Vertex AI, our team developed several machine learning models that automatically detect actions and objects within a given video using an approach detailed in Section 5.

Videos will be automatically tagged on import by custom action recognition and object tracking models trained in Vertex AI as well as a pre-trained speech-to-text model. These initial auto-generated tags will be used to enhance KO discoverability through search and to segment video KOs into moments.

In order to provide support for the creation and curation of KOs, we will utilize (or create) a number of tools for visualizing streams. These tools will be used during the creation and import of KOs as well as in the curation of our RKS system as users interact with KOs.

For the purpose of identifying key events and objects within a datastream, we have found a tagging tool called Datavyu [2], that allows for various streams of multimodal data to be visualized and tagged. In Fig 1, you can see a human user tagging a video of a person making a recipe. A key strength of Datavyu is the ability to load and synchronize multiple streams of data in multiple modalities. This means we can sync video, audio, and text. For the example shown in Fig 1, we can sync up written recipe steps with their location within a video of the recipe being executed. These visualization tools augment human input with our AI generated synching.

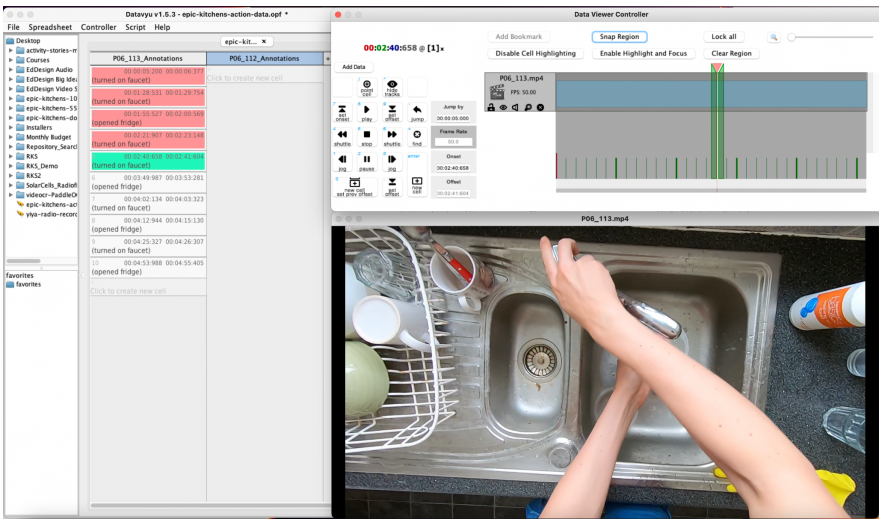

Figure 1: Example of identifying turning on a faucet in the Epic Kitchen's videos as a knowledge object for completing a recipe.

As a KO is used, continuous analysis occurs and additional tags are appended to the KO. For example, as users comment on the KO, NLP will detect sections of the video in which comments suggest a potential problem (See Fig 2, comment flags are shown in the orange in video). This allows for quick human curation of any suspected issues within a KO. Explicitly reported problems or official changes to how a procedure should be carried out also result in tags appended to moments in the video KO (See Fig 2, the red and blue bars in the above video, represent reported problems and flagged areas respectively). The combination of up-front KO tagging on import and continuous tagging enables automated KO repository maintenance.

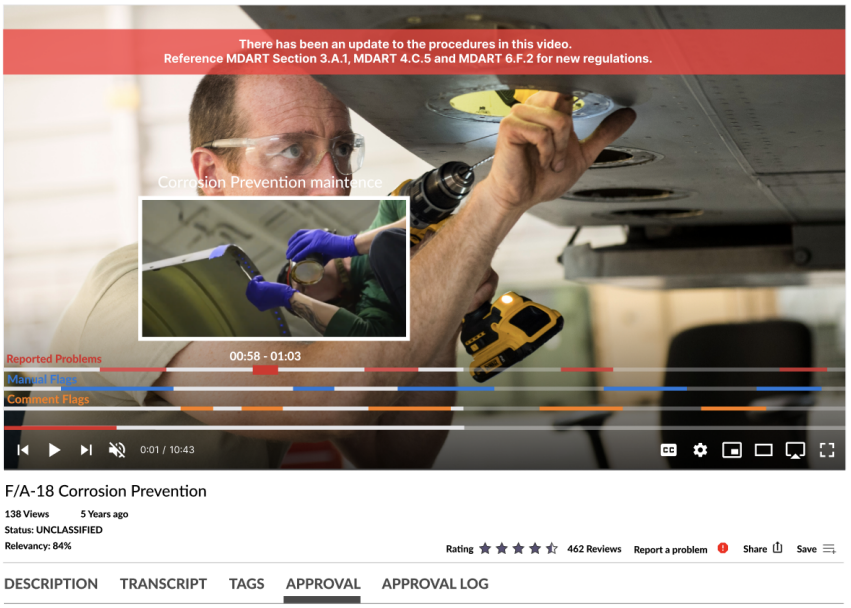

Figure 2: Example of a video portion of a KO that has user flags identifying potential problems and comments that users have generated.

5. Real-Time Knowledge Sharing (RKS) Overview

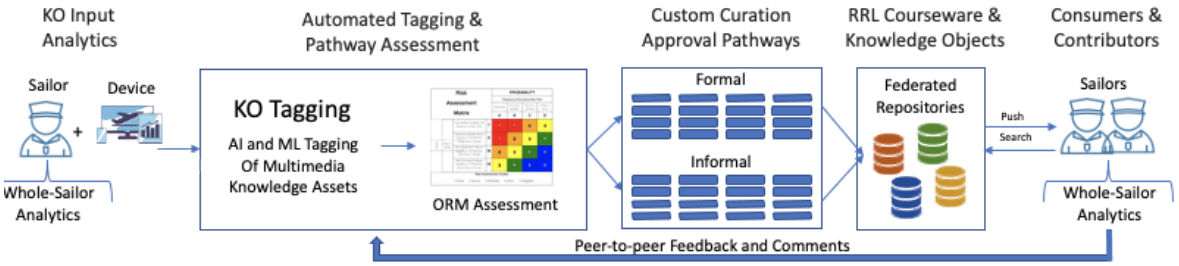

Figure 3: Diagram that represents the overarching functionality of RKS, which demonstrates how videos relevant to the consumers' needs are distributed to consumers in the context of the Navy.

Figure 3 demonstrates how our KO repository, Real-Time Knowledge Sharing (RKS), will distribute tutorials to the consumers relevant to their needs and interests. Uploaded content, in the form of a KO, will be processed through an initial automated tagging and pathway assessment, in which the corresponding content will be tagged with informative tags on the actions and objects that occur within the video through custom curated Machine Learning models which specialize in action recognition and object tracking. After this preliminary stage, the related KO is required to proceed to an approval stage, which is typically mandated by most government organizations. The KO will be thoroughly evaluated in conjunction with the associated official manual and can be further validated by designated approvers within an organization. Once the KO passes this stage, it will be appended to the KO repository, ready for distribution to the consumers and other stakeholders.

The initial addition of a KO to the KO repository does not indicate that it is fixed and insusceptible to change. As previously discussed, there are multiple circumstances in which a KO must undergo major revisions, which include 1) official updates or changes within the correlated official manual 2) potential negative comments from consumers 3) and overall reported problems from consumers and other stakeholders. If any of these concerns arise, the associated KO will be processed through the automated tagging and pathway assessment once more, where new tags may be included in the KO

and its existing tags will be revised to reflect these changes, updates, and concerns. Once the KO is amended in the automated tagging and pathway assessment, it will cycle through the same initial procedure to be reinstated in the KO repository.

6. Initial Implementation and Discussion

Inspired by success in auto-labeling videos in the Epic Kitchens project [1], our team leveraged the Epic Kitchens dataset for model prototyping in Vertex AI. Our team's action recognition model was trained to identify moments in which a refrigerator is opened and a sink faucet is turned on. With 21 training videos and 5 test videos sourced from a diverse set of kitchens within the Epic Kitchens dataset, our model has 91.7% Recall, 58.3% Precision, and 0.802 Average Precision when the confidence threshold is set to .98 and the precision length window is 3 seconds.

Similarly, our team trained an object tracking model to identify and track refrigerators and sinks within a given Epic Kitchens video. Leveraging 25 training videos and 5 test videos from a diverse set of kitchens, our team initially achieved 6.6% Recall, 100% Precision, and 0.401 Average Precision with a confidence threshold of 0.6 and IoU threshold of 0.5. In an effort to improve the model's low Recall, our team established a stricter data labeling protocol and created a new dataset that contains videos from a single kitchen to ensure that the definitions of "sink" and "refrigerator" remain constant and the model is not confused by differences between sinks in diverse kitchens within the Epic Kitchens dataset. After training and testing the new object tracking model with only videos from a single participant, our Recall was 14.5%, Precision was 100%, and Average Precision was 0.318 when the confidence threshold was 0.6 and IoU threshold was 0.5.

Though there was a slight improvement in increase Recall, we realize that there needs to be significant refinements made to significantly increase this metric to advance the accuracy of our object tracking model. Our team continues to refine labeling practices and expand our training datasets to increase the accuracy of our object tracking and action detection models.

7. Conclusion and Future Work

This research represents an initial work towards aligning text based manuals with video based learning generated by different creators. We have proposed a data model built on the concept of a knowledge object that describes a high level task and can provide the data structures to align different modalities of learning content. Our approach utilizes NLP methods to extract the steps towards task completion from a manual and locates and aligns the steps with videos of the tasks. Utilizing object and action detection we want to quickly identify the manual steps for the purpose of allowing searching within a KO. We continue to improve our models testing on various tasks and plan to create a repository of KOs for others to use in the near future.

8. Acknowledgements

This work was supported by US Navy STTR Phase I Contract #N68335-21-C-0438.

9. References

[1] Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., ... & Wray, M. (2020). The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 43(11), 4125-4141.

[2] Datavyu Team (2014). Datavyu: A Video Coding Tool. Databrary Project, New York University. URL http://datavyu.org.

[3] Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2), 44-51.

[4] Hölbling, G., Thalhammer, A., & Kosch, H. (2010, June). Content-based tag generation to enable a tag-based collaborative tv-recommendation system. In *Proceedings of the 8th European Conference on Interactive TV and Video* (pp. 273-282).
[5] Liu, R., Stamper, J., Davenport, J., Crossley, S., McNamara, D., Nzinga, K., & Sherin, B. (2019). Learning linkages: Integrating data streams of multiple modalities and timescales. *Journal of Computer Assisted Learning*, 35(1), 99-109.
[6] Liu, R., Stamper, J. C., & Davenport, J. (2018). A novel method for the in-depth multimodal analysis of student learning trajectories in intelligent tutoring systems. *Journal of Learning Analytics*, 5(1), 41-54.
[7] Liu, Ran, and Stamper, John C (2017). Multimodal Data Collection and Analysis of Collaborative Learning through an Intelligent Tutoring System. In *MMLA-CrossLAK@ LAK*. 47-52.
[8] Merkt, M., Weigand, S., Heier, A., & Schwan, S. (2011). Learning with videos vs. learning with print: The role of interactive features. *Learning and Instruction*, 21(6), 687-704.
[9] Parra, E., Escobar-Avila, J., & Haiduc, S. (2018, May). Automatic tag recommendation for software development video tutorials. In *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)* (pp. 222-22210). IEEE.
[10] Salim, F., Haider, F., Tasdemir, S. B. Y., Naghashi, V., Tengiz, I., Cengiz, K., ... & van Beijnum, B. J. (2019, October). A searching and automatic video tagging tool for events of interest during volleyball training sessions. In *2019 International Conference on Multimodal Interaction* (pp. 501-503).
[11] Yonezawa, T., Wang, Y., Kawai, Y., & Sumiya, K. (2020, March). Dynamic video tag cloud: A cooking support system for recipe short videos. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (pp. 122-123).