Compressing Multi-Modal Temporal Knowledge Graphs of Videos

Shusaku Egami[1], Takanori Ugai[2,1] and Ken Fukuda[1,*]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
[2]Fujitsu Limited, Kanagawa, Japan

Abstract

The construction of multi-modal temporal knowledge graphs (MMTKGs) that ground non-symbolic and time-series data, such as videos, into entities in the graph is still in the early stages. Hence, there is a lack of discussion about compressing and publishing MMTKG with huge data size. In this paper, we propose compression methods for MMTKGs of videos based on splitting images and inference rules and conduct experiments to evaluate their performance. As a result, our methods reduced the size of the MMTKGs by 27.7-36.1%. This study contributes to reducing the cost of distributing large MMTKGs on the web.

Keywords

Multi-Modal Knowledge Graph, RDF Compression, Video Dataset, Temporal Knowledge Graph

1. Introduction

Multi-modal knowledge graphs (MMKGs) [1, 2], which ground non-symbolic data into symbolic entities, have attracted attention as datasets for semantic and conceptual processing across modalities. However, constructing and publishing multi-modal temporal knowledge graphs (MMTKG) that ground multi-modal and time-series data, such as videos, into entities in the graph is still in the early stages.

Typical MMKGs describe multi-modal contents by URLs or file paths. This approach may not be suitable for the permanent publication of MMKGs as the multi-modal contents may become inaccessible due to broken links. This issue could potentially be resolved by encoding the file's binary data as an entity in the KG [3, 4]. However, building an MMTKG that describes the content of a video in fine-grained time intervals, such as in seconds or video frames, would result in huge data size, making it expensive to publish and share.

We proposed methods compressing MMTKGs of videos and conducted experiments to determine their effectiveness. We focused on two types of MMTKGs: KGs with video frame images encoded in Base64 and KGs with entire video files encoded in Base64. The proposed methods include differential compression based on knowledge representation of splitting video frame images and reduction of redundant triples based on inference rules. The results demonstrated that our compression methods reduced the size of the MMTKGs by 27.7-36.1%. This study contributes to reducing the cost of distributing large MMTKGs on the web.

Posters, Demos, and Industry Tracks at ISWC 2024, November 13-15, 2024, Baltimore, USA
*Corresponding author.
s-egami@aist.go.jp (S. Egami); ugai@fujitsu.com (T. Ugai); ken.fukuda@aist.go.jp (K. Fukuda)

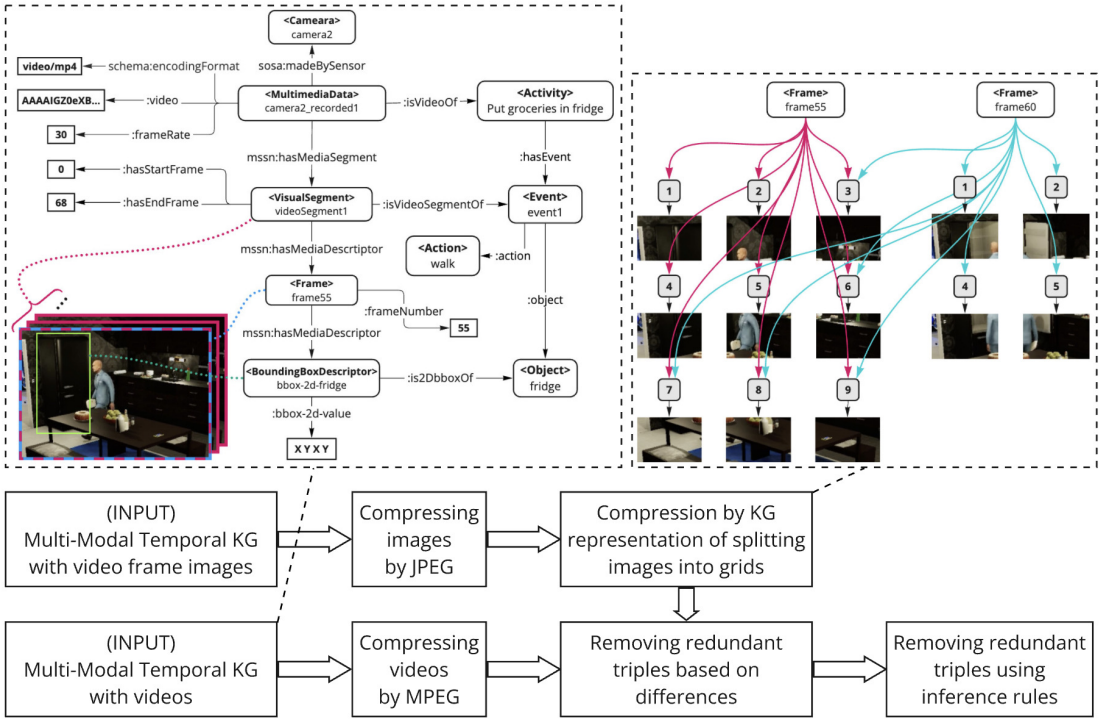

Figure 1: Overview of multi-modal temporal knowledge graph compression

2. Related Work

Zhu et al. [1] and Chen et al. [2] comprehensively surveyed and summarized works on MMKGs. Typical multimodal knowledge graphs are MMpedia [5] and IMGpedia [6], which ground images to entities in the graph. VisionKG [7] is an MMKG containing bounding boxes (bboxes) of objects extracted from various image datasets such as MS-COCO [8], CIFAR [9], and PASCAL VOC [10]. These MMKGs represent images by URIs or file paths. Studies on video KGs have evolved in the context of video indexing and retrieval [11, 12, 13]. VEKG [14] is an MMKG based on the extracted events from videos, bboxes, and image features. However, the data is not publicly available. There have been a lot of studies of compression methods for KGs [15]. However, MMKGs for videos are not covered.

3. Approach

MMKGs usually describe images and videos by URIs or file paths, which causes broken links to multi-modal files. Thus, we focus on permanently accessible MMTKGs that embed multi-modal files in a KG as an entity, and propose compression methods for these MMTKGs.

3.1. Data preparation

As an example, we constructed MMTKGs of indoor daily activities from multi-modal data of videos, text, and JSON output by VirtualHome-AIST[1] [16], as shown in the upper left of Figure 1. The multi-modal data was output every five frames. The dataset contains over 3,500 videos, which include both fixed camera views and third-person views of the camera moving. The average video length is 64.2 seconds, with a maximum of 268.9 seconds and a minimum of 12.5 seconds. We prepared two types of MMTKGs: a KG with every five video frame images encoded in Base64 described as literal values (i.e., image-embedded MMTKG), and a KG with videos encoded in Base64 described as literal values (i.e., video-embedded MMTKG). We reused the Multimedia Semantic Sensor Network (MSSN) ontology [17] and VirtualHome2KG [18] ontology for schema design.

3.2. MMTKG compression

3.2.1. Compressing image-embedded MMTKG

If the MMTKG contains video frame image data, each video frame image is first compressed as a JPEG. Next, each image is split into a grid. The grid image is encoded in Base64 and described in the knowledge representation as shown in the upper right of Figure 1. Here, if there is no difference between the grid image of the current frame and the grid image at the same position in the previous frame, the entity and the literal value of the current grid image are not created, and those of the previous frame are reused.

3.2.2. Compressing video-embedded MMTKG

We adopted MPEG-4 [19] to reduce the video data size. Each video frame entity has a frame number instead of having a Base64 value, and the video entity has a Base64 value for the compressed video. It is possible to extract arbitrary frame images from the video using FFmpeg [20]. The MMTKG size can be further reduced, but long videos take a longer time to decompress.

3.2.3. Removing redundant triples using inference rules

The MMTKGs have redundant triples if the 2D bboxes are not changed. We reduced the number of entities and triples by referring to the previous entities if the current 2D bboxes have not changed since the previous frame.

Moreover, inspired by the approach of removing triples that can be inferred from the rules [21], we create only the relation equivalentFrame(e_{pf}, e_{cf}) between previous frame entity e_{pf} and current frame entity e_{cf} when all 2D bboxes are not changed from the previous frame. We defined the rule as follows: hasMediaDescriptor(e_{pf}, e_{bbox}) $\wedge$ equivalentFrame(e_{cf}, e_{pf}) $\rightarrow$ hasMediaDescriptor(e_{cf}, e_{bbox}). Similarly, for grid images, we removed triples that can be inferred from the following rule: image(e_{pf}, e_{image}) $\wedge$ equivalentImage(e_{cf}, e_{pf}) $\rightarrow$ image(e_{cf}, e_{image}). Note that the image property here refers to a split image.

[1] https://github.com/aistairc/virtualhome_aist

Table 1
Image-embedded MMTKG

MMTKG	# of triples	Size [GB]
raw	134,945,485	62.0
3×3 grid	64,242,296	41.8 (-32.5%)
4×4 grid	78,384,156	39.6 (-36.1%)
5×5 grid	96,401,621	39.9 (-35.6%)

Table 2
Video-embedded MMTKG

MMTKG	# of triples	Size [GB]
raw	131,786,665	17.3
w/o redundant triples	37,646,681	12.5 (-27.7%)
w/o redundant triples and triples can be inferred	36,284,402	12.4 (-28.3%)

4. Result

Tables 1 and 2 show the results of the compression experiments. Our methods achieved data size reductions of 36.1% for image-embedded MMTKG and 28.3% for video-embedded MMTKG. There is a trade-off between the number of grid divisions and the number of triples. The best strategy is 4×4. In this study, we experimented with $n \times n$ grid divisions; however, experiments with $n \times m$ grid divisions are also necessary for a more detailed analysis. We published MMTKGs in a permanently accessible format.[2] In addition, tools for decoding and extracting images and videos from compressed MMTKG are available.[3]

5. Discussion

We proposed compression methods for two types of MMTKGs: image-embedded and video-embedded MMTKGs. The former MMTKGs can display arbitrary images on the web using HTML <img> tags without decoding the videos. The latter MMTKGs can apply video compression methods, and if the video is decoded, any frame can be extracted based on the frame number of the image. These MMTKGs can help create benchmark datasets for vision-language models since it is possible to extract arbitrary text and images using SPARQL queries [16]. The compression method for image-embedded MMTKGs might be effective for image stream data in which no video file is created. In contrast, the compression method for video-embedded MMTKGs is more effective when video files are available. Our compression methods for MMTKGs are effective for fixed-camera view videos but are less effective for first-person view videos.

6. Conclusion

We proposed compression methods for two types of permanently available MMTKGs in which video data are directly embedded as literal values. As a result, our methods achieved data size reductions of 36.1% for image-embedded MMTKG and 28.3% for video-embedded MMTKG. The two MMTKG datasets and the tools are available on GitHub. In the future, we will consider combining our methods with other RDF compression methods [22, 23].

Acknowledgments
This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), and JSPS KAKENHI Grant Number JP22K18008 and JP23H03688.

[2] https://github.com/aistairc/vhakg
[3] https://github.com/aistairc/vhakg-tools

References

[1] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, N. J. Yuan, Multi-Modal Knowledge Graph Construction and Application: A Survey, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 36 (2024).

[2] Z. Chen, Y. Zhang, Y. Fang, Y. Geng, L. Guo, X. Chen, Q. Li, W. Zhang, J. Chen, Y. Zhu, et al., Knowledge graphs meet multi-modal learning: A comprehensive survey, arXiv preprint arXiv:2402.05391 (2024).

[3] X. Wilcke, P. Bloem, V. De Boer, The knowledge graph as the default data model for learning on heterogeneous knowledge, Data Science 1 (2017) 39–57.

[4] P. Bloem, X. Wilcke, L. van Berkel, V. de Boer, kgbench: A collection of knowledge graph datasets for evaluating relational and multimodal machine learning, in: R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2021, pp. 614–630.

[5] Y. Wu, X. Wu, J. Li, Y. Zhang, H. Wang, W. Du, Z. He, J. Liu, T. Ruan, MMpedia: A Large-Scale Multi-modal Knowledge Graph, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 18–37. doi:10.1007/978-3-031-47243-5_2.

[6] S. Ferrada, B. Bustos, A. Hogan, IMGpedia: A Linked Dataset with Content-Based Analysis of Wikimedia Images, in: C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web – ISWC 2017, Springer International Publishing, Cham, 2017, pp. 84–93. doi:10.1007/978-3-319-68204-4_8.

[7] J. Yuan, A. Le-Tuan, M. Nguyen-Duc, T.-K. Tran, M. Hauswirth, D. Le-Phuoc, VisionKG: Unleashing the Power of Visual Datasets via Knowledge Graph, in: A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2024, pp. 75–93. doi:10.1007/978-3-031-60635-9_5.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[9] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes (VOC) Challenge, International Journal of Computer Vision 88 (2010) 303–338. doi:10.1007/s11263-009-0275-4.

[11] L. F. Sikos, Rdf-powered semantic video annotation tools with concept mapping to linked data for next-generation video indexing: a comprehensive review, Multimedia Tools and Applications 76 (2017) 14437–14460.

[12] K. Fukuda, J. Vizcarra, S. Nishimura, Massive semantic video annotation in high-end customer service, in: F. F.-H. Nah, K. Siau (Eds.), HCI in Business, Government and Organizations, Springer International Publishing, Cham, 2020, pp. 46–58.

[13] J. Vizcarra, S. Nishimura, K. Fukuda, Ontology-based human behavior indexing with multimodal video data, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), 2021, pp. 262–267. doi:10.1109/ICSC50631.2021.00052.

[14] P. Yadav, E. Curry, Vekg: Video event knowledge graph to represent video streams for complex event pattern matching, in: 2019 First International Conference on Graph Computing (GC), 2019, pp. 13–20. doi:10.1109/GC46384.2019.00011.

[15] M. Besta, T. Hoefler, Survey and taxonomy of lossless graph compression and space-efficient graph representations, 2019. arXiv:1806.01799.

[16] S. Egami, T. Ugai, S. N. N. Htun, K. Fukuda, VHAKG: A multi-modal knowledge graph based on

synchronized multi-view videos of daily activities, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024. To appear.
[17] C. Angsuchotmetee, R. Chbeir, Y. Cardinale, MSSN-Onto: An ontology-based approach for flexible event processing in Multimedia Sensor Networks, Future Generation Computer Systems 108 (2020) 1140–1158. doi:10.1016/j.future.2018.01.044.
[18] S. Egami, T. Ugai, M. Oono, K. Kitamura, K. Fukuda, Synthesizing Event-Centric Knowledge Graphs of Daily Activities Using Virtual Space, IEEE Access 11 (2023) 23857–23873. doi:10.1109/ACCESS.2023.3253807.
[19] T. Ebrahimi, C. Horne, Mpeg-4 natural video coding – an overview, Signal Processing: Image Communication 15 (2000) 365–385. doi:https://doi.org/10.1016/S0923-5965(99)00054-5.
[20] FFmpeg Team, FFmpeg, 2000. URL: https://ffmpeg.org/, accessed: 2024-05-27.
[21] A. K. Joshi, P. Hitzler, G. Dong, Logical linked data compression, in: P. Cimiano, O. Corcho, V. Presutti, L. Hollink, S. Rudolph (Eds.), The Semantic Web: Semantics and Big Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 170–184.
[22] J. D. Fernández, M. A. Martínez-Prieto, C. Gutierrez, Compact representation of large rdf data sets for publishing and exchange, in: P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, B. Glimm (Eds.), The Semantic Web – ISWC 2010, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 193–208.
[23] N. Fernández, J. Arias, L. Sánchez, D. Fuentes-Lorenzo, Ó. Corcho, Rdsz: An approach for lossless rdf stream compression, in: V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, A. Tordai (Eds.), The Semantic Web: Trends and Challenges, Springer International Publishing, Cham, 2014, pp. 52–67.