Neuro-symbolic Visual Reasoning for Multimedia Event Processing: Overview, Prospects and Challenges

Muhammad Jaleed Khan, Edward Curry

SFI Centre for Research Training in Artificial Intelligence, Data Science Institute, National University of Ireland Galway, Galway, Ireland.

Abstract

Efficient multimedia event processing is a key enabler for real-time and complex decision making in streaming media. The need for expressive queries to detect high-level human-understandable spatial and temporal events in multimedia streams is inevitable due to the explosive growth of multimedia data in smart cities and internet. The recent work in stream reasoning, event processing and visual reasoning inspires the integration of visual and commonsense reasoning in multimedia event processing, which would improve and enhance multimedia event processing in terms of expressivity of event rules and queries. This can be achieved through careful integration of knowledge about entities, relations and rules from rich knowledge bases via reasoning over multimedia streams within an event processing engine. The prospects of neuro-symbolic visual reasoning within multimedia event processing are promising, however, there are several associated challenges that are highlighted in this paper.

Keywords

multimedia event processing, visual reasoning, commonsense reasoning, video stream processing, spatiotemporal events

1. Introduction

Internet of multimedia things (IoMT), data analytics and artificial intelligence are continuously improving smart cities and urban environments with their ever-increasing applications ranging from traffic management to public safety. As middleware between internet of things and real-time applications, complex event processing (CEP) systems process structured data streams from multiple producers and detect complex events queried by subscribers in real-time. The enormous increase in image and video content surveillance cameras and other sources in IoMT applications posed several challenges in real-time processing of multimedia events, which motivated researchers in this area to extend the existing CEP engines and to devise new CEP frameworks to support unstructured multimedia streams. Over the past few years, several efforts have been made to mitigate the challenges in multimedia event processing by developing techniques for extension of existing CEP engines for multimedia events [1] and development of end-to-end CEP frameworks for multimedia streams [2]. On the other hand, the research in computer vision has focused on complimenting object detection with human-like visual reasoning that allows for prediction of mean-

Proceedings of the CIKM 2020 Workshops, October 19-20, Galway, Ireland.

m.khan12@nuigalway.ie (M.J. Khan); edward.curry@nuigalway.ie (E. Curry)
https://www.linkedin.com/in/mjaleedkhan/ (M.J. Khan); http://edwardcurry.org/ (E. Curry)
0000-0003-4727-4722 (M.J. Khan); 0000-0001-8236-6433 (E. Curry)

ingful and useful semantic relations among detected objects based on analogy and commonsense (CS) knowledge [3, 4].

2. Background

In this paper, we discuss the background, prospects and challenges related to leveraging the existing visual and commonsense reasoning to enhance multimedia event processing in terms of its applicability and expressivity of multimedia event queries. The *motivation* for development of an end-to-end multimedia event processing system supporting automated reasoning over multimedia streams comes from its potential real-time applications in smart cities, internet and sports. Fig. 1 shows an example of traffic congestion event detected using visual and commonsense reasoning over the objects and relations among the objects in the video stream. A conceptual level design and a motivational example of a novel CEP framework supporting visual and commonsense reasoning is presented in Fig. 2.

This section presents a review of the recent work in stream reasoning, multimedia event processing and visual reasoning that could be complementary within a proposed neuro-symbolic multimedia event processing system with support for visual reasoning.

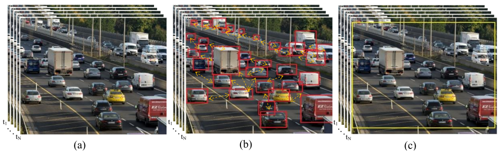

Figure 1: (a) Example of video stream in smart city. (b) Detection of objects and relations. (c) High-level event of traffic congestion detected as a result of automated reasoning.

2.1. Reasoning over Streams and Knowledge Graph

Emerging from the semantic web, streaming data is conventionally modelled according to RDF [8], a graph representation. The real-time processing of RDF streams is performed in time-dependent windows that control the access to the stream, each containing a small part of the stream over which a task needs to be performed at a certain time instant. Reasoning is performed by applying RDF Schema rules to the graph using SPARQL query language or its variants. Reasoning over knowledge graphs (KG) provides new relations among entities to enrich the knowledge graph and improve its applicability [9]. Neuro-symbolic computing combines symbolic and statistical approaches, i.e. knowledge is represented in symbolic form, whereas learning and reasoning are performed by DNN [10], which has shown its efficacy in object detection [11] as well as enhanced feature learning via knowledge infusion in DNN layers from knowledge bases [12]. Temporal KG allows time-aware representation and tracking of entities and relations [13].

2.2. Multimedia Event Representation and Processing

CEP engines inherently lacked the support for unstructured multimedia events, which was mitigated by a generalized approach for handling multimedia events as native events in CEP engines as presented in [1]. Angsuchotmetee et al. [14] has presented an ontological approach for modeling complex events and multimedia data with syntactic and semantic interoperability in multimedia sensor networks, which allows subscribers to define application-specific complex events while keeping the low-level network repre-

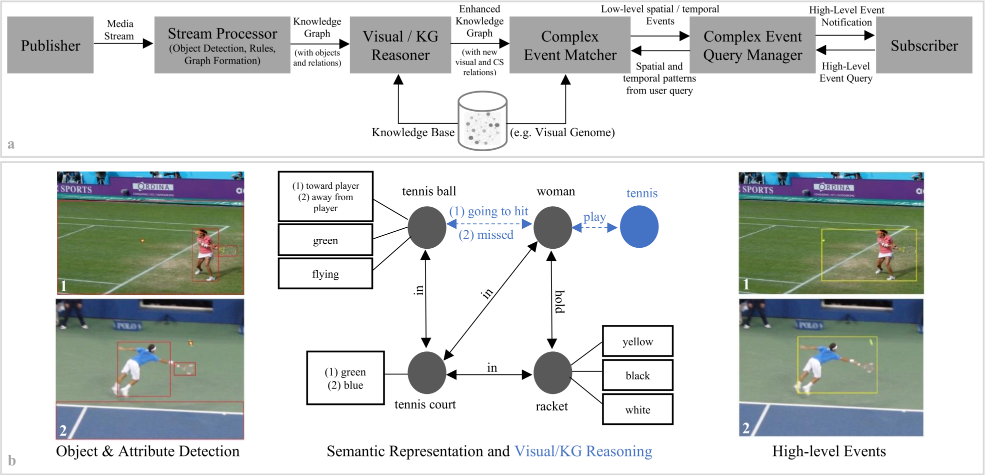
Publisher
Media Stream
Stream Processor
(Object Detection, Rules, Graph Formation)
Knowledge Graph
(with objects and relations)
Visual / KG Reasoner
Enhanced Knowledge Graph
(with new visual and CS relations)
Complex Event Matcher
Low-level spatial / temporal Events
Spatial and temporal patterns from user query
Complex Event Query Manager
High-Level Event Notification
High-Level Event Query
Subscriber
Knowledge Base
(e.g. Visual Genome)
a
(1) toward player (2) away from player
green
flying
tennis ball
woman
tennis
(1) going to hit
(2) missed
play
in
in
hold
(1) green (2) blue
tennis court
in
racket
yellow
black
white
1
2
Object & Attribute Detection
Semantic Representation and Visual/KG Reasoning
High-level Events
b

Figure 2: (a) Conceptual level block diagram of a CEP framework supporting visual reasoning. The input stream of images (or video frames) is received from a publisher, the objects are detected using DNN and rule-based relations [5] are represented using a graph, which is followed by automated reasoning that adds new visual relations from a knowledge base [6] and validates those relations using commonsense knowledge [7]. The matcher performs spatial and temporal event matching on these detected objects and relations with the spatial and temporal patterns in high-level events queried by the subscriber. (b) An example of visual reasoning in multimedia event processing. Suppose a subscriber is interested in the event where tennis player is either "hitting" or "missing" a shot. This event is not explicitly defined via rules but it can be predicted via automated reasoning over detected objects and predicted relations. (Image credits: Visual Genome [6])

sentation generic. Aslam et al. [15] leveraged domain adaption and online transfer learning in multimedia event processing to extend support for unknown events. Knowledge graph is suitable for semantic representation and reasoning over video streams due to its scalability and maintainability [16], as demonstrated in [5]. VidCEP [2], a CEP framework for detection of spatiotemporal video events expressed by subscriber-defined queries, includes a graph-based representation, Video Event Query Language (VEQL) and a complex event matcher for video data.

2.3. Visual and Commonsense Reasoning

In addition to the objects and their attributes in images, detection of relations among these objects is crucial for scene understanding for which compositional models [17], visual phrase models [18] and DNN based relational networks [19] are available. Visual and semantic embeddings aid large scale visual relation detection, such as Zhang et al. [4] employed both visual and textual features to leverage the interactions between objects for relation detection. Similarly, Peyre et al. [3] added a visual phrase embedding space during learning to enable analogical reasoning for unseen relations and to improve robustness to appearance variations of visual relations. Table 1 presents some knowledge bases publicly available for visual reasoning. Wan et al. [7] proposed the use of commonsense knowledge graph along with the visual features to enhance visual relation detection. Rajani et al. [20] leverage human reasoning and language models to generate human-like explanations for DNN-based commonsense question answering. There are various commonsense reasoning methods and datasets available for visual commonsense reasoning [21] and story completion [22].

3. Neuro-symbolic Visual Reasoning in Multimedia Event Processing

3.1. Prospects

The current multimedia event representation methods use knowledge graph to represent the detected objects, their attributes and relations among the objects in video streams. Pre-defined spatial-temporal rules are used to form relations among the objects. However, the complex relations that exist among real-world objects also depend on semantic facts and situational variables that can not be explicitly specified for every possible event as rules. The statistical reasoning methods and knowledge bases discussed in Section 2 have great potential to complement the rule-based relation formation in multimedia event processing by injecting some semantic knowledge and reasoning to extract more semantically meaningful relations among objects. This advancement will allow subscribers to define abstract or high-level human-understandable event query rules that can be decomposed into spatial and temporal patterns. The spatiotemporal matching of the queried high-level events will be performed on the objects, rule-based relations and relations extracted using visual reasoning. The subscriber will be instantly notified of the high-level event as a combined detection of those spatial-temporal patterns. The idea of developing an end-to-end multimedia event processing system supporting visual reasoning over video streams (Fig. 2) poses several challenges that are discussed in the next section. This novel approach will give more expressive power to subscribers in querying complex events in multimedia streams, and thus increase the scope of real-time applications of multimedia event processing in smart city applications as well as internet media streaming applications.

3.2. Challenges

1. Suitable representation for reasoning It is crucial to select a generalized and scalable model to represent events and effectively perform automated reasoning to derive more meaningful and expressive spatiotemporal events.

2. Expressive query definition and matching Providing a generic and human-friendly format to subscribers for writing expressive and high-level queries would require new constructs. Matching queries with the low-level events and relations along with reasoning via knowledge bases requires efficient retrieval within the complex event matcher. Real-world complex events can share similar patterns, occur as a cluster of similar events or occur in a hierarchical manner, which requires generalized, adaptive and scalable spatiotemporal constructs to query such events.

3. Labeling and training samples of visual relations There can be a large numbers of objects and possible relations among them in images, which can result in a large number of categories of relations. It is difficult

Table 1
Available Knowledge Bases for Visual Reasoning

Knowledge Base	#Images	#Entity Categories	#Entity Instances	#Relation Categories	#Relation Instances
Open Images V4 [23]	9,200,000	600	15,400,000	57	375,000
YAGO 4 [24]	–	10,124	64,000,000	–	2 billion
Visual Genome [6]	108,077	33,877	3,843,636	42,374	2,269,617
COCO-a [25]	10,000	81	74,000	156	207,000
VisKE [18]	–	1,884	–	1,158	12,593

to annotate all possible relations and to have balanced categories of relations in the training data. For example, Visual Genome [6] has a huge number of relations with unbalanced instances of each relation.

4. Consistent integration of knowledge bases The object labels in datasets for object detection and entity labels in knowledge bases (e.g. person, human, man) are not always the same. Similarly, knowledge bases have different labels for the same entity, different names for the same attribute (e.g. birthPlace and placeOfBirth) or relation (e.g. 'at left' and 'to left of'). This can cause inconsistency or redundancy while integrating relations from the knowledge bases. It is important to select the knowledge base and dataset that are consistent and suitable for the combined use of both object detection and visual reasoning.

5. Supporting rare or unseen visual relations Apart from the common relations, very rare or unseen relations among objects also appear in certain scenes. It is nearly impossible to collect sufficient training samples for all possible seen and unseen relations. Handling such relations while evaluating the models is also a challenge.

6. Temporal processing of objects and relations The recent methods on this subject address complex inference tasks by decomposing images or scenes into objects and visual relations among the objects. The temporal events and temporal tracking of the detected objects and predicted relations has not been explored much, which is crucial for spatiotemporal event processing.

Acknowledgement

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

[1] A. Aslam, E. Curry, Towards a generalized approach for deep neural network based event processing for the internet of multimedia things, IEEE Access 6 (2018) 25573–25587.

[2] P. Yadav, E. Curry, Vidcep: Complex event processing framework to detect spatiotemporal patterns in video streams, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 2513–2522.

[3] J. Peyre, I. Laptev, C. Schmid, J. Sivic, Detecting unseen visual relations using analogies, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1981–1990.

[4] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, M. Elhoseiny, Large-scale visual relationship understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 9185–9194.

[5] 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), IEEE, 2019, pp. 13–20.

[6] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (2017) 32–73.

[7] H. Wan, J. Ou, B. Wang, J. Du, J. Z. Pan, J. Zeng, Iterative visual relationship detection via commonsense knowledge graph, in: Joint International Semantic Technology Conference, Springer, 2019, pp. 210–225.

[8] Rdf 1.1 concepts and abstract syntax (2014).

[9] X. Chen, S. Jia, Y. Xiang, A review: Knowledge reasoning over knowledge graph, Expert Systems with Applications 141 (2020) 112948.

[10] W. Li, G. Qi, Q. Ji, Hybrid reasoning in knowledge graphs:

Combing symbolic reasoning and statistical reasoning, Semantic Web (2020) 1–10.

[11] Y. Fang, K. Kuan, J. Lin, C. Tan, V. Chandrasekhar, Object detection meets knowledge graphs (2017).

[12] U. Kursuncu, M. Gaur, A. Sheth, Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning, arXiv preprint arXiv:1912.00512 (2019).

[13] A. García-Durán, S. Dumančić, M. Niepert, Learning sequence encoders for temporal knowledge graph completion, arXiv preprint arXiv:1809.03202 (2018).

[14] 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.

[15] A. Aslam, E. Curry, Reducing response time for multimedia event processing using domain adaptation, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 261–265.

[16] L. Greco, P. Ritrovato, M. Vento, On the use of semantic technologies for video analysis, Journal of Ambient Intelligence and Humanized Computing (2020).

[17] Y. Li, W. Ouyang, X. Wang, X. Tang, Vip-cnn: Visual phrase guided convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1347–1356.

[18] F. Sadeghi, S. K. Kumar Divvala, A. Farhadi, Viske: Visual knowledge extraction and question answering by visual verification of relation phrases, in: CVPR 2015, ????, pp. 1456–1464.

[19] B. Dai, Y. Zhang, D. Lin, Detecting visual relationships with deep relational networks, in: Proceedings of the IEEE conference on computer vision and Pattern recognition, 2017, pp. 3076–3086.

[20] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).

[21] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[22] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, Hellaswag: Can a machine really finish your sentence?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[23] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al., The open images dataset v4, International Journal of Computer Vision (2020) 1–26.

[24] T. P. Tanon, G. Weikum, F. Suchanek, Yago 4: A reasonable knowledge base, in: European Semantic Web Conference, Springer, 2020, pp. 583–596.

[25] M. R. Ronchi, P. Perona, Describing common human visual actions in images, arXiv preprint arXiv:1506.02203 (2015).