

A Conceptual Architecture for Deep-learning-based Video-Objects Retrieving System

Dinh-Lam Pham^{1,2}, Viet-Vu Vu², Byeongnam Yoon¹, Kyoung-Sook Kim¹, Kwanghoon Pio Kim¹

¹Contents Convergence Software Research Institute, Kyonggi University, South Korea

²Information Technology Institute, Vietnam National University Hanoi, Vietnam

{phamdinhlam, vuvietvu}@vnu.edu.vn, {phamdinhlam,tomayoon, khmjmc, kwang}@kgu.ac.kr

Abstract— Finding content in videos or video-objects retrieving systems has been a matter of concern for many years until now. Regularly, to retrieve and explore content in videos, the traditional solution is that we oblige to use the human eyes and examine those videos manually until we obtain the content that appears in the video. To speed up the search, we can fast-forward the video or predict when the content appeared. Nevertheless, this method often consumes a lot of time, especially when we need to retrieve content in a large number of Videos. In this paper, we introduce a deep-learning-based architecture that allows retrieving objects in videos instantly to search objects in CCTV cameras. This solution promises to bring high efficiency in finding video content, thereby saving costs and human resources to operate related systems.

Keywords— The video-objects retrieving system, video content retrieval, CCTV Camera, deep learning-based, object detections

I. INTRODUCTION

In today's modern social life, using CCTV cameras to support the local government to monitor urban security has become popular in most countries. We can see CCTV cameras appearing everywhere on the streets, public areas, bus stations, etc. The formation and development of CCTV cameras started in the 1950s of the last century. At that time, with the limited technology and expensive cost of CCTV systems, only a few large cities in rich countries could afford to install CCTV surveillance systems [1][2].

Fortunately, with the development of technology, the cost of hardware and software to install a CCTV system has been dramatically reduced, creating favorable conditions for building a CCTV camera system in a wide area. Along with the popularity of the CCTV systems is the problem of managing a massive amount of videos generated every day. To monitor and access the content of these videos, we need manual human support. In recent years, with the significant development of Deep learning algorithms, many solutions have been introduced and applied to help us manage CCTV systems effectively.

With the development of the Convolutional neural network introduced by Prof. Yann Lecun [3], many algorithms have been built to support detecting objects in images, thereby

applying to object detection in Video. Some highly efficient algorithms include YOLO (You only look once) [4] and SSD (Single shot multibox detector) [5]. By using deep learning-based algorithms, we can efficiently identify common objects appearing in videos (One of the popular data used to train and detect objects in images is the COCO [6] (common objects in context) dataset). Applying artificial intelligence and deep learning to identify content in videos produces positive effects in the management, operation, and use of CCTV systems [7][8]. Instead of having to use naked eyes to monitor and review video content, we can use deep learning algorithms to analyze the video content, convert it to text-video format, and then search for the content through text describing the video content that has been built.

In this paper, we introduce an architecture that supports locating objects in video based on deep learning. Notably, we present an architecture that uses the YOLO algorithm to automatically detect objects in videos and then contextualize video content based on detected objects. The contextualized data will be stored in a JSV format, obeyed a schema, and then indexed through the Apache Solr system. After that, agents who are tasked with analyzing video content (E.g., police department officers) can quickly retrieve content from these videos.

The remainder of this paper is organized as follows. The following section discusses the related works and motivation for introducing our architecture. After that, in the Architecture section, we present and explain in detail the architecture and its component for supporting video-objects retrieval. Finally, the last section conclusions the content of this paper and future works.

II. RELATED WORKS AND MOTIVATION

In the last decade, the application of deep learning algorithms in video content recognition has made significant progress. Researchers are increasingly developing neural networks and applying them in the recognition and analysis of objects in videos, thereby helping us extract information from stored videos more effectively. Some related studies include:

Applying sequence-to-sequence model to generate text from video [9]. Retrieving video object by trajectory and appearance [10]. Using context-aware concept to realize the content in the video [11][12][13][14]. Using the LSTM (Long-short term memory) model to understand the video content and objects [15][16][17][18]. Or using some others techniques to generate video captioning, video-text [19][20][21][22].

The studies mentioned above used a variety of approaches for analyzing the video objects. However, the general goal of these researches is to help us understand the content of the video better, thereby conducting analysis and finding information based on the video content data. In this paper, we want to present the architecture to detect video objects and hence contextualize video contexts to retrieve content in the video. With the architecture, we can create a system that finds content in videos efficiently. This architecture will save costs and improve the efficiency of CCTV system administration.

III. ARCHITECTURE AND DATA SCHEMA

1. Architecture

Figure 1 describes the proposed architecture and its modules. As shown in Figure1, the architecture contains three main Modules:

- The YOLO object detection module (A): This module uses the YOLO algorithm for detecting the objects in videos. Video types can come from different sources, e.g., CCTV, Social media video, etc.
- The Video object contextualization module (B): This module stores detected objects in the JSV (Json video) format following a specific schema. The concept in the schema is divided by video files, clips, objects in clips, and their properties.
- The Indexing module (C): This module uses Apache Solr [23] for indexing the content in JSV files. After indexing JSV content, we can use this module for retrieving content and getting relevant frames, clips.

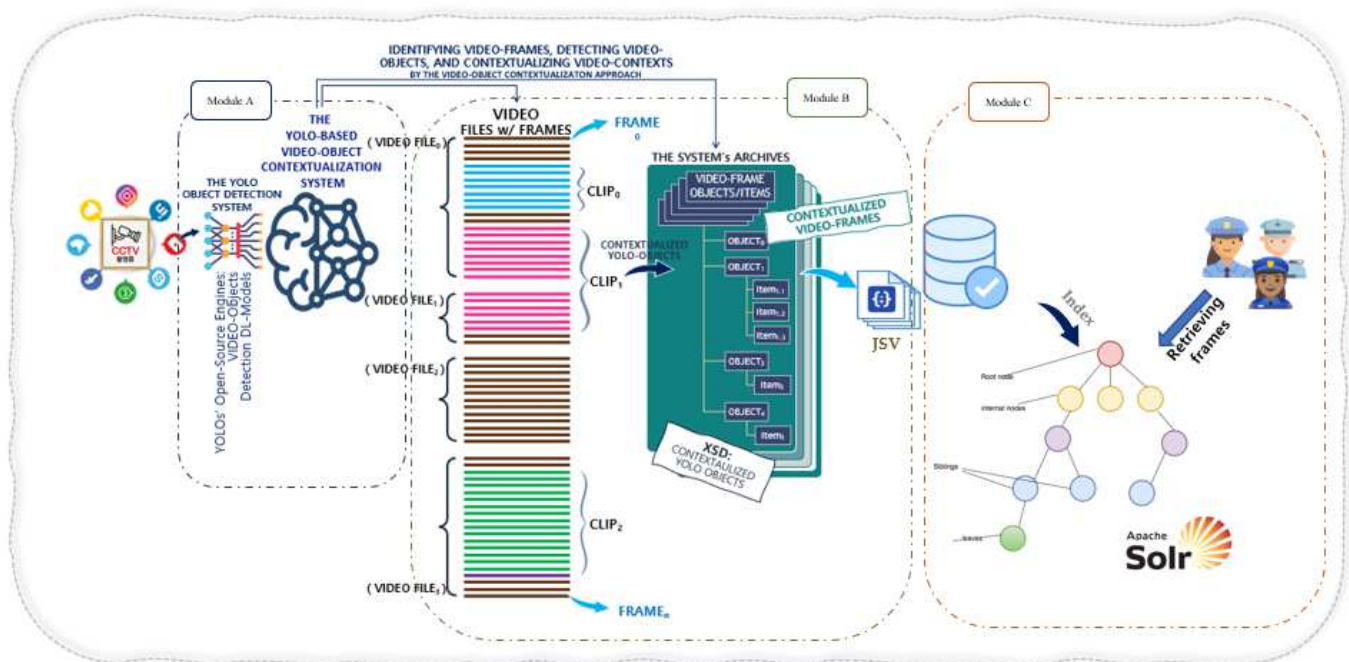


Figure 1. The Architecture for Retrieving Video-Objects with YOLO and Apache Solr.

The architecture procedure is described as follows: First, in Module A, video files are gathered from different sources such as CCTV systems, black boxes in cars, social networks, etc. Then, for each Frame in the video, the YOLO algorithm will detect the objects. After collecting information about detected objects, we proceed with the contextualization process at module B. In module B, a JSV structure is specified to store data according to the formation. This JSV structure includes the following components:

- Metadata: It contains descriptive information about the video such as creation time, location, fps, etc.
- Frame ID: The index of the Frame in the video.
- Detected objects in each Frame: The list of detected objects in each Frame. This message includes the object type, location.
- Properties of each object.

Then, in Module C, we use Apache Solr to perform content indexing in the generated JSV files. With Apache Solr's supporting features such as full-text search, real-time indexing, clustering, we can instantly retrieve and access the Frame containing the search objects.

2. Contextualization Data Schema

In the proposal architecture, we use a JSV format, a specific schema for storing the contextualization data of detected objects from the input videos. The name JSV comes from the combination of JSON and Video. In Figure 2, we present the

schema of JSV with its entities and relations. As shown in Figure 2, a JSV file contains:

- Metadata: Storing the brief information of the video file.
- Frame: Storing information of detected objects in each Frame by using the entity named Property, E.g., the ID of Frame, the location of each detected object.
- Attribute: Represent the characteristics of entities in JSV.

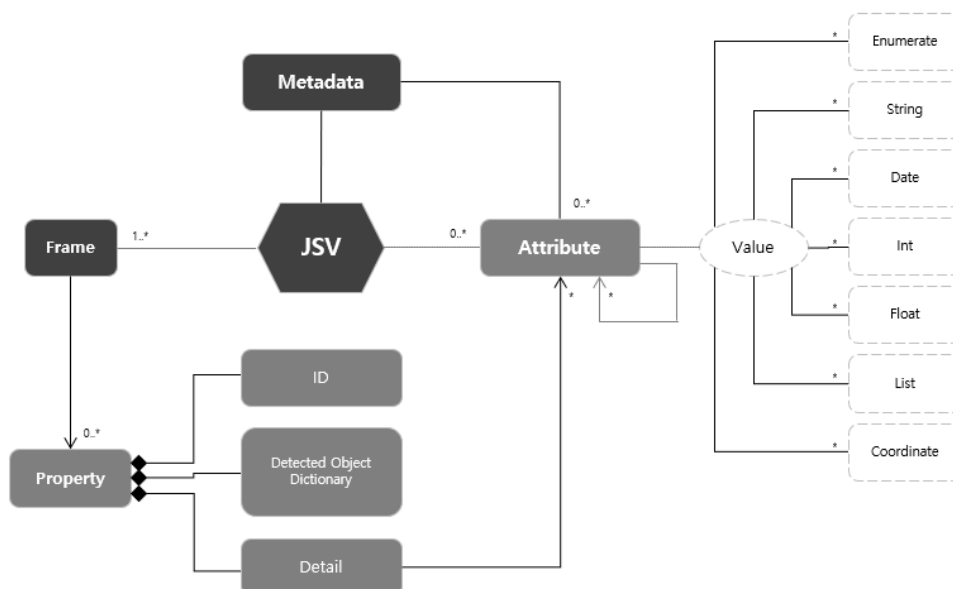


Figure 2. The data schema storing contextualization data.

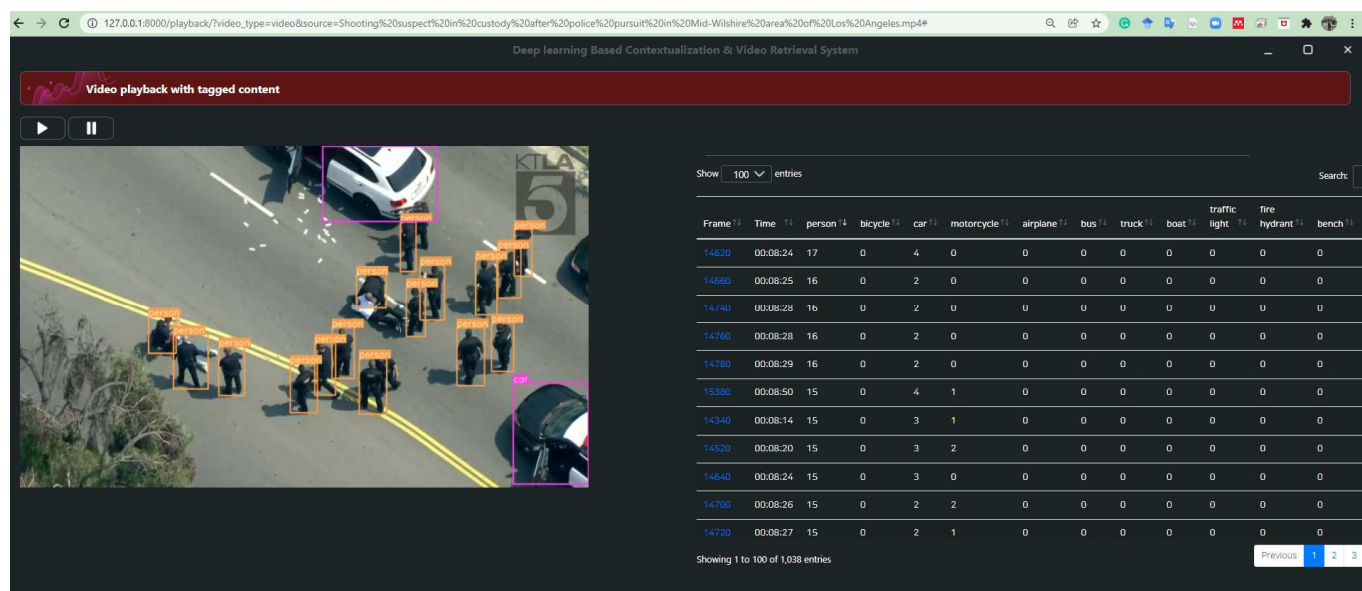


Figure 3. Retrieving frames containing more than 10 persons and 2 cars from a CCTV video file.

By using the JSV format, we can store all the information of detected objects and index data by using **Module C** in the Architecture.

IV. EXPERIMENTAL RESULTS

Using the proposal architecture in the preview section, we conduct building a system for retrieving video objects. **Figure 3** shows the retrieving frames with the condition containing more than 10 persons and 2 cars from a CCTV video file. The video was recorded from a Helicopter about the police pursuit shooting suspects in Los Angeles, United States, in August 2020 [24].

V. CONCLUSIONS

In this paper, we have introduced an architecture generating contextualization video data by using the YOLO algorithm and JSON format. After contextualizing data, we use the Apache Solr to index and search content to gather object information in each video frame. Using this approach architecture, we can build the video retrieval system for helping the agents who are tasked with analyzing video content. In future works, we will create the system based on this architecture and put it to real-life use.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number 2020R1A6A1A03040583).

REFERENCES

- [1] C. A. Williams, "Police Surveillance and the Emergence of CCTV in the 1960s," *Crime Prev. Community Saf.* 2003 53, vol. 5, no. 3, pp. 27–37, Jul. 2003.
- [2] C. Norris, M. McCahill, and D. Wood, "The Growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space," *Surveill. Soc.*, vol. 2, no. 2/3, pp. 110–135, Sep. 2002.
- [3] Y. Lecun, Y. Bengio, and R. 4g332, "Convolutional Networks for Images, Speech, and Time-Series."
- [4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 779–788.
- [5] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.
- [6] "COCO - Common Objects in Context." [Online]. Available: <https://cocodataset.org/#home>. [Accessed: 21-Nov-2021].
- [7] X. He and L. Deng, "Deep learning for image-to-text generation: A technical overview," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 109–116, Nov. 2017.
- [8] M. Amaresh and S. Chitrakala, "Video captioning using deep learning: An overview of methods, datasets and metrics," *Proc. 2019 IEEE Int. Conf. Commun. Signal Process. ICCSP 2019*, pp. 656–661, Apr. 2019.
- [9] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to Sequence - Video to Text." pp. 4534–4542, 2015.
- [10] Y. H. Lai and C. K. Yang, "Video object retrieval by trajectory and appearance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 1026–1037, Jun. 2015.
- [11] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-Aware Surveillance Video Summarization," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5469–5478, Nov. 2016.
- [12] S. Niklaus and F. Liu, "Context-Aware Synthesis for Video Frame Interpolation." pp. 1701–1710, 2018.
- [13] W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, "ContextVP: Fully Context-Aware Video Prediction." pp. 753–769, 2018.
- [14] W. Xiang, D. Q. Zhang, H. Yu, and V. Athitsos, "Context-aware single-shot detector," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, 2018, vol. 2018-Janua, pp. 1784–1793.
- [15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video Captioning with Attention-Based LSTM and Semantic Consistency," *IEEE Trans. Multimed.*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.
- [16] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features," *IEEE Access*, vol. 6, pp. 1155–1166, Nov. 2017.
- [17] I. Md. Sanzidul, S. M. Sadia Sultana, S. Abujar, and S. A. Hossain, "Sequence-to-sequence Bangla Sentence Generation with LSTM Recurrent Neural Networks," *Procedia Comput. Sci.*, vol. 152, pp. 51–58, Jan. 2019.
- [18] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.
- [19] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video Captioning via Hierarchical Reinforcement Learning." pp. 4213–4222, 2018.
- [20] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Text-to-Clip Video Retrieval with Early Fusion and Re-Captioning," *arXiv Prepr. arXiv1804.05113*, 2018.
- [21] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-Temporal Dynamics and Semantic Attribute Enriched Visual Encoding for Video Captioning." pp. 12487–12496, 2019.
- [22] L. Zhu and Y. Yang, "ActBERT: Learning Global-Local Video-Text Representations." pp. 8746–8755, 2020.
- [23] "Welcome to Apache Solr - Apache Solr." [Online]. Available: <https://solr.apache.org/>. [Accessed: 23-Nov-2021].
- [24] "Shooting suspect in custody after police pursuit in Mid-Wilshire area of Los Angeles - YouTube." [Online]. Available: <https://www.youtube.com/watch?v=mYr-mGZYdrl>. [Accessed: 07-Dec-2021].



Dinh-Lam Pham is a research professor in the Contents Convergence Software Research Institute at Kyonggi University, South Korea. He received B.S. degree and M.S. degree in computer science from Tainguyen University in 2008 and 2010, respectively, and received Ph.D. degree in Computer Science from Kyonggi University in 2021. His research interests include video retrieving system, deep learning, workflow systems, BPM, workflow-supported social and affiliation networks discovery and analysis, process mining, deep-learning-based process predicting.



Assoc. Prof. Dr. **Viet-Vu Vu** received the B.S. degree in Computer Science from Ha Noi University of Education in 2000, a M.S. degree in Computer Science from Ha Noi University of Technology in 2004, and a Doctor Degree in Computer Science from Paris 6 University in 2011. He is a researcher at Information Technology Institute, Vietnam National University, Hanoi. His research interests include clustering, active learning, semi-supervised clustering, and E-government applications.

Byeongnam Yoon (M'97) He became a Member of IEEE in 1997. He was born in Seoul Korea 1949. He got the PhD in computer science, Chungnam National University, Dejeon city, Korea, 1997. He worked for the Sperryrand UNIVAC as a Computer Specialist 1974 - 1978, Samsung as a Manager of Telecommunications Section 1978 -1982, Electronics & Telecommunications Research Institute (ETRI) as a Principal Researcher 1982 - 1999, National Information society Agency (NIA) as a

Senior Executive Director General 1999 - 2010, Kyonggi University as an Associate Professor Faculty of Computer Science 2010 - 2016. Vietnam National University, Hanoi as a Invited professor 2017 - 2020. His research area include a Telecommunications, Internet, Software, Web programming & security, e-Government, Enterprise Architecture, Work Flow, System Work Method. SPiCE, CMMI, BPM, etc

Kyoung-Sook Kim is an adjunct professor of the department of computer science and engineering at Kyonggi University, South Korea. She received her B.S. degree in computer science from Kyonggi University in 1984. She also received her M.S. and Ph.D. degrees in computer engineering from Kyunghee University in 2001 and 2019, respectively. She had worked as a researcher and developer at FAST Systems Inc. and OSI Corp. in South Korea. Her research interests include large-scale database systems and applications,

enterprise information systems, real-time databases, distributed databases, temporal databases, federate databases, object-oriented database management systems, and the Internet of Things collaboration architectures and services.

Kwanghoon Pio Kim is a full professor of computer science department and the founder and supervisor of the collaboration technology research laboratory at Kyonggi University, South Korea. He received B.S. degree in computer science from Kyonggi University in 1984. And he received M.S. degree in computer science from Chungang University in 1986. He also received his M.S. and Ph.D. degrees from the computer science department at University of Colorado Boulder, in 1994 and 1998, respectively. At Kyonggi University, he is Dean of the

Computerization and Informatics Institute, and the director of the contents convergence software research center, as well. He had worked as researcher and developer at Aztek Engineering, American Educational Products Inc., and IBM in USA, as well as at Electronics and Telecommunications Research Institute (ETRI) in South Korea. In present, he is a vice-chair of the BPM Korea Forum. He has been in charge of a country-chair (Korea) and ERC vice-chair of the Workflow Management Coalition. He has also been on the editorial board of the journal of KSII, and the committee member of the several conferences and workshops. His research interests include groupware, workflow systems, BPM, CSCW, collaboration theory, Grid/P2P distributed systems, process warehousing and mining, workflow-supported social networks discovery and analysis, process-aware information systems, data intensive workflows, and process-aware Internet of Things.