CONTENTS CONVERGENCE SOFTWARE RESEARCH INSTITUTE AT KGU

HPC Asia 2024: International Conference on High Performance Computing
In Asia-Pacific Region, Jan. 25-27, 2024, Nagoya, JAPAN

KGU KYONGGI UNIVERSITY

A Bigdata Acquisition Framework of Deep-Learning-based CCTV-Video Contextualization Machines

Eun-Bee Cho & Batt Chaya*
{eunbee0508, tsbaachka95}@kgu.ac.kr
Data and Process Engineering Research Lab.
Department of Public Safety Bigdata
Graduate School of KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Kyung-Hee Sun†
kh_sun@kgu.ac.kr
Contents Convergence Software R.I.
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Dinh-Lam Pham‡
phamdinhlam@kgu.ac.kr
Contents Convergence Software R,I,
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Kyoung-Sook Kim§
khmjmc@kgu.ac.kr
Contents Convergence Software R.I.
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Jeong-Hyun Chang¶
crime_tiger564@kgu.ac.kr
Contents Convergence Software R.I.
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

Kwanghoon Pio Kim‖
kwang@kgu.ac.kr
Data and Process Engineering Research Lab.
Division of AI Computer Science and Engineering
KYONGGI UNIVERSITY
Suwon-si, Gyeonggi-do, South Korea

ABSTRACT

This paper proposes a conceptual framework of CCTV-video contextualization machines and implements its concrete system by developing a video-object detection deep-learning model based on YOLO neural network architecture [2][3]. The primary functionality of the proposed framework is for detecting active contextual clues, like objects, motions, and physical environs, on every CCTV-video frame and codifying the detected clues [4] into a new formation of structured code-format named as COME-Code\footnote{Note that the COME-Code is an abbreviation for contextual objects, motions, and environs of the classified active contextual clues.}. In other words, the detected active contextual clues are codified into the textual forms of objects, motions, and environs with their properties, and structured by the markup-styled formats of the systematic COME-Code scheme. Consequently, through the proposed CCTV-video contextualization machines, we can initiate not only a new era of CCTV-surveillance Bigdata achieves and their engineering disciplines, but also a new paradigm of CCTV-driven crime-prevention services that are detecting, predicting, and preventing criminal situations and behaviors [1][5], but also providing intelligence-led policing and predictive patrol scheduling operations in real-time. Finally, we verify the feasibility and functional correctness of the proposed framework by developing a CCTV-video contextualization system that is able to detect video-objects of the active contextual clues on every CCTV-video frame under the technological support of the YOLO object detection deep learning models and produce the JSON-formatted COME-Code datasets corresponding to the object-based active contextual clues on the CCTV-video frames. Additionally, as one of the future works, we will develop a CCTV-video retrieval system based upon the deep learning driven CCTV-video contextualization machines proposed in this paper.

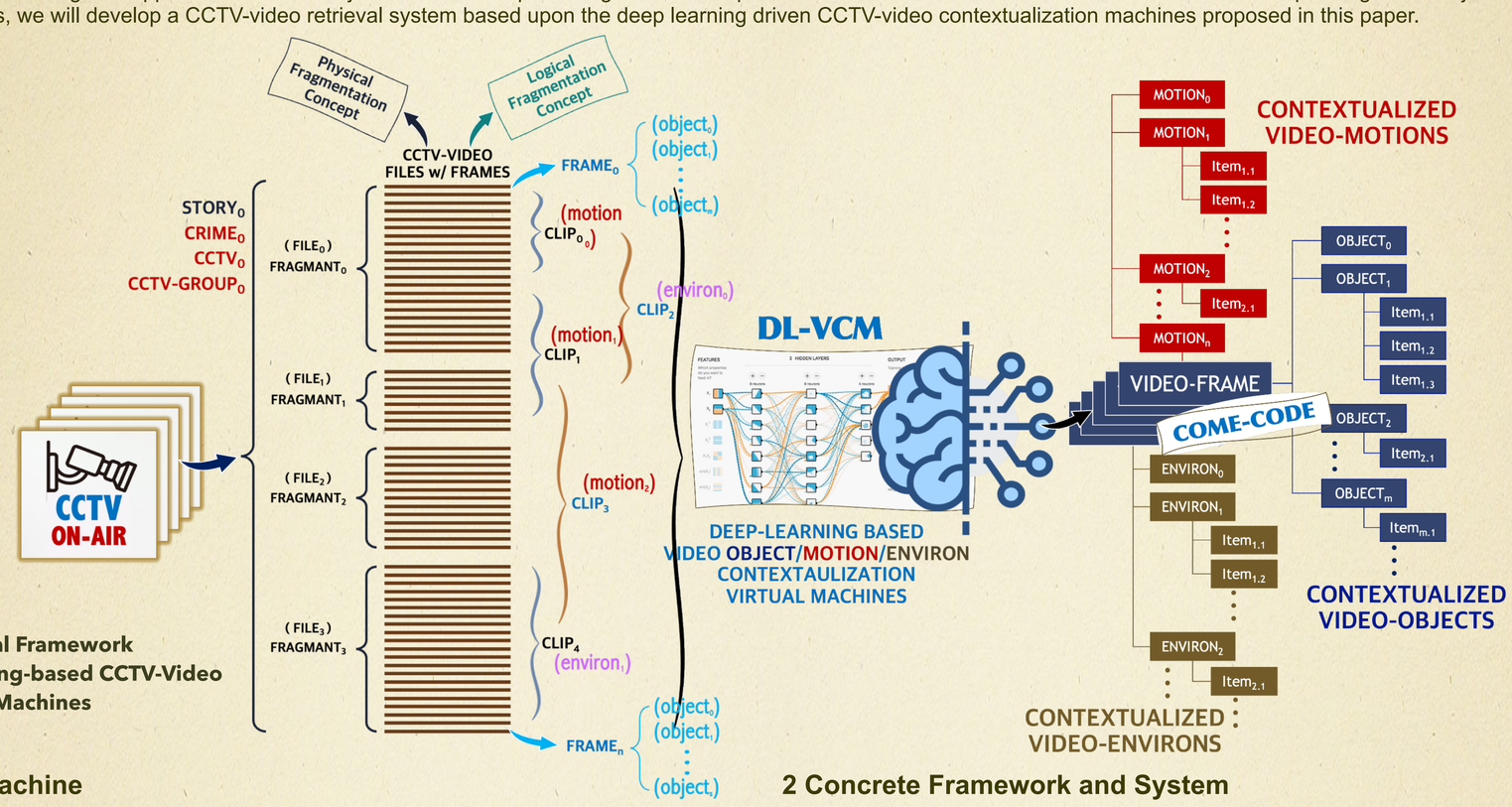

Fig. 1. A Conceptual Framework of the Deep-Learning-based CCTV-Video Contextualization Machines

1 CCTV-Video Contextualization Machine

The essential component of the proposed framework is the CCTV-video contextualization machine supported by the novel and innovative concept of the CCTV-video contextual clue detection approaches based upon the deep neural network models. The functional components of the proposed machine are as follows: first, the CCTV-video manager is identifying video-frames from the input video clippings and files, second, the YOLO video-object detector is detecting YOLO-objects on each of the CCTV-video-frames, and third, the transformer is contextualizing the detected YOLO-objects and their properties into the context-objects represented in a textual formation of XML-schema and JSON formats. Finally, all the contextualized CCTV-video's YOLO-objects are transferred and stored on a cloud-based archive under the name of the active video-contexts bigdata.

1.1 YOLO-Objects and Properties. The machine is fundamentally based upon the real-time YOLO video-object detection framework [2][3]. So far, a series of the YOLO systems has been spawned from the framework, and these systems are well known as the fastest video-object detectors that detect, in real time, 80 different categories (COCO Classes Dataset) of video-objects on every video-frame. The authors' research group has successfully developed a couple of additional deep neural network models so as to detect such properties that are the supplementary characteristics of a corresponding detected object. Consequently, the functional goal of the proposed system is to contextualize every CCTV-video frame by detecting the contextual clues (which is called as YOLO-objects, in particular) with their innate properties as well as their supplementary characteristics and transforming them into a textual formation of XML schema format including JSON format.

1.2 COME-Code: Contextual Clues in the Standardized Data Format. The eventual output of the CCTV-video-object contextualization machine is a contextual clue dataset of a corresponding CCTV-video clip, each of which is coded as COME-Code in an XML schema structure. The COME-Code is for formatting the detected video-objects into the corresponding contextualized video-objects on all the video-frames of the CCTV-video clippings. The functional components of the proposed system produce the COME-code bigdataset contextualized from a file of CCTV-video streamings in realtime, and the codes are formed with the data-schema of the meta-models, such as the deep learning model, video content model, video context model, and crime prevention model. Note that it is necessary to differentiate a logical group of video-frames from a physical group of video-frames in a way of fragmenting the identified video-frames out of the input CCTV-video streaming file. The physical group of video-frames is called as Video-Fragments, while the logical group of video-frames is called as Video-Clippings. These terminologies of video-fragments and video-clippings are usefully applied as the detection range-units of the video-objects' behaviors and situations. At last, all the video-objects and their properties are detected on each of the video-frames is contextualized, formatted, and stored in a COME-code formatted bigdataset. The COME-code bigdataset can be eventually transformed into any type of the XSD format, JSON format, and others.

2 Concrete Framework and System

The essential functionality of the CCTV-video object contextualization system is concretized with two functions: detection and contextualization. The former is to detect objects on each video-frame of the input CCTV-video clipping file, and the latter is to contextualize the detected objects and their properties in a formation of the JSV format presented in the previous subsection. More specifically, the detection function is supported by the YOLO system, while the contextualization function is implemented from scratch by the authors' research group, for themselves. In other words, the video-object contextualization system is realized by integrating the YOLO's detection functionality onto the contextualization functionality, as shown in the conceptual and functional framework of Fig. 2.

As stated in the previous subsection, the YOLO system can detect the 80 different categories of video-objects in a unified fashion of real-time. The primary principle of our approach is to make the best use of the YOLO's detection ability. The concrete framework illustrated in Fig. 2 starts from a CCTV equipment capturing a series of video frames of the public street-view clipping in realtime, which becomes eventually input video-frames. From these input CCTV-video frames, the YOLO-object detection function detects and identifies a group of persons and a single car on each video-frame of the public street-view clipping. We can easily become aware that the YOLO system is able of identify those video-objects being indicated with the color-lined boxes of different categories: car in deep-blue-lined box and person in pale-blue-lined box. Additionally, the YOLO system provides a series of valuable properties of the detected video-objects such as position with two points of (X, Y) coordinates with width and height properties, confidential probability, colors, and others. By using these basic properties of the detected CCTV-video objects, we can make more delicate detection functions to be used for guesstimating motions and situations on the CCTV-video frames as precisely as possible.

Next is about the CCTV-video object contextualization system that are performed via two essential operations: Transformation and Guesstimate. Transformation fulfills an operational command that transforms the detected YOLO-objects and their properties into the contextual clues (e.g. objects, motions, and environs) that are formatted in a formation of the JSV format and XML tagging format, as well; Guesstimate executes a set of analytical guesstimate functions, each of which operates a guesstimate function either analytically deciding the additional property of the discovered CCTV-video object or estimating the contextual clues (motion or environ) on a frame-sequence of the identified video-clippings. Note that the meaning of contextualization implies to give their own semantical as well as contextual clues for characterizing and describing the detected CCTV-video objects. Fig. 2 is to depict the functional framework of the CCTV-video contextualization machine. Conclusively, in the next subsection, we verified the operability and capability of the essential functionality of the CCTV-video contextualization system by applying them onto a real CCTV-video clipping file.

3 Experimental Validation

Based upon the concrete framework described in the previous subsection, the authors' research group implemented the CCTV-video contextualization system basically supporting the CCTV-video contextual clue (YOLO-object) detection functionality and the transformation functionality, as well. We tried to verify the functional correctness of the system through applying to a sample file of CCTV-video clippings captured from a real CCTV device installed at a street in Suwon. Fig. 3 shows a group of captured screens, each of which is produced by executing each operation of the functional framework of the CCTV-video object contextualization system, respectively. A captured CCTV-video frame in the left-most of the figure visualizes a group of YOLO-objects detected and identified with boxes by the YOLO-based deep learning system; The captured screen in the middle visualizes a tree-structured metadata containing all the CCTV-video frames and their properties; The two right-most captured screens visualize the contextual clues (YOLO-objects) in the JSON-formatted COME-Code and the enlarged contextual clues of YOLO-objects and their properties corresponding to the specific CCTV-video frame (i.e. FrameID = 10), respectively.

Summarily, we carried out an experimental verification to prove the functional correctness of the fundamental operations such as YOLO-object detection, frame identification, COME-Code trans- formation, and contextual clue guesstimate to be executed by the CCTV-video contextualization system. Consequently, it is certified for the conceptual architecture and its functional framework proposed in this paper to be operable and reasonable in not only performing the CCTV-video con- textualization and bigdata analysis activities but also practically being applied to the CCTV-video surveillance platforms and systems.

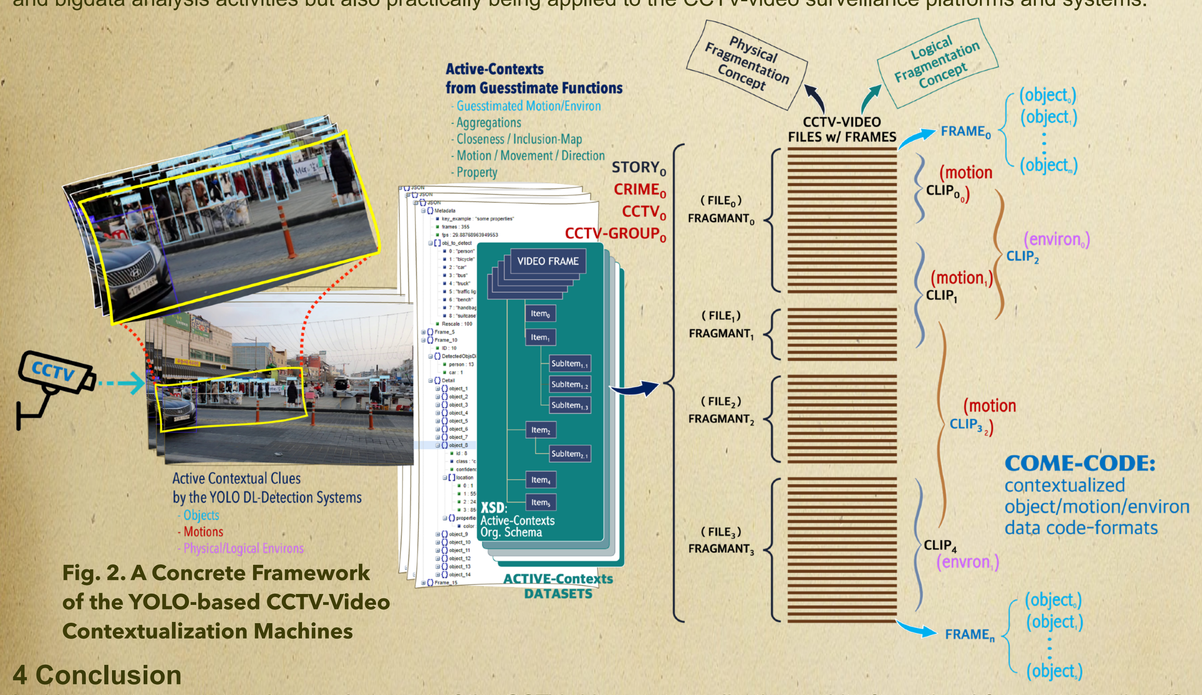

Fig. 2. A Concrete Framework of the YOLO-based CCTV-Video Contextualization Machines

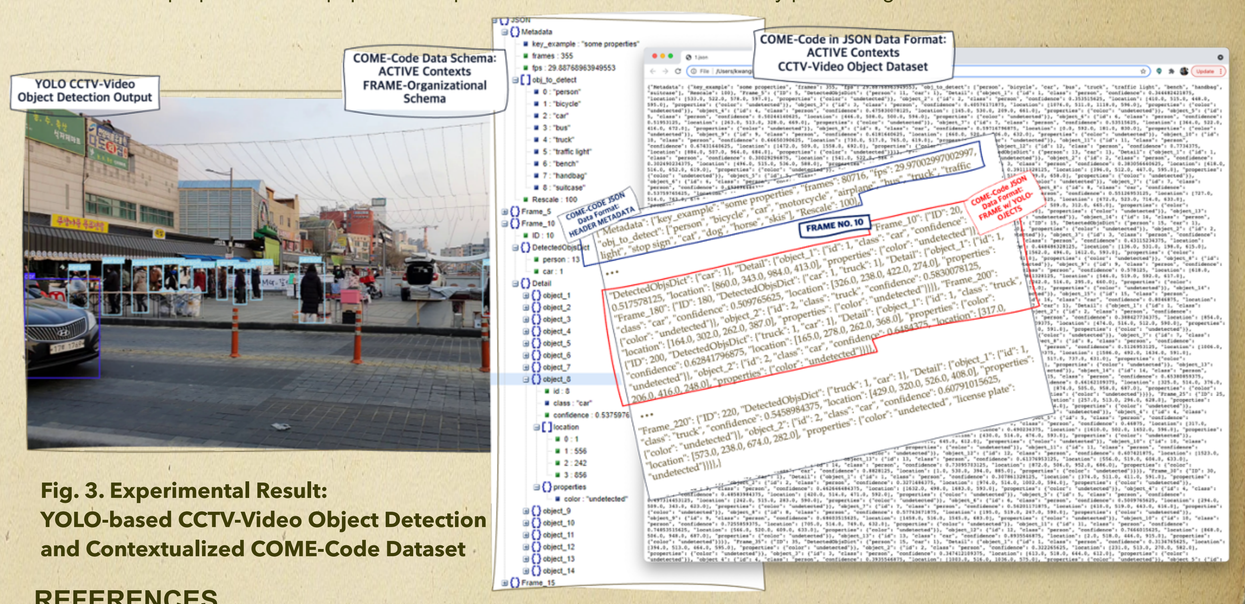

Fig. 3. Experimental Result: YOLO-based CCTV-Video Object Detection and Contextualized COME-Code Dataset

4 Conclusion

In this paper, we proposed a novel concept of the CCTV-video contextualization and its func- tional framework and verified the functional correctness via an experimental verification example. Consequently, the proposed conceptual architecture and its implemented system are tangible and applicable as a meaningful tool for the CCTV-video surveillance platforms by successfully implementing a CCTV-video object contextualization system with the cutting-edge deep learning approach, YOLO. The authors' research group has our confidence on this system's expansibility and applicability in many video-related bigdata engineering platforms and services. Additionally, we strongly believe that the huge amount of the contextualized CCTV-video bigdata actively collected from the CCTV-video clippings and devices ought to be a very valuable and impeccable clue to efficiently and effectively resolve those various social safety problems issued on the video data-flooding era.

REFERENCES

[1] Hamad Ali Abosaq, Muhammad Ramzan, Faisal Althobiani, Adnan Abid, Khalid Mahmood Aamir, Hesham Abdushkour, Muhammad Irfan, Mohammad E. Gommosani, Saleh Mohammed Ghonaim, V. R. Shamji, and Saifur Rahman. 2023. Unusual Driver Behavior Detection in Videos Using Deep Learning Models. Sensors 23, 1 (January 2023), 311–330.

[2] Tanvir Ahmad, Yinglong Ma, Muhammad Yahya, Belal Ahmad, Shah Nazir, and Amin ul Haq. 2020. Object Detection through Modified YOLO Neural Network. Scientific Programming 2020 (2020), 8403262.

[3] Tausif Diwan, G. Anirudh, and Jitendra V. Tembhurne. 2023. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimedia Tools and Applications 82 (March 2023), 9243–9275.

[4] Lisa Ehrlinger, Johannes Schrott, Martin Melichar, Nicolas Kirchmayr, and Wolfram Wöß. 2021. Data Catalogs: A Systematic Literature Review and Guidelines to Implementation. Database and Expert Systems Applications - DEXA 2021 Workshops. DEXA 2021. Communications in Computer and Information Science 1479 (September 2021), 148–158.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. Computer Vision and Pattern Recognition (cs.CV), arXiv:1812.03982 [cs.CV] https://doi.org/10.48550/arXiv.1812.03982 (2019).

ACKNOWLEDGEMENTS
This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number 2020R1A6A1A03040583).
The research outcomes in this paper belong to the Contents Convergence Software Research Institute in Kyonggi University.

NRF 한국연구재단 Ministry of Education