지능형 비디오 감시를 위한 온톨로지 기반의 고차원 컨텍스트 추론 기법

{aftab, jkhanbk1, azher, numan, Irfan, yklee}@khu.ac.kr

Higher-Level Context Inferences based on Ontology For Intelligent Video Surveillance

Aftab Alam, Jawad Khan, MD Azher Uddin, Muhammad Numan Khan, Irfan Ullah, Young-Koo Lee

Department of Computer Science and Engineering, Kyung Hee University

ABSTRACT

There is increasing reliance on the intelligent CCTV systems for effective analysis and interpretation of the streaming data with the intentions to recognize activities and to ensure public safety. Monitoring videos captured by surveillance cameras is always a difficult and time-consuming task. There is a need for automated analysis using computer vision methods in order to recognize/predict abnormal activities and assist authorities. Once, videos are processed using computer vision technologies; another problem is how this data is indexed for search, analysis, and real-time alerts since a large number of cameras continuously capture videos resulting vast amounts of data. In order to address this issue, in this paper, we propose a generic architecture for distributed intelligent surveillance and is composed of four layers. The first layer acquisition large number of the video streams from for device independent video stream data sources. In the second layer, we use computer vision algorithms for semantic video annotation while exploiting the distributed in-memory computing engine. The third layer is used to persist the video stream and the to manage the intermediate results being produced by the second layer. Finally, the intermediate results are mapped to the RDF according to domain-specific application.

1. INTRODUCTION

Since its inception, security cameras have played a significant role in the security system and become ubiquitous with the evaluation of the technology. Real-time video stream monitoring or searching is a challenging task. Studies show that after 22 minutes of continuous video monitoring, up to 95% of activities are missed. Computer vision technologies thus are needed to analyze surveillance videos, segment and recognize abnormal activities automatically. Large-scale real-time video stream analytics is an active research area and has attracted the attention of the research community to bring intelligence by analyzing and investigating the salient aspect of the streaming videos. The streaming video is systematically analyzed with the intention to anticipate or detect abnormal activities either in real-time or offline to ensure public safety. While analyzing and processing real-time large-scale video streaming is a multi-level activity and needs processing power. One feasible and optimal solution is to exploit distributed computing technologies.

In video search and analysis, efficient and intelligent retrieval of videos based on their semantic contents is essential. Despite the tools for semantically annotating individuals or objects (e.g., person, car, etc.) in videos [1], crowd behavior annotation models and tools are the new directions of research. In [2], they present a platform to annotate and then search for traffic videos using ontologies. However, the whole process is manual, which is time-consuming and does not scale. In [3], Semantic Web and crowded scene analysis is combined. Object-based crowd behavior analysis is performed, and a tracking ontology is proposed to track annotated objects (e.g., Person, Car) in order to find people falling or aggression to a person. However, in their ontology design, existing multimedia standards have not been applied.

Similarly, Multimedia annotation is a popular research topic, and several standards have been developed over the years. Instead of creating new semantic models, first, we investigated existing multimedia ontologies. We aim to extend existing maturated models whereas possible since re-usability is an important part of ontology engineering.

Motivated by the limitations of existing work, in this paper, we propose a generic, scalable architecture for real-time intelligent video stream analytics while exploiting state-

of-the-art technologies. The proposed architecture is composed of four main layers, i.e., The Video Stream Acquisition Layer, the Video Annotation Layer, Distributed Big data store, and Knowledge Curation Layer. The first layer, acquire large-scale video stream from real-time video stream data sources. The second layer is responsible to annotate the process and annotate the complex events in the video while using distributing computing engine. The produced intermediate results and video are maintained in the third layer. Finally, the Data Curation Layer, map the Intermediate Results to RDB for intelligent reasoning.

The main contributions of this paper are to propose a novel architecture for large-scale intelligent video analytics in the cloud while using distributed computing, computer vision, and semantic web technologies. We organized the rest of this paper as follows – section II presents the proposed architecture, and use cases in section III. Finally, we discuss the conclusion in section IV.

2. PROPOSED System Architecture

We propose a generic architecture for distributed intelligent surveillance video analysis in the cloud while exploiting state-of-the-art cloud computing technologies. The block diagram of the proposed architecture is shown in Fig.1. The proposed framework is composed of four main components, which are explained in the following subsections.

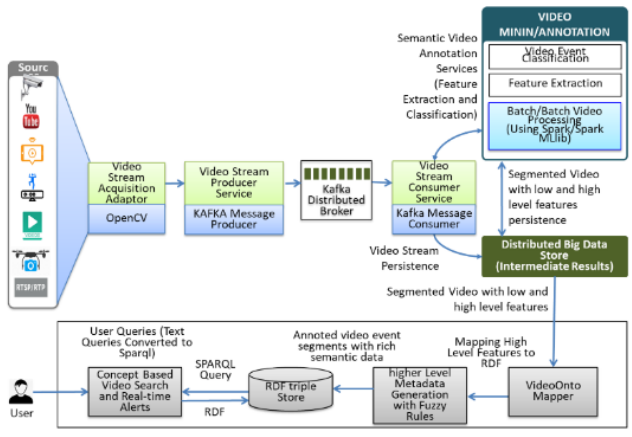

Figure 1: Proposed Intelligent Surveillance architecture.

A. *Video Data Acquisition*

The real-time video stream needs to be collected from the source device and forwarded to the executors for on the fly processing against the subscribed video annotation service. Handling a tremendous amount of video streams, both processing and storage are subject to lose. To handle, large-scale video stream acquisition in real-time and to ensure the scalability and fault-tolerance, we develop Video Stream Acquisition Adopter. Component while assuming a distributed messaging system. This component decodes the video stream, detects the frames, and then performs some necessary operations on each frame such as meta-data extraction and frame resizing, which is then converted to a formal message. These messages are then serialized in the form of mini-batches, compressed, and sent to the Broker to Kafka Topic t.

As the acquired video streams are now residing in the Kafka Broker in different Queues in the form of mini-batches. The Video Stream Consumer Service is used to read the mini-batches of the video stream from the respective topic in the Kafka Broker to distribute video analytics (using Spark[1]) and to persist the same to Distributed Big Data Store (HBase[2]). In the next section, we explain how we process the video for annotations.

B. *Distributed in-memory Video Annotation*

The objective of video annotation is to infer semantic rich tags from videos, where it can be helpful to bring the higher-level information to persons or detect activities of interest. The semantic annotation framework is illustrated in Fig. 2. This framework is composed of two main components, i.e., Training and Testing phases. In the Training phase, the users are allowed to train the model. First, extract features (e.g., poselets, optical flow, and local space-time regions) from video sequences. These features, together with labels, are fed into a classifier to train the model. In the Testing phase, the unseen video sequences or video streams (Acquired from the Kafka Server) are classified into pre-defined action categories. During this process, videos are also extracted to the same features used in the training process.

The proposed system support some packages for feature extraction and annotation while using in-memory distributed computing, i.e., Spark. In particular, we provide various video processing and computer vision libraries that help users extract the features to serve for their desired applications. For example, with our provided libraries, CNN [4], which perform well on a variety of datasets, can be utilized.

[1] https://spark.apache.org/

[2] https://hbase.apache.org/

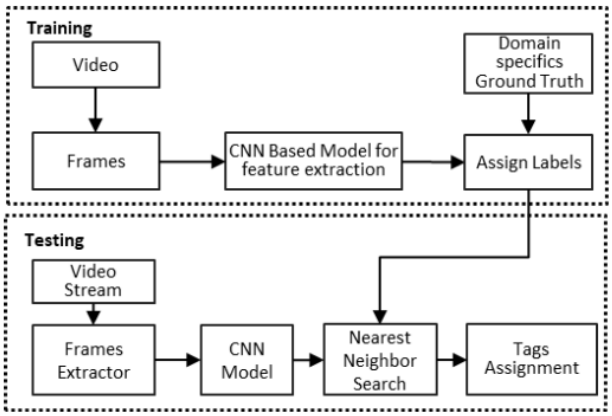

Figure 2: Framework for semantic video annotation.

C. BDCL- Intermediate Results Management

The BDPL component is responsible for providing permanent and distributed big-data persistence to both the structured and unstructured data of the proposed platform. When a mini-batch of the video is classified and tagged in the second step then the respective high-level features, (also known as Intermediate Results) are persistent to BDCL.

D. Knowledge Curation Layer

The KCL is responsible for mapping the IR (video annotation stored in the BDCL) to RDF store for searching and browsing. This Layer is composed of four components i-e. Video Onto Mapper, Higher Level Metadata Generator, RDF Triple Store, and Video Browser.

The VideoOnto Mapper and RDF Generator component obtained the output (Intermediate Results) of the distributed computer vision module stored in the DBDS and mapped the respective ontology.

Then RDF is generated according to the domain-specific semantic metadata model. To create domain-specific ontology, we will extend existing semantic multimedia related web standards, i.e., Media Resource Ontology[3], MPEG-7[4] Ontology, Annotation Ontology[5], and Provenance Ontology[6].

In the second step, i.e., Higher Level Metadata Generation with SWRL Rules, rules are applied to extract higher-level context. The respective domain specifically generated ontology is then persisted to an RDF triple store (i.e., database), which can be utilized by a search interface for accessing video datasets while using the Video Browser.

3. Use Cases

The proposed architecture can be used in many different use cases, such as abnormal crowd behavior searching and analysis (offline analytics), real-time traffic monitoring, security, and surveillance. The generated rich semantic meta-data can be used for search and analysis, for example, statistical data of abnormal events in a locality. Similarly, search for crowd movements and abnormal events can be retrieved against the semantic rich query while considering different required granularity levels.

4. CONCLUSION

We proposed an architecture for large-scale intelligent video analytics while using distributed computing, computer vision, and semantic web technology. The proposed system is a layered architecture and is composed of four layers. We then elaborate on each layer's functionality. In the future, we are going to implement the proposed architecture will also investigate research issues of each layer.

ACKNOWLEDGEMENT

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2016-0-00406, SIAT CCTV Cloud Platform).

REFERENCES

[1] Fernández, C., Baiget, P., Roca, X., & Gonzalez, J. (2007, September). Semantic annotation of complex human scenes for multimedia surveillance. In Congress of the Italian Association for Artificial Intelligence (pp. 698-709). Springer, Berlin, Heidelberg.

[2] Z. Xu, L. Mei, Y. Liu, H. Zhang, C. Hu. Crowd Sensing Based Semantic Annotation of Surveillance Videos, Int. Journal of Distributed Sensor Networks, 1-9, 2015.

[3] L. Greco,P. Ritrovato, A. Saggese, M. Vento. Abnormal Event Recognition: A Hybrid Approach Using Semantic Web Technologies. In Computer Vision and Pattern Recognition Workshops (CVPR), 2016.

[4] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).

[3] https://www.w3.org/TR/mediaont-10/

[4] https://www.w3.org/2005/Incubator/mmsem/XGR-mpeg7/

[5] https://www.w3.org/ns/oa

[6] https://www.w3.org/TR/prov-o/