SUPEREYE: SMART ADVERTISEMENT INSERTION FOR ONLINE VIDEO STREAMING

* Utku Bulkan, Tasos Dagiuklas, Muddesar Iqbal

***Abstract*— Without any doubt, state of art advertisement insertion mechanisms along with the requested video content has emerged to be the most crucial part of online video delivery ecosystems. There are several widely deployed methods which have been used since the first days of video streaming such as tag word matching between video content versus advertisement content along with manual matching-based approaches. Conventional but non-scalable and context independent methods cannot fulfil the requirements of an online video platform when there are several millions of user generated videos along with premium content and advertisement of varying production quality. In such environment, a content aware advertisement insertion framework is required based on object recognition, machine learning and artificial intelligence to understand the context of the video and match appropriate advertisement and stitch the advertisement at the most convenient moment of the target video content. In this paper, SuperEye; a deployment ready, content aware, scalable, distributed advertisement insertion framework for a 5G oriented online video platform is designed and developed. The foundational object analyzing mechanism of the underlying system examine each particular context that is part of the wider video and advertisement catalogue using object recognition while generating a time-lapse map of all objects that are detected through the video. Based upon this information, the framework matches the most significant object that is detected for a particular interval and associates the advertisement with similar properties. Additionally, this novel technique does not require any watch history or personalized data related to the user, but primarily interested in only the current requested content information. Therefore, this framework can work along with any type of recommendation engine or rank based association algorithm. The proposed framework is independent of the user information and regarding the subjective user results collected, successful video to ad match ratios of SuperEye significantly exceed the current implementations of YouTube, Vimeo and DailyMotion.**

***Index Terms*—Online Video Streaming, Advertisement Insertion, Object Detection, Object Recognition, Behavior Detection, Quality of Experience, Artificial Intelligence**

I. INTRODUCTION

The close relationship of online video streaming and advertisement insertion has evolved to dominate today's world wide web as the primary origin of profit source. The origin of the revenue is generally based on matching correct target consumer with most relevant advertisement. Each successful video advertisement engagement results in a win-win situation for all the parties that build up the online video ecosystem: consumer, content provider, advert owner and the broker which refers to the online video streaming platform in this context.

The consumer requests content from the online video streaming webservices related with genre and relevance interests. Quality of Experience (QoE) is the success of these webservices that can be measured with active watch duration of consumers, which can be directly taken as a degree of customer satisfaction [1].

The content provider creates either premium or user-generated videos to increase channel's reputation and the number of online subscribers. Independent of the size of the content provider, this can be either a multinational news channel on YouTube or a local entertainment channel on Vimeo. The goal is to increase audience watch duration which will attract more advertisement engagement and statistically this concludes as the success of the online video channel [2].

The advert owner intends to reach wide audience with relevant interests about the advertised product. Following the primary intention of reaching a wider audience, correct customer engagement must be established in the shortest feasible campaign duration which will keep the advertisement budget in the lowest possible levels.

The broker hosts the online video streaming platform to provide mechanisms for content providers to broadcast their content in order to provide relevant videos to the consumers, while providing an advertisement dashboard to instantiate a campaign for the advert owner. More consumers with higher active watch durations mean better online video services which concurrently attract higher number of advert owners.

This creates a strong satisfaction-based relationship between the parties that forms the online video ecosystem. However, there is only one fact that prevents the consumer's overall QoE [3]. Consumers intend to watch the provider's content and not the advertisement! So, the primary task of brokers is to find a "smart" way to make the advertisements fit seamless into the actual content that is targeted primarily.

Frontrunners webservices such as YouTube [4], DailyMotion [5] rely primarily on the watch history of the user when selecting appropriate advertisement. The watch history of the consumer is associated with the tag words that are used to describe the video content. These tag words for each video are

* U. B. Author (phone: +44 (0)20 7815 7465; e-mail: bulkanu@lsbu.ac.uk).
T. D. Author (e-mail: tdagiuklas@lsbu.ac.uk).
M. I. Author (e-mail: m.iqbal@lsbu.ac.uk).

assigned by the content provider at the time of uploading the files to the webservice (or publishing the content as a scheduled event). However, the accuracy is debatable when defining a 10+ minutes video with only a couple of words. The videos include a lot of visual information, which is hard to define with only a few words. According to YouTube statistics [6], the average video is labelled with 3 to 4 tag words for an estimated catalogue of 7 billion videos as of January 2020.

The historical user specific interest data of the previous watch sessions are generated on this relevant video tag information. Proprietary advertisement insertion mechanisms select the most convenient advertisement from the ad catalogue while matching the video tag. Also, it must be underlined that these "couple of words" represent the context of the video and does not pinpoint the actual moment of the visual illustration of the relevant tag word that is being perceived along the timeline of the video content.

In addition to all these facts, there is the infamous issue [8] which established GDPR [7] as a mainstream subject. This has become a ready to implode reality for any webservice since the first days of the shenanigans that has been taking place for Cambridge Analytica and Facebook in 2016 [8].

Contradictory to 2000s and 2010s, the ingenious audience of 2020s intends to share minimum amount of their personal data, ideally nil! Nowadays, even the average audience does not prefer the idea of a webservice that knows everything about them, both from interests and susceptibility perspective. The rising success of user data friendly search engines [9] such as DuckDuckGo [10] can be seen as the evidence of such consumer behavior for protecting personal data.

Online video services have evolved to a point that all advertisement related mechanisms should be reconsidered based upon a different point of view regarding the following facts: the inefficiency of video tag word mechanisms [11], inadequacy of video to advertisement context engagement [12], implausible calculation for the successful advertisement insertion moment into the video timeline [13] and protecting personal information in alignment with GDPR [14] where anonymity of the user is guaranteed unless explicitly shared.

This paper provides a novel perspective solution for all of the issues that is being originated from the conventional video tagging based advertisement insertion methodologies. The number of tag words for an average video is far from being sufficient to represent the video.

This might provide thousands of tag words for a one-hour video content. However, more importantly, this will result in a situation where the video can be labelled throughout the timeline! With each consecutive frame labelled in a way to represent the context of the image, the same labelling mechanism can also be applied to the advertisement catalogue.

Now, we have a frame-by-frame labelled video content catalogue along with a catalogue of frame-by-frame labelled advertisement content. The following task is to match the relevancy of the video tags with advertisement tags among a potential of thousands of distinctive labels.

This method can achieve a successful advertisement campaign at the correct insertion moment for such an example occasion: consider an entertainment video in YouTube [4] with multiple animals in it, but from frame-by-frame labelling perspective the video is labelled as a cat video, and luckily one of the advert owners has introduced a cat food advertisement recently.

Although there are other animals visible in the video, the moments where the cat figurine is consecutive in the timeline would provide the most convenient moment to insert cat food advertisement, as this will catch the attention of the audience immediately.

As the audience active watch duration increases, this means that the user's interest in alignment with the content of the video. This provides a positive indicator for a successful advertisement engagement even without any previous personal watch history data!

The analysis for the advertisement insertion mechanism focuses only on the current content. If the user watches the content, it means that she/he is interested. If the user watches the content, the advertisement insertion mechanism must provide a related advertisement concentrating on the content. The only possible overall way is to achieve this is to rely on a frame-by-frame tagging mechanism.

In this paper, an object recognition framework based online video advertisement insertion system has been designed and developed which creates a frame-by-frame timeline tag mapping for each video in the catalogue, either content or advertisement.

The timeline correlated tag information is used for content and advertisement association. This is achieved by matching most relevant tag words which also counterparts the most relevant duration throughout the content. This way achieves not only relevant advertisement and content matching, also solves the issue of selecting the appropriate advertisement insertion moment.

The rest of the paper is given as follows: Section II outlines the related works in online advertisement insertion domain. Section III provides a state of art information about object detection and object recognition models. Following that, Section IV introduces the object recognition and tag word indexing mechanism that is based on machine learning and artificial intelligence.

Section V provides the overall SuperEye platform implementation using the microservices model Section VI declares an explicit list of novelties and contributions of this paper. Section VII discusses the strategies to cover different advertisement durations and formulations regarding the content specific parameters.

Section VIII hints use case scenarios and paradigms the experimentation for system validation. Section IX debates the results and presents a comparison both versus different object detection models and against state of art advertisement insertion methodologies that has been used to validate results.

Section X discusses the overall results and finally Section XI provides the conclusions and future of the modern online video advertisement systems that should assist anticipated 5G web streaming systems.

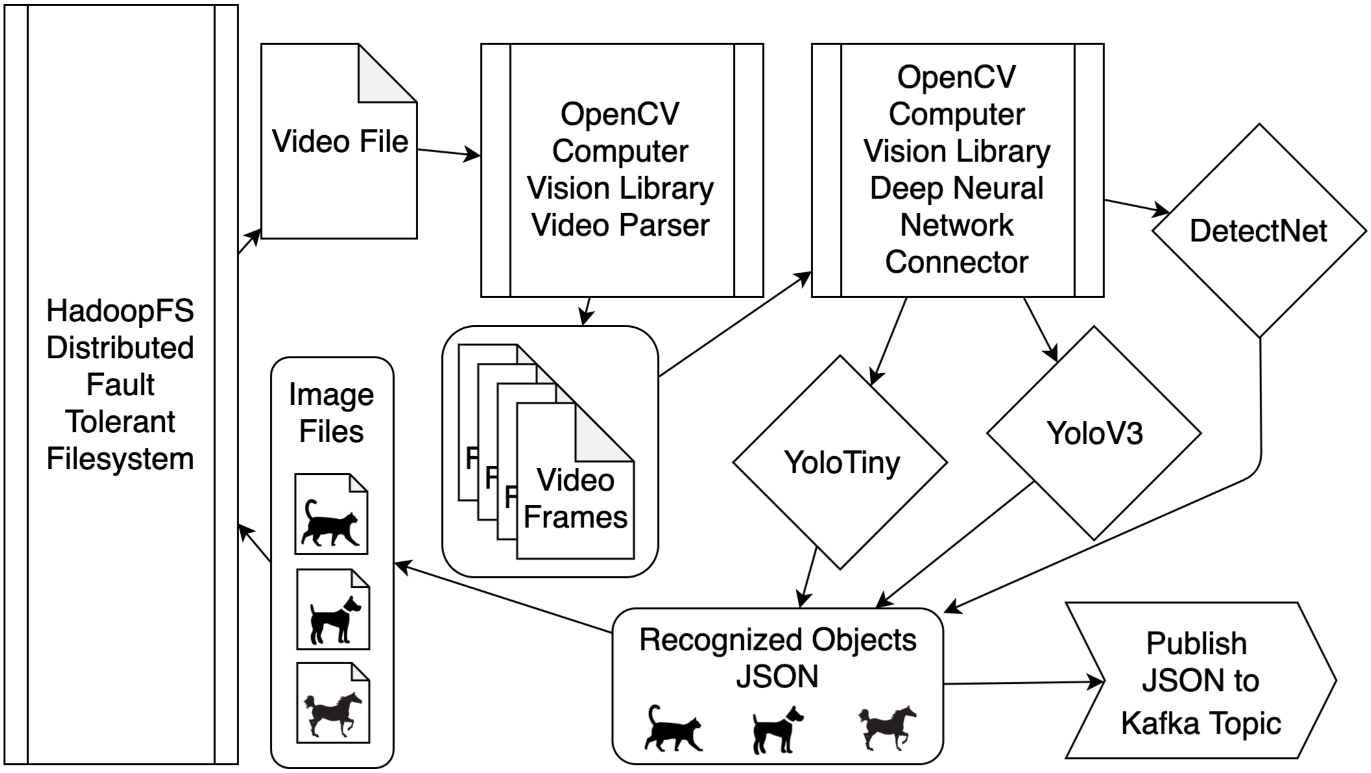
HadoopFS Distributed Fault Tolerant Filesystem
Video File
OpenCV Computer Vision Library Video Parser
OpenCV Computer Vision Library Deep Neural Network Connector
DetectNet
Video Frames
Image Files
YoloTiny
YoloV3
Recognized Objects JSON
Publish JSON to Kafka Topic

Figure 1. Architecture and Algorithm I for SuperEye Core Functionality, Object Detection and Recognition Microservice

II. State of Art object Detection and object Recognition Models

During the last decade, object detection [15] and object recognition [16] methodologies have evolved in a rapid fashion to fulfil the expectations of long anticipated modern 21st century computer vision [17] and artificial intelligence systems [18]. Unarguably, a significant leap forward was mandatory in many aspects of science and technology to develop computer vision systems analogous to human vision capabilities such as achieving real-time object recognition capable software for mass production ready edge mini supercomputers.

Parallel to the quest for any other major scientific target, there has been many scientific achievements required through the path to eureka for advanced object recognition, which principally has taken place in mathematical aspect, especially revolving around machine learning methodologies. The scientific achievements evaluate old but well-known statistical problems in a faster, yet with even a higher precision such as R Convolutional Neural Networks (RCNN) [19]. Support Vector Machines (SVM) [20] superseding the predecessor approaches such as Artificial Neural Networks (ANN) [21] or Kth Nearest Neighbor (KNN) [22]. The significant advances in machine learning methodologies have made this achievable through only parallel processing and with the recent rise of Graphical Processing Unit (GPU) [23] based computing by introducing new computational implementations of mathematical approaches to the well-known statistical problems which require immense processing power.

The brand-new computational methodologies are just mimicking the nature, particularly the eye-brain correlation [24], which forms the foundation of vision capability in many mammals. Although the details of how mammals store required object specific parameters in brain still holds many mysteries for scientific world, computational methods store the object specific data in machine learning mathematical models. Then these models are compared with actual image using GPU powered parallel processing methods.

Training models for object detection with frameworks such as Tensorflow [25] or Caffe [26] has already become widely accessible through well-known public data stores such as github.com [27]. Nowadays, within a couple of google search, object models can easily be found even on public domains that can distinguish different objects with a wide variety of ranges from 10s of objects to several thousands. The sizes of the machine learning models and their relevant footprint in memory changes regarding the number of training data that was fed into the machine learning model, considering the demand and motivation of the training process.

Then comes the part of the connection of these machine learning models with the computer vision frameworks. The two frontrunners NVIDIA [28] and Intel [29] support OpenCV [30] (unofficially) and OpenVino [31] as their primary computer vision framework.

These frameworks have native capability to connect machine learning models that are generated with Onyx, DarkNet, Caffe and Tensorflow [25] using a friendly Application Programming Interface (API).

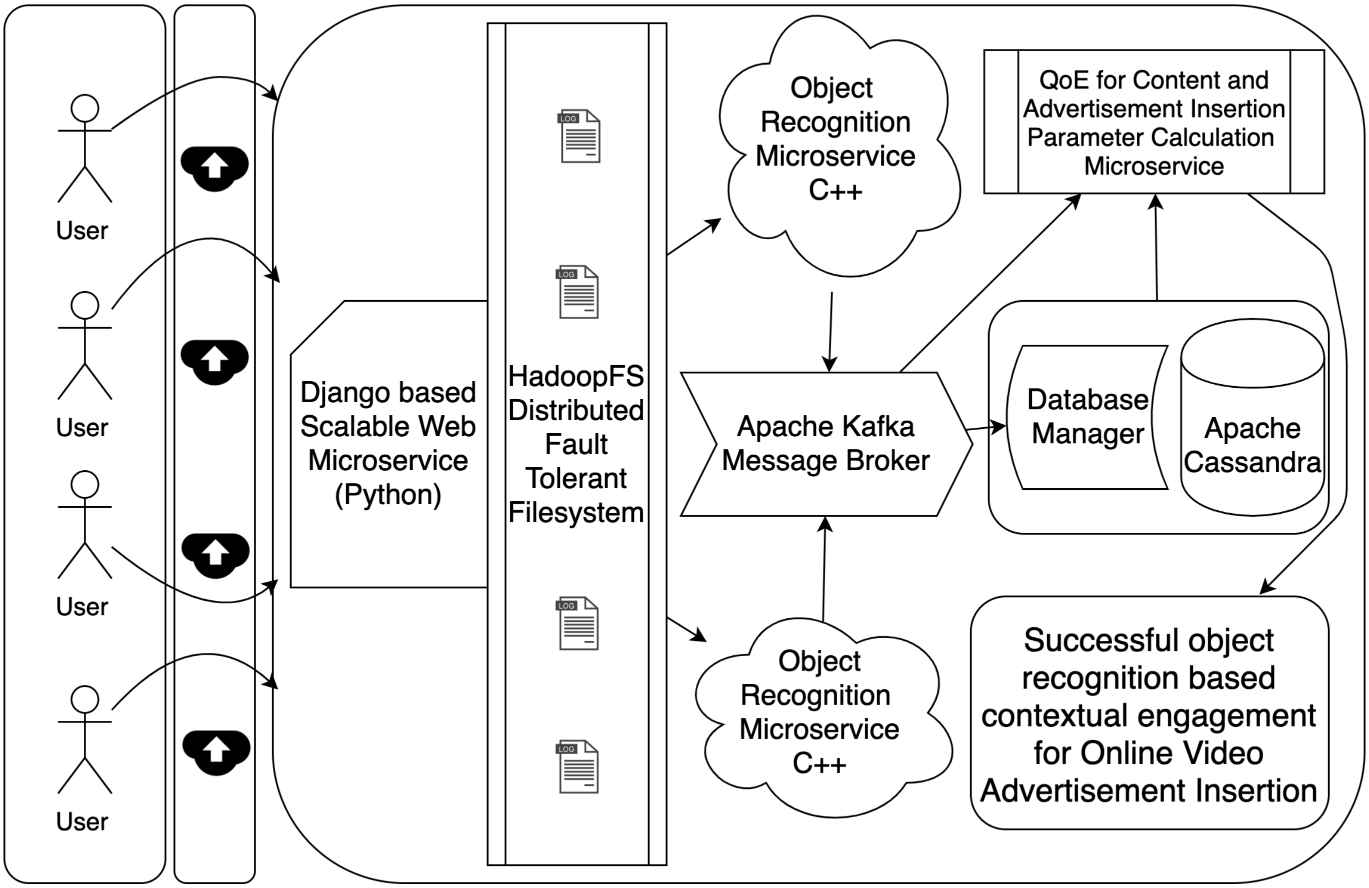
User
User
User
User
Django based Scalable Web Microservice (Python)
HadoopFS Distributed Fault Tolerant Filesystem
LOG
LOG
LOG
LOG
Object Recognition Microservice C++
Object Recognition Microservice C++
Apache Kafka Message Broker
QoE for Content and Advertisement Insertion Parameter Calculation Microservice
Database Manager
Apache Cassandra
Successful object recognition based contextual engagement for Online Video Advertisement Insertion

Figure 2. Overall Architecture for SuperEye Object Recognition based Contextual Engagement Online Advertisement Insertion System

After a short-phase integration, regarding the object-oriented approach of these frameworks, pipelines that support multiple machine learning models can easily be implemented on different demands. Another approach is to provide transfer learning for specific machine learning models, where the models are fine-tuned for quite specific object recognition capabilities such as detecting the species from a distinctive family of flowers or cats.

Regarding these innovations in object recognitions models and the advances in GPU based computing, indexing online video content has become achievable even for small sized video webservices. In accordance with this motivation, this paper has designed and developed an object recognition catalogue system to index all the content that an online video service provides to match with relative advertisement context. Automated Video Tag Indexing for an online video content forms the fundamental foundation of this framework.

III. Object Recognition and Video Tag Word Indexing

Automated Video Tag Indexing microservice for the online video content forms the fundamental foundation of the proposed framework. The steps are given as pseudocode for the object recognition microservice as Algorithm I. Using object detection and object recognition machine learning models, each reference frame [32] in the video content is processed and associated with several number of different objects where applicable.

Algorithm I
Object Recognition Microservice for
Automated Video Tag Indexing

```
1. Load Video File that is Uploaded through Web Dashboard
2. Calculate Unique Video Hash for the video
3. Instantiate object recognition machine learning model
4. Foreach I Frame in Video Content
5. Distinguish object(s) in video, store object data as Image in Distributed File Sytem HadoopFS.
6. Update object recognition data in Distributed database Apache Cassandra.
7. Provide Autogenerated Tag Based Presentation of the video
8. End ForEach
```

Regarding the detected objects, a timeline map is generated for the video where the image data are stored for each object in a distributed file system HadoopFS [33], while the recognized object with relative timestamp is stored in a distributed database Apache Cassandra [34] where the video content is transcoded and prepared for Adaptive Bitrate (ABR) [35] making it ready to be accessible through a choice of compatible Content Delivery Network (CDN) [36].

The source code for the object recognition microservice is written in C++ using OpenCV as the primary computer vision library for image and matrix operations. Deep Neural Network (DNN) connector functionality of OpenCV is used for importing machine learning models YoloTiny, YoloV3, DetectNet, Coco [37] and RCNN that are trained with frameworks such as Caffe and Tensorflow. BACKEND_GPU is selected for OpenCV neural network operations to employ parallel processing capability that is natively supported within the principal computer vision library. The final output of the automated video tag indexing microservice is the JavaScript Object (json) output where each frame is associated with tags wherever available and a json output example is provided with Table II.

IV. Overall Architecture for SuperEye platform Microservices

Object recognition microservice that is discussed in detail in section IV forms the core of the SuperEye smart advertisement system that is powered by machine learning and artificial intelligence.

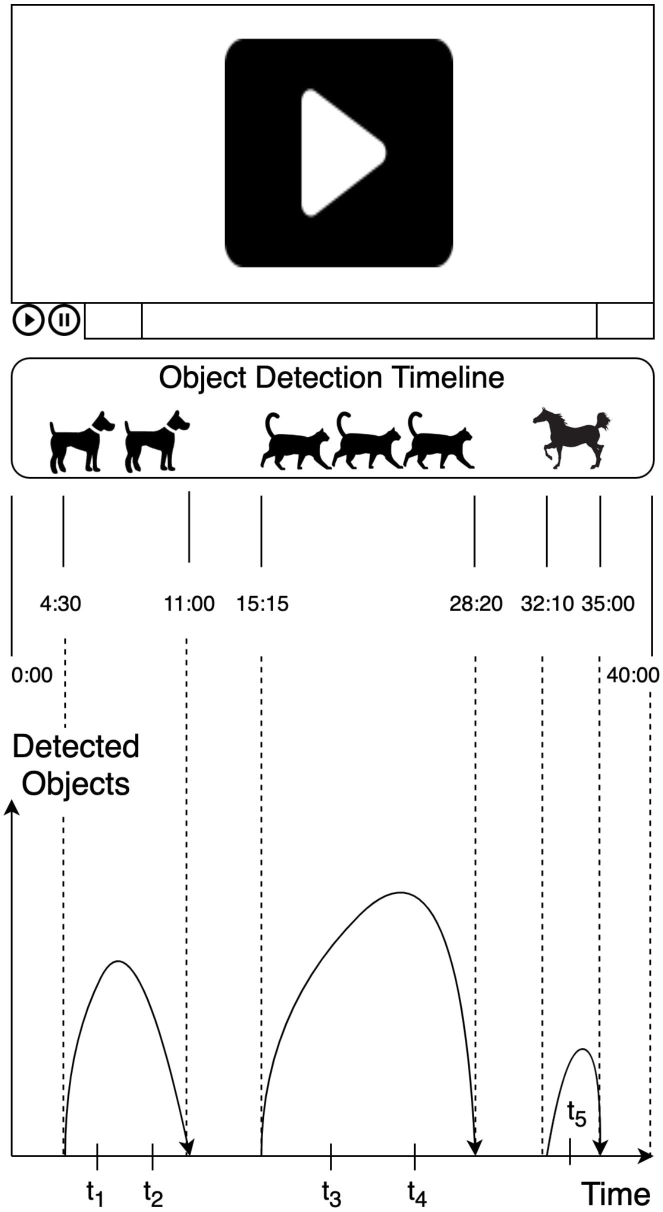

Figure 3. Detected Objects vs Timeline Visual for a video content of frequent type on the web, cat-dog videos. The recognized objects for the video have been provided in a timeline while visualizing the most significant object for the watch duration.

In this section, the rest of the system regarding internal messaging mechanisms between object recognition microservice and the web framework for the video upload dashboard that is targeted for both content and advertisement owner will be presented.

First step, the user (either content owner or advertisement owner) uploads the content using their credentials associated with their account. The content is uploaded through dashboard that is implemented with the Django Python framework. The framework instantiates a containerized object recognition microservice docker image to process the video with the configured machine learning model. Concurrently, wherever applicable, new frames are processed, and JSON data are updated regarding new object recognition information that is determined in Algorithm I using steps 5 and 6. The dashboard is also updated using live timeline visualization patterns. Conclusively, when all reference frames are processed, video JSON data in database are marked as ready for use and the docker container is also terminated. The system is built by using Google App Engine and can execute as many objects recognition microservice docker containers as possible concurrently for different videos. These videos are submitted through the web service using the microservice architecture and is aligned with the cloud resource deployment strategies. Fig 3. visualizes an indexed video content of 40 minutes, where between 4:30 - 11:00 the most significant detected object detected is a dog, between 15:15 – 28:20 is a cat and between 32:10 – 35:00 is a horse. The two-dimensional plot of detected objects vs time function provides an understanding of the contextual data in a graphical fashion presented in Fig 3.

Parallel to the first step of the content owner, the advert owner also uses the same webservice to parse the recognized objects in advertisement content and generate and indexed timeline of detections, which are arranged as JSON output. Fig. 4 represents the indexed advertisement catalogue, where the left hand-side content is a dog biscuit advertisement, middle content is a funny raven video and right hand-side content is cat food advertisement.

Second step, for each video content, timeline priority of the detected tag words is reordered in ascending sorted list. This list provides an understanding for labelling the video from overall point of view. At the end of the second step, the video is parsed and the number of frames including particular objects has been recorded. In order to achieve a better understanding and visualize current comprehension of the particular content, Fig. 3 and Fig 4 can be considered.

Third step is based on matching the sorted tag words in the lists for the detected objects in the video hints possible contextual engagement with relative advertisement content with same detected objects.

The current data for the object recognition have been abstracted by using a frequently available content type on web, a cat-dog video. Though, there are different types of objects detected during the object recognition phase, due to the knowledge of the number of objects counts in a consecutive frame-by-frame analysis. There is a distinctive understanding of the video and quite good chance of finding correct moment of advertisement insertion.

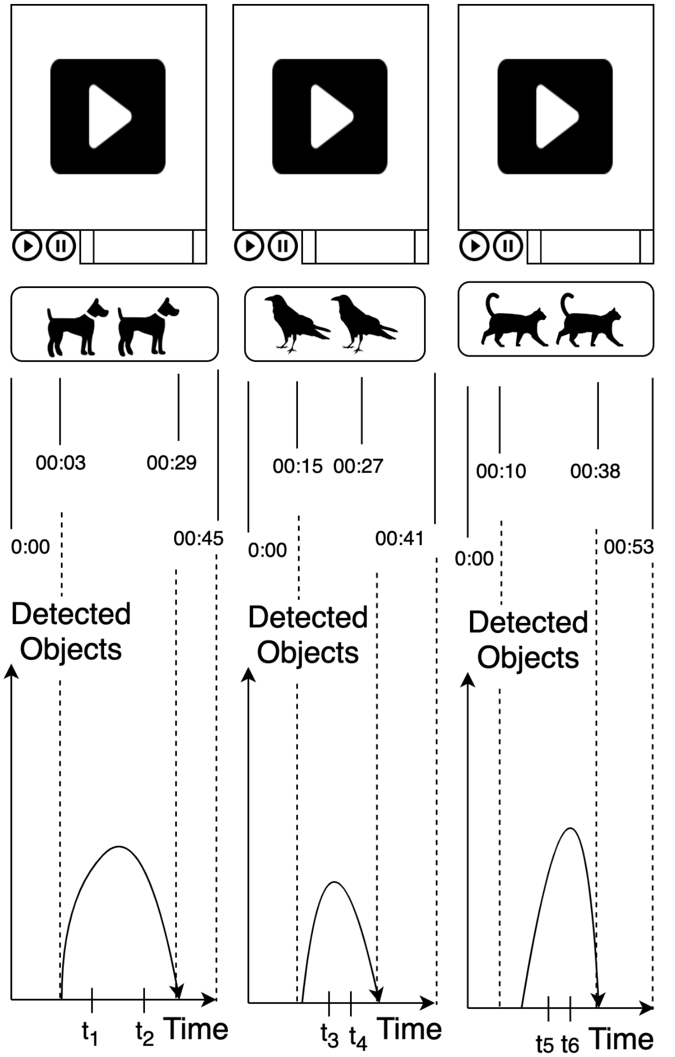

Figure 4. In accordance with video content, advertisement contents are analyzed and an object detection timeline for each advertisement in the catalogue is indexed.

Without loss of generality, in this case the proposed framework aims to insert a dog biscuit advertisement between timestamps t1 and t2 and correspondingly a cat food advertisement between t3 and t4, following the automatic fashion that has been described in the previous section. This corresponds to the most conveying moments as the primary contextual focus of the content and that specific interval is distinctive.

These two well-defined moments are the best candidate for stitching a contextual matching advertisement. Perceptibly, there needs to be other mathematical decision mechanisms about the calculations regarding the number of advertisements that is to be inserted for content and advertisement durations which will be discussed in detail in Section VII.

V. Novelties and Contributions of SuperEye, Smart Advertisement Insertion System

This paper provides a novel technique in the context of smart advertisement insertion for online video broadcasting systems. In order to achieve this an online video indexing platform has been designed and developed. The service is accessible through the url "**https://www.supereye.co.uk**". Below are the brief ideas regarding the approach that is being mentioned throughout the manuscript that can overcome the bottlenecks for today's advertisement insertion mechanism.

1. The methodology is based on the contextual engagement of the video content to the advertisement content. The underlying idea for clarification of a user's interest area is simple, if the user watches the content, advertisement related to content should be shown. Although still compatible and applicable, relying on the track of previous user web searches or video watch history is not required which achieves advertisement insertion in a GDPR safe attitude.
2. With this method, the advertisement can be inserted at the most convenient moment with the most relevant context, because both the video and advertisement context are labelled in a frame-by-frame approach. In terms of orthodox strategies, this can "only" be achieved manually with a human operator even when online video services are considered.
3. Videos can be auto labelled while superseding user assigned limited number of tag words. By using this methodology, in accordance with the content length, videos can easily be labelled by hundreds even thousands of relevant concepts.
4. Frame-by-frame labelling can cluster the videos partially rather than tagging the video as a whole entity.
5. The strategy guarantees to show contextual related advertisement to the requested video content from a frame precise point of view.

VI. Content vs Advertisement duration parameters

Apart from contextual engagement, there are other important parameters related with advertisement and video association, which must be considered to achieve a successful advertisement insertion system.

Some of those parameters are oriented with the duration of advertisement and video or the capability to skip the ad. For instance, showing five different 30s advertisements for a 3-minute video would disturb user watch experience and will be totally unacceptable. In contradictory, stitching a single advertisement for a 2-hour premium content might result in a non-profitable situation for the content provider and also advertisement broker.

In this section, the methodology and mathematical model for finding the most appropriate interval for advertisement insertion is presented. This mathematical model provides a basis for respecting advertisement duration versus video content duration related parameters such as content to advertisement ratio, advertisement duration, user's total watch session duration and number of stitched advertisements.

Regarding conventional online video, the content duration represented by l_c can be classified in three different groups.

Group 1: $l_c < 4min$ content can be short videos, funny clips or music videos.

Group 2 $4min < l_c < 10min$ can be web blogs, news or short movies.

Group 3 $l_c > 10min$ can be classified as movies, series or sports.

All of these varying l_c values require different advertisement insertion approach. For instance, it would be unacceptable to insert advertisement in the middle of a music clip of $l_c < 4min$ type, where the advertisement must be shown prior to the content. In addition to that, for a $l_c > 10min$ content, maximum possible acceptable amount of advertisement must be inserted in order to achieve a profitable service without disturbing customer satisfaction.

The first decision that needs to be taken is to calculate the most convenient number of advertisements for each value of l_c, for each duration type [38]. The number of stitched advertisements is represented by "n" and can be estimated with Eq. 1 and Eq. 2 where the value of n can be computed with linear relation to the content duration, "l_c" following the proprietary but easy to follow brief conventions of YouTube.

$$n = \begin{cases} l_c < 4min\,, 1 \\ 4min < l_c < 10min, N(l_c) \\ l_c > 10min, N(l_c) \end{cases} \tag{1}$$

$$N(l_c) = \kappa\, l_c \tag{2}$$

In order to associate content and advertisement duration related parameters and find the most convenient interval for advertisement insertion,

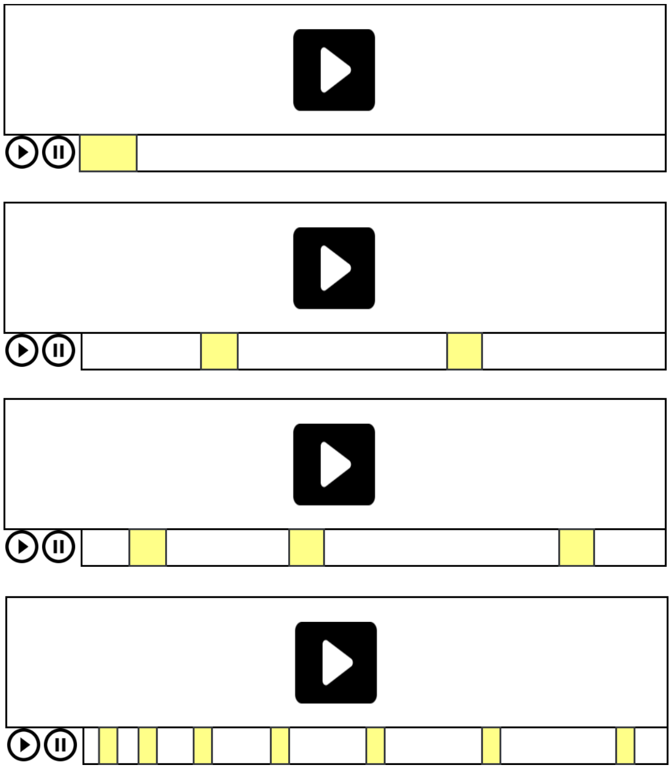

Figure 5. Advertisement insertion calculations using Eq. 3 for different content duration, from top to bottom; 3min, 3-10min, 60min, 60+ min. yellow rectangles along the video content timeline represents the duration of advertisement that has been inserted into the content.

Concurrently, as the L increases, the advertisement insertion becomes less frequent in order to avoid Quality of Experience (QoE) for Advertisement Insertion degradation.

$$for\ each\ n > 1;$$

$$t_{a_i}(n) = \left(\frac{\lambda_i}{n}\sum_{k=0}^{n-1}\frac{l_{a_i}}{T_{max}}\right)\sin\left(\kappa\frac{\phi}{L}\right) \tag{3}$$

Eq. 3 is presented as a sinusoidal function $t_{a_i}(n)$ where l_{a_i} is ith advertisement duration, λ_i is the ratio of the advertisement duration, T_{max} for maximum watch session duration and finally, L refers to the whole watch session experience duration.

The nature of the sinusoidal function provides a repeating behavior where certain frequencies introduce multiple advertisement into the content.

TABLE I
LIST OF NOTATIONS

Notation	Meaning
l_c	Content duration
T_{max}	Maximum watch session duration for a particular user
ϕ	The duration for Skippable advertisement ability
ε	The acceptable duration delta zone for next advertisement insertion
λ_i	The ratio of duration of ith advertisement to the duration of content,
n	Number of advertisements stitched during a watch session
L	Whole watch session experience duration
l_{a_i}	Duration for the ith advertisement inserted to the content
κ	Constant for the sinusoidal function
t_{a_i}	Timestamp representing the moment of ith Ad insertion

VII. PRACTICAL USE CASE SCENARIOS AND EXPERIMENTATION AGAINST SECTOR'S FRONTRUNNERS, YOUTUBE, VIMEO AND DAILYMOTION

This section will present the experimentation methodology that has been employed for evaluating the success of SuperEye, smart object recognition-based advertisement insertion framework. In order to cover a wide variety of video and advertisement content, a catalogue covering an extensive range of genres has been generated. Simultaneously, the content catalogue spans several different durations in order to cover the three groups considered in this study: $l_c < 4min$, $4min < l_c < 10min$, $l_c > 10min$ conditions. From the catalogue, some of the selected content and advertisement video information have been presented in Table I along with genre, duration and primary dominant recognized object throughout the context.

A. Subjects, equipment and test content

The experimentation methodology that is retained in this work, is based on a platform allowing the users to watch online video content through the web service accessing random video contents with their smart devices where advertisements have been inserted into their watch experience regarding the content and watch session parameters. At the end of each video session, users have been provided a survey [39] that consists of questions regarding to the relevance of both duration parameters and contextual engagement of the content versus inserted advertisements.

TABLE II
INFORMATION REGARDING SELECTED VIDEO AND ADVERTISEMENT CONTENT FROM SUPEREYE OBJECT RECOGNITION BASED ADVERTISEMENT INSERTION FRAMEWORK VIDEO CATALOGUE

Video Content	Duration (seconds)	Genre	Most significant recognized object
stonehenge-doc	890	documentary	Rock
thor-tlr2	142	action	People
thetheoryofeverthing	104	biography	People
kedi-doc	2114	drama	Cat
skyfall-tlr2	151	crime	Car
theintern-tlr2	179	comedy	People
independenceday-tlr2	191	science fiction	Spacecraft
applepay	88	advertisement	Phone
bayercat	42	advertisement	Cat
iphone	35	advertisement	Phone
mercedes	74	advertisement	Chicken
messydog	29	advertisement	Dog
samsung	42	advertisement	Television
vodafone	87	advertisement	Dog

Subjects who have participated in the research are undergraduate and postgraduate students attending computer science and data science programmes at London South Bank University at the time of the experimentation. A total of 24 test subjects have participated for the testing evaluation in 3 different 60 minutes sessions. Testers have used 12 different consumer devices including a variety of mobile phones; Samsung S3, S4, S5, Note 3, Note 4, Sony Xperia XZ which have resolution of 1920x1080, HTC 10 (2560x1440) and personal computers; Dell Latitude e6410 (1280×800), Macbook (2560x1600), HP Elitebook8460 (1366x768), Probook 430 (1366x768) where either Firefox or Safari browsers have been executed depending on the operating system of the particular device. All test consumer equipment that has been used via crowdsourcing received service from the proposed video object detection and web services that executes on Google App Engine running via Docker and Django framework on a serverless cloud platform.

In terms of test content, the subjects have been provided a collection of 30 three minutes, 10 three to twenty minutes and 10 twenty plus minutes as video content catalogue. Additionally, a separate catalogue of 30 different publicly available advertisement content ranging from 30 seconds to 2 minutes have been used. Information regarding some of the selected videos and advertisement content has been presented in Table I. All the content can also be accessed through the object recognition portal that is associated with this paper. Relevant access information has been provided in Section IV.

The test subjects have provided their advertisement insertion experience through online questionnaire which have been prepared in alignment with ITU standard P.912 [39] where they have responded about the relevance of the content to advertisement, the suitability of the advertisement insertion moments throughout the watch experience and acceptability of the number of advertisements for a particular content duration.

B. Object Recognition Models

The foundational microservice of this architecture is the object recognition mechanism that has been discussed in Section III. Due to the pipeline-oriented design for the microservice, the object recognition mechanism can benefit from multiple machine learning models including but not limited to Yolo-tiny, YoloV3, DetectNet and GoogleNet, which can be connected to OpenCV through Caffe and Tensorflow backends.

The technical performance of these models varies with respect to couple of parameters, the size of the model, the type of the machine learning model, the number of objects that can be distinguished per each ML model and the efficient implementation of memory, L1, L2 usage, cache miss performance along with GPU support. Object recognition capabilities mainly rely on the number of object classes that can be distinguished. however correct labelling and confidence plays another important factor. Brief information regarding these models and the recognition capabilities are declared in Table III.

TABLE III
OBJECT RECOGNITION MACHINE LEARNING MODEL COMPARISON FOR SUPEREYE OBJECT RECOGNITION BASED ADVERTISEMENT INSERTION FRAMEWORK

Object Recognition ML Model	Size (in MB)	Number of object classes	Frame per second (in ms)
Yolo-Tiny	230	80	9
YoloV3	88	120	18
DetectNet	104	300	78
GoogleNet	180	500	44
DarkNet	240	500	75

The subjects have been provided random video content with varying durations from the catalogue that has been mentioned in Section VIII that provide a watch experience similar to conventional online video services.

For each subject, the experimentation has been repeated using SuperEye while interchanging the object recognition models that has been listed in Table II and for YouTube, Vimeo and DailyMotion.

For each particular watch session, the number of advertisements, the moment for inserting advertisement using Eq. 3 and the relevance of advertisement have been evaluated with the object association mechanism that has been presented in Fig. 2.

The comparison regarding user's evaluation of the success of the online video advertisement insertion system has been provided in Fig. 6 as Quality of Experience for Advertisement Insertion for varying session durations and object recognition models.

C. Comparing Object Recognition Models using SuperEye

The outcomes of the experimentation that has been clarified in Section VIII.C with three different l_c constraints.

Regarding constraint $l_c = 3min$, following a 100-minute watch session, the performance of Yolo model exceeds the performance of the other models DetectNet, GoogleNet, DarkNet and Yolo-Tiny.

Unexpectedly, the underlying reason might be the lower number of object classification capability of Yolo when compared to the other methods except for Yolo-Tiny. Indexing with auto tag words for short videos such as music videos and short web blogs requires advertisement prior to the content and selecting appropriate location during the content is not applicable for this kind of content.

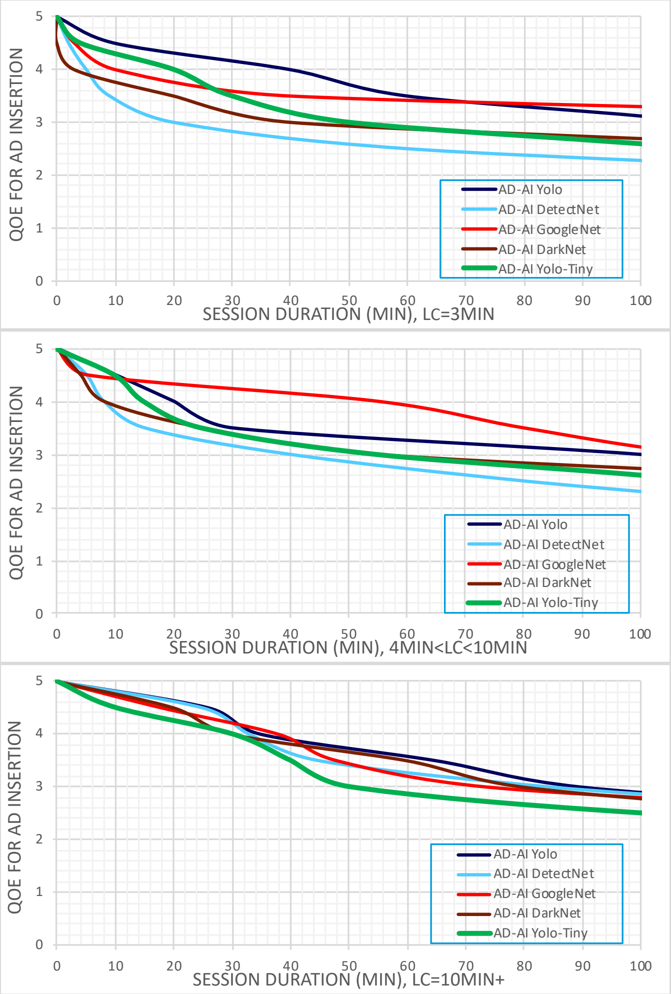

Figure 6. Quality of Experience vs Session Duration for Varying Session Durations using Different Object Recognition Models

For $4min < l_c < 10min$, GoogleNet performance is better when compared to the rest of the models where the QoE first starts to degrade after 60 minutes of experimentation. From the overall point of view, GoogleNet has a slightly better object classification capability than DarkNet and DetectNet, yet for the video catalogue, the same confidence index provided better object recognition rates which might be the motivating reason for the performance. The constraint $l_c > 10min$ provides a watch session experience similar to a single concrete long content such as a conventional movie or a documentary. The advertisement insertion rate is similar to the fourth type in Figure 5 and most of the object recognition methods provide a similar performance while Yolo provides a slightly better QoE especially following the first 40 minutes of the experimentation. Through the outcomes of this constraint, we can say that when the number of the classes for object classification exceeds about 100 distinctive types. Therefore, the object recognition system can provide a similar behavior for each machine learning model.

D. Comparing SuperEye vs Youtube, Vimeo and DailyMotion

Online video broadcasting platforms execute proprietary video analyzing mechanisms for many aspects of the service. It is certain that this technology will be dominantly and inevitably applicable for all online video services. Yet, it is not explicitly clear until which level YouTube or DailyMotion provides object recognition or autonomous classification mechanisms for current online advertisement insertion services.

Nevertheless, comparing the success of YouTube, Vimeo and DailyMotion is still possible by creating the same content video playlist and expect the proprietary mechanisms to insert advertisement using their own algorithms and compare them to the success rate of SuperEye. In this environment, no previous user web search history has been provided to the online video broadcasting system by relying on "incognito mode" of browsers where it is natively supported by major browsing applications such as Firefox, Chrome and Edge.

Under same circumstances, a similar experimentation is executed for each of the video services including SuperEye. The technique for comparing SuperEye versus other online video streaming services has been clarified explicitly in [38] with detail, especially about the playlist generation and comparison strategies. Hence, this concept will not be repeated in while stating that the same strategy has been employed in this manuscript.

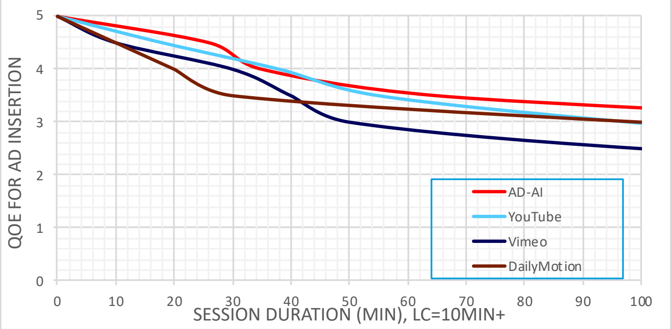

Figure 7.Quality of Experience for Online Video Advertisement Insertion vs Session Duration, for SuperEye, YouTube, Vimeo and DailyMotion

Spanning a total of 4320 minutes of test sessions for each product, the cumulative user experience has been measured in a quantitive approach where Fig. 7 shows moving average for the sessions that has been explicitly mentioned in section VII. The users have evaluated their experience in an online scoring questionnaire out of 5 levels and this data reflects the success rate of advertisement insertions during the watch experience. Due to the nature of QoE surveys that are presented at the end of multimedia applications, such as Skype or WhatsApp [3] calls or YouTube [4] happiness questionnaires, in Fig. 7, the overall results are presented in a function of time vs QoE.

Regarding QoE for online advertisement, SuperEye has provided a better performance against YouTube, Vimeo and DailyMotion for mixed content where $l_c > 10min$ as provided in Fig. 7. The following results have been achieved by estimating the average of the QoE collected from subjective questionnaire with the motivation of [39].

VIII. Conclusions and Future of The Modern Online Advertisement Insertion Systems

In this work, a novel online video advertisement system has been presented using object recognition. Such approach can possibly shape the next decade of the future of video streaming services.

The technology world is evolving in a way that every business becomes a software business. And a significant amount of these software businesses is one way, or another based on or connected with video streaming technologies, including but not limited to social media, news, education, music, cinema and many other forms of entertainment. Regarding the primary origin of the profit for any video streaming business it will not be very difficult to state that the next 10 years of online video streaming will be dominated by smart online video insertion mechanisms. Throughout the manuscript, parameters for achieving a successful online video advertisement insertion mechanism are given from multiple perspectives including both duration and contextual oriented arguments. As vividly given with this work, customer engagement rates for content aware advertisement architectures can easily generate better results in terms of correct audience targeting and increasing both Quality of Experience and profitability for the business. Overall results that have been declared in Section IX provides a good understanding of the future of the online video advertisement and how it will evolve parallel to the achievements in object detection and object recognition.

Conflict of Interest Disclosure

Authors: * Bulkan, U., Dagiuklas T. and Iqbal, M. declare that they have no conflict of interest.

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

[1] Hossfeld, T. et al, “Quantification of YouTube QoE via crowdsourcing”, 2011 IEEE International Symposium on Multimedia, Dana Point CA, USA, 2011.

[2] Wamser, F. et al, “Utilizing buffered YouTube playtime for QoE-oriented scheduling in OFDMA networks”, 24th International Teletraffic Congress (ITC 24), 2012, Krakow, Poland.

[3] Fiedler, M. et al, “A generic quantitative relationship between quality of experience and quality of service”, IEEE Network, Volume: 24, Issue: 2, 2010, USA.

[4] Ketyko, I. et al, “QoE measurement of mobile YouTube video streaming”, MoViD Proceedings of the 3rd workshop on Mobile video delivery, October 2010, Italy.

[5] Bouraqia, K. et al, “Quality of Experience for Streaming Services: Measurements, Challenges and Insights”, IEEE Access, Volume: 8, 2020.

[6] Khan, U. A, et al, “Movie Tags Prediction and Segmentation Using Deep Learning”, IEEE Access, 2020, USA.

[7] Foulsham, M, “Living with the New General Data Protection Regulation (GDPR)”, Financial Compliance, Palgrave Macmillan, Cham, 2019, USA.

[8] Isaak, J. et al, “User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection”, IEEE Computer, 2018, USA.

[9] Osterle, H., “Life with Machine Intelligence”, Springer Nature Life Engineering, 2019, Switzerland.

[10] Marthews, A. et al, “Privacy policy and competition”, Economic Studies at Brookings, December 2019.

[11] Checkley, G., “Video recommendation based on video titles”, Patent No: US10387431B2, 2020, USA.

[12] Liss, B., “Interactive advertising system with tracking of viewer's engagement”, Patent No: US10514823B1, 2020, USA.

[13] Eldering, C. A, “Advertisement insertion techniques for digital video streams”, Patent No: US6704930B1, 2020, USA.

[14] Urban T. et al, “A Study on Subject Data Access in Online Advertising After the GDPR”, Data Privacy Management, Cryptocurrencies and Blockchain Technology, Springer, Cham, 2019, Luxembourg.

[15] Wen, L. et al, “UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking”, Elsevier Computer Vision and Image Understanding, 2020, USA.

[16] Gong, W. et al, “Visual signal representation for fast and robust object recognition”, IEEE 18th European Control Conference(ECC), 2019, Italy.

[17] Wang, X., “Supervised Learning for Data Classification Based Object Recognition”, Machine Learning-based Natural Scene Recognition for Mobile Robot Localization in An Unknown Environment, Springer, 2020, Singapore.

[18] Zhaoping, L., “Artificial and Natural Intelligence: From Invention to Discovery”, ScienceDirect Neuron, 2019, Beijing.

[19] Guo, S. et al, “Foreign Object Detection of Transmission Lines Based on Faster R-CNN”, Springer Information Science and Applications pp 269-275, 2019.

[20] Wu, X. et al, “SVM-based image partitioning for vision recognition of AGV guide paths under complex illumination conditions”, Robotics and Computer-Integrated Manufacturing, 2019, China.

[21] Hayakawa, Y. et al, “Feature Extraction of Video Using Artificial Neural Network”, Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications, 2020, Japan.

[22] Harijanto, B. et al, “Recognition of the character on the map captured by the camera using k-nearest neighbor”, IOP Conference Series: Materials Science and Engineering, 2020, Japan.

[23] Çambay, V. Y. et al, “Object Detection on FPGAs and GPUs by Using Accelerated Deep Learning”, 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, Malatya, Turkey.

[24] Bryan, R. N et al, “Medical Image Analysis: Human and Machine”, Academic Radiology Special Review Volume 27, Issue 1, P76-8, 2020, USA.

[25] Leon, V. et al, “A TensorFlow Extension Framework for Optimized Generation of Hardware CNN Inference Engines”, Special Issue MOCAST: Modern Circuits and Systems Technologies on Electronics, 2019, Athens.

[26] Fu, J. et al, “Contextual deconvolution network for semantic segmentation”, Pattern Recognition, Volume 101, 2020, China.

[27] Susmitha, A. V. V, “Smart Recognition System for Business Predictions (You Only Look Once – V3) Unified, Real-Time Object Detection”, Springer Internet of Things for Industry 4.0 pp 137-146, 2019, India.

[28] Mohamed, K. S., “Parallel Computing: OpenMP, MPI, and CUDA”, Springer Neuromorphic Computing and Beyond pp 63-93, 2020, Egypt.

[29] Dutta, S. et al, “Crowd Behavior Analysis and Alert System Using Image Processing”, Emerging Technology in Modelling and Graphics pp 721-729, 2019, India.

[30] Duan, R. et al, “Object Recognition and Localization Base on Binocular Vision”, IGTA Image and Graphics Technologies and Applications pp 300-309, 2019, China.

[31] Kustikova, V. et al, “Intel Distribution of OpenVINO Toolkit: A Case Study of Semantic Segmentation”, AIST 2019: Analysis of Images, Social Networks and Texts pp 11-23, 2019, Russia.

[32] Apatean, A et al, “An Intelligent Eye-Detection Based, Vocal E-Book Reader for The Internet Of Things”, Acta Technica Napocensis Electronics and Telecommunications, Volume 57, Number 2, 2016, Romania.

[33] Ahad, M. A. et al, “Handling Small Size Files in Hadoop: Challenges, Opportunities, and Review”, Soft Computing in Data Analytics pp 653-663, 2018, India.

[34] Dharavath, R. et al, “Capturing Anomalies of Cassandra Performance with Increase in Data Volume: A NoSQL Analytical Approach”, Advances in Data Science and Management pp 3-20, 2020, India.

[35] Licciardello, M. et al, “Understanding video streaming algorithms in the wild”, Networking and Internet Architecture, 2020, Germany.

[36] Pathan, M. et al, “Content Delivery Networks: State of the Art, Insights, and Imperatives”, Content Delivery Networks pp 3-32, 2020, India.

[37] Qian, X. et al, “Object Detection in Remote Sensing Images Based on Improved Bounding Box Regression and Multi-Level Features Fusion”, Image Processing and Spatial Neighborhoods for Remote Sensing Data Analysis, 2020, China.

[38] Bulkan, U. et al, “Modelling Quality of Experience for Online Video Advertisement Insertion”, IEEE Transactions on Broadcasting, 2020, United Kingdom.

[39] ITU-T, Telecommunication Standardization Sector OF ITU, P.912 (03-2016) “Subjective video quality assessment methods for recognition tasks”, “Series P: Terminals and Subjective And Objective Assessment Methods”, Audiovisual quality in multimedia services. (online) https://www.itu.int/rec/T-REC-P.912