Received: February 1, 2021. Revised: March 10, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Multi-Object Semantic Video Detection and Indexing Using a 3D Deep Learning Model

Eslam Mofreh[1]* Amr Abozeid[1,2] Hesham Farouk[3] Kamal A. ElDahshan[1]

[1]*Mathematics Department, Faculty of Sciences, Al-Azhar University, Cairo, Egypt*
[2]*Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Saudi Arabia*
[3]*Computers and Systems Department, Electronic Research Institute, Cairo, Egypt*
* Corresponding author's Email: eslammofreh@azhar.edu.eg

Abstract: Within the exponential growth in raw data production, attributed in no small part to social media - Facebook, Youtube, and others. Video is proving to be the most important data type thanks to the substantial amount of raw data it contains, requiring an efficient way to be understood, organized, structured, and stored for ease of retrieval. Hence, an efficient video indexing architecture is thus crucial for video datasets. This paper proposes an efficient Multi-Object Semantic Video Detection (MOSD) that leverages the deep learning power to achieve effective indexing on the semantic concept level. MOSD is multi-detection network of video semantics in multiple frames. MOSD exploits a 3D convolution operation to do multiple detections among multiple frames with higher performance. The detected semantics then structured and used for indexing the video segments. MOSD has been trained and evaluated on ImageNet VID dataset and has been compared to peers. MOSD showed efficiency in exploiting the temporal context of a video to do simultaneous detections of consecutive frames which speeds up the detection of semantic objects. MOSD also showed performance efficiency in terms of mAP which is 85.2%.

Keywords: Deep learning, Video indexing, Sematic video indexing, Video object detection, 3D convolution.

1. Introduction

Nowadays, video is the primary as well as the comprehensive medium for the exchange of information. It is also widely used in several significant domains such as education, surveillance, entertainment, medicine, and others [1]. The ubiquity of smart devices, advances in processing power, and markedly improving internet connection have supported the rapid spread of video data. Moreover, the video characterizes with [2]:

1- Non-defined prior structure.
2- Rich in raw data.
3- Repetitive nature of the frames.
4- Needs large storage capacity.
5- Include video formation issues, e.g., viewpoint change, illumination variation, motion blur, occlusion, etc.

For these reasons, video is considered the most important multimedia type, and structuring its data for effective storage is a critical concern [1]. Over the past decade or so, considerable efforts have enriched the literature with significant breadths covered thanks to deep learning paradigms. The Convolutional Neural Network (CNN) has achieved exceptional results which have attracted the attention of researchers worldwide. In 2010, the Large Scale Visual Recognition Challenge (ILSVRC) had launched and deep learning paradigms have widely spread since AlexNet, 2012 winner, is considered as the first considerable work introduced to CNN and since then, the number of research contributions in deep learning architectures in real-world problems has been rapidly growing. Deep learning architectures have also achieved exceptional success in the context of video as well.

The process of managing and organizing video datasets becomes crucial. The key concern is that for the large amount of video datasets, the process of manually annotating is no longer being an effective

way as it is very time-consuming[1]. One of the most important computer vision topics is video indexing. Video indexing is a way to assign an index for the video segment for effective organization of the video dataset and effective retrieval.

Semantic Video Indexing (Semantic-VID) is the process of exploring a set of expressive semantic concepts of video frames and assigning it to the video [3]. Semantic-VID output depends on the level of indexing that the model seeks to achieve. At the highest level of indexing, there are both the video label and/or the video shots' labels. The intermediate level has more semantic depth of the video; it comprises a set of objects, actions, and activities. The lowest level is the most representative of the video. It represents the video by a dense representation of annotations and captions.

VID comprises basic subtasks such as video classification, object detection, action/activity recognition, and so on. It may comprises other tasks (e.g., facial recognition) depending on the application. VID architectures can roughly be classified into two main sub-categories: 1) Conventional (Handcrafted based) architectures and 2) Deep learning-based architectures.

The difference lies in the features and how they are extracted to ultimately reach the video index. Deep learning techniques are efficient in extracting the semantic concepts in one step unlike conventional ones which extract the video features and then annotate the frames using the extracted features.

This paper introduces an effective deep learning-based semantic video indexing architecture. Multi-Object Semantic Video Detection (MOSD) is mainly concerned with detecting multiple semantics from multi-frames simultaneously. MOSD uses a Convolutional Neural Network (CNN) in abstracting the video frames' features in different levels. It forms a paramedical feature representation of the frames' features. This paramedical representation is then used to detect the semantic categories from the frames. MOSD utilizes a 3D convolution to do multiple detections among multiple frames simultaneously. It trained and evaluated on the ILSVRC 2015 VID dataset [4].

The paper is organized as follows: Section 2 covers the related work of video indexing, while Sections 3 and 4, respectively, introduce a newly proposed MOSD model and the implementation details. Section 5 provides a dataset overview, and Section 6 introduces the experimental results and the results discussion. The paper ends with its conclusion and acknowledgments.

2. Related work

Video Indexing (VID) is the process of generating an expressive index that describes the video segment for efficient storage and retrieval purposes[2]. As indicated in figure 1, video indexing architectures can be divided into two categories: features-based and semantic-based.

Video Indexing architectures can also be divided into two categories: conventional methods and deep learning methods. Conventional video indexing methods create an index based on the video's high-level features. Using conventional machine learning techniques like SVM, several of these methods attempt to bridge the gap between a video's features and its semantics. While Deep Learning-based approaches derive video semantics in a single step by extracting features and classifying them into semantic classes.

2.1 Features based Indexing

Feature based indexing is a method that utilizes high-level features in indexing video segments. Histogram of Oriented Gradients (HOG) detector is one of the most important features-based indexing architectures [N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886-893 Vol. 1, doi: 10.1109/CVPR.2005.177.5]. HOG is an example of feature-based methods which works by counting the occurrence of gradient orientation of an image, and describes the image object appearance and shape by intensity gradient and edge direction. Another important feature is the Scale–Invariant Feature Transform (SIFT). SIFT describes high-level features of an object, which locates certain interest points and assigns an invariant feature to them regardless the image scale, noise, or illumination. After calculating and storing SIFT features reference images, the new SIFT is calculated and compared with the stored ones and therefore identifying key points in the new to filter out the best matches. Abozaid et al. proposed a Global Dominant SIFT (GD-SIFT) descriptor for video indexing and retrieval [1]. These approaches have a gap between the extracted high-level features and the semantics, making them unsuitable for semantic video indexing. They lack an interpretation from the human perspective of the semantic index, as they rely on the video's features rather than its semantics (such as person and car).

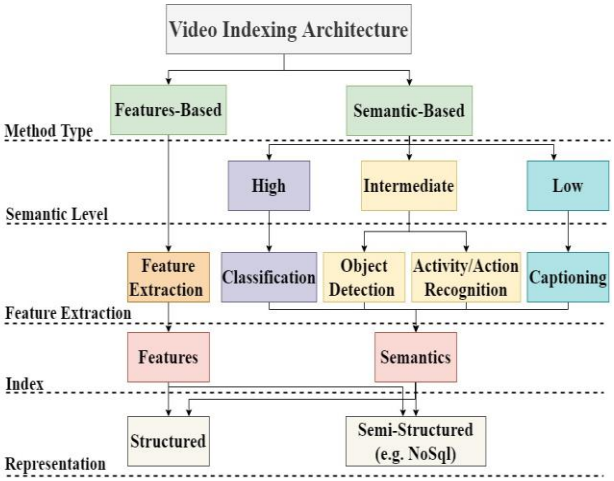


Figure. 1 Video indexing architecture classification

2.2 Semantic-based indexing:

Semantics based indexing is video indexing by using a set of meaningful semantic concepts [6]. The essence of semantic VID lies in understanding the video context from a human perspective and structuring it in an efficient manner for easy retrieval purposes. Semantics concepts can roughly be classified into three levels:

i. High level: to extract single semantic label for a video and/or semantic labels for its shots.
ii. Intermediate level: to extract semantic objects, actions, or activities out of the video.
iii. Low level: to extract a dense representation of the semantic concepts (e.g., annotations and captions).

Semantic index extraction undergoes with set of tasks which are video classification, video object detection (VOD)/Recognition, video actions/events recognition, and video captioning / annotating. The literature has a lot of investigation regarding these tasks and it had a great evolution path, especially in the Deep Learning era. Convolutional Neural Network (CNN) introduces amazing architectures to semantically detect the semantic concept in a single step instead of the conventional architectures taking multiple stages to extract the semantics. Ghatak and Bhattacharjee proposed a multi-stage video indexing technique based on the Viola Jones algorithm [7], although others rely on the CNN model, which only requires a single step of semantic detection [6, 8, 9].

Semantic indexing of a video is one of three main levels: 1) High-Level Semantic Index, 2) Intermediate Level Semantic Index, and 3) Low-Level Semantic Index.

2.2.1. Extraction of high level semantics:

Picking an indexing task depends mainly on the video dataset nature which may have a very diverse content in which a single semantic label is quite enough to describe the content. In other cases this is not adequate to efficiently index the video and higher semantic levels, are more preferable.

Video classification is the process of assigning a semantic label that is relevant to the content of the video segment. A good semantic label is the one that best describe the video content. This is the most basic indexing strategy and the simplest one.

Kumar et al. proposed a classification model based on automated comprehension of human motions for obtaining semantic labels from sports' videos [10]. Savran et al proposed a CNN-RNN model for extraction of semantic labels of the video [11]. Their architecture utilized both spatial and temporal information of the video for better classification results.

High-level indexing techniques are appropriate for video datasets with highly diverse content (such as one clip of a match, another one for a school lecture, another for traffic, etc.). For retrieval purposes, indexing a video with a single label would suffice in these cases. However, in real-world datasets with substantially more similar content, more indexing layers are required. As a result, intermediate and low levels have been come into place to go deeply into the content and characterize the video using semantics like objects, events, and/or activities.

2.2.2. Extraction of intermediate-level semantics

More sophisticated intermediate levels of indexing are usually useful for efficient describing most of the datasets. Intermediate levels of video index comprise of the extraction of semantic objects, actions, or activities out of a video. Video object detection and action/activity recognition attracted the researchers' interest to the deep learning evolution. Undoubtedly, these evolutions added a lot to video indexing research. As shown in figure-1, extraction methods of intermediate-level Features' tasks can be divided into: Video Object Detection and Video Action / Activity Recognition.

Video Object Detection (VOD): It is a challenging task which revolves around detecting both semantic object category and location. Two main tasks are used in detecting the semantics objects; object classification and localization. It is roughly divided into two main types: Direct Detector and Multi-Stage Detector.

In Multi-Stage Detector, a set of regions are proposed, and then each region is classified using a CNN architecture. Girshick et al. have proposed RCNN which uses selective search as a region

proposal, and then a pre-trained CNN architecture (such as VGG) is then fine–tuned to classify the regions to semantic objects [12]. RCCN boosts performance of object detection with mean average precision (mAP) with 58.5%. K. He and X. Zhang have proposed Spatial Pyramid Pooling Networks (SPPNet) which mainly provides Spatial Pyramid Pooling (SPP) that generates a fixed length representation of regions without rescaling the regions [13]. SPPNet improves speed without sacrificing accuracy, as it has a mAP of 59.2%. Girshick has proposed a Fast RCNN which unifies the three modules used by RCNN into one [14]. Rather than depending on selective search and an edge box, S. Ren et al. have proposed a Faster RCNN which integrates Region Proposal Network (RPN) into a CNN [15].

In Direct Detector one model is used for both classification and bounding box regression. Szegedy et al. was the first contributor to propose DetectorNet which treats object detection as a regression problem. He used AlexNet and replaced the Softmax layer with a regression one [16]. Sermanet et al. have proposed the OverFeat model which is a single-stage object detection one, where both classification and localization are achieved simultaneously [17]. The first effort to develop a real-time object detector was the You Only Look Once (YOLO) network. Many enhancements were introduced to it, and it evolved till Yolov4 and Yolov5 [18]. Liu et al. have proposed Single Shot MultiBox Detector (SSD) with its pyramidal hierarchy when extracting features with a CNN and that improves the detection results [19]. Kang et al. have introduced the T-CNN model as an example of methods that work on bounding box level using precomputed optical flow fields and object tracking to propagate bounding boxes to nearby frames [20]. Han et al. have built Seq-NMS that improved detection by utilizing high score detection from nearby frames [21]. B. Hatem et al. have proposed Seq-Bbox which built tubelets by linking b-boxes across frames to improve detection [22]. Chen et al. proposed GigaDet, an object detection model [23]. The proposed model is composed of a patch generation network (PGN) which is used to discover feasible areas holding objects and decide the optimal resize ratio of each patch. Then the generated patches are then used by a decorated detector (DecDet) to perform detection.

While methods working on feature level are an improvement, Zhu et al. have introduced Flow-Guided Feature Aggregation (FGFA) model which uses optical flow and features extracted from nearby frames for improving detection [24]. Feichtenhoter et al have proposed both Detect-to-Track and Track-to-Detect (D&T) models which are used simultaneously for detection and tracking [25]. Yuning Chai has proposed a Patch-Work model for detecting objects from a video by using specialized memory that retrieves lost context [26]. Patchwork adopts Q-learning based policy that intelligently selects sub-windows to be treated in subsequent frames. Fujitake and Sugimoto proposed a video object detection method based on the Generative Adversarial Network (GAN) to accomplish identification and content synthesis [27]. The architecture utilized an encoder-decoder network that decoded the encoded features one at a time. The encoder and decoder are recurrent encoders and decoders in the network. ResNet and Feature Pyramid Network (FPN) are used to create the encoder and decoder, respectively. This model achieved a mAP of 73.1%.

In detection and segmentation-based tasks, Region of Interest Align, or RoIAlign, is a method for obtaining a small feature map from each RoI which is firstly proposed by He et al [28]. RoIAlign computes the precise values of the input features at four regularly sampled locations in each RoI bin using bilinear interpolation, and the result is then aggregated using max or average. However, RoI Align, continues to extract features from a single-frame feature map for proposals, resulting in derived RoI features that lack temporal information from movies.

Gong et al. proposed Temporal RoI Align, which is an enhanced version of the RoIAlign that exploits the video's temporal information [29]. Temporal RoI Align works by firstly extract the RoI features from the target frame. Then, for target frame proposals, Most Similar RoI Align automatically collects the most similar RoI features from support frames feature maps. Then, in order to create final temporal RoI features, a temporal attention mechanism is used to aggregate the RoI features and the most similar RoI features. Temporal RoI Align succeeded in incorporating the temporal information of a video however, it still lacks consideration of different scales of the object to further enhance the detection through the temporal multi-scaled representation of the object.

Video Action/Activity Recognition: Video actions and activity recognition to get higher indexing levels are crucial for building a robust index. An object action produces a video, while one or more actions in a given period of time produce an event, which makes video actions and events similar concepts. It is crucial to effectively use both video temporal and context information without loss of information. There are three categories of models used for action recognition: i) spatiotemporal

networks; ii) temporal coherency networks; and iii) multiple stream networks.

i) Spatio-temporal networks

In terms of video, temporal information should be considered for action recognition. Ng et al. (2015) proposed the temporal pooling and found that max temporal pooling was more beneficial [30]. Varol et al. (2016) investigated the enhanced effect when increasing the temporal duration of the input and combining the results of different temporal durations of video, since by adding the temporal dimension, the parameters get increased and this will affect the 3D convolution operation performance [31]. Yang and Zou proposed a deep learning network model based on spatiotemporal features fusion (FSTFN). Both the spatial and temporal information are utilized through composition of two networks composed of CNN and LSTM [32].

ii) Temporal coherency networks

The concept of temporal coherency of a video is that each of the consecutive set of frames is semantically and dynamically coherent. A video is said to be coherent if:

- The video frames are in their appropriate temporal order
- The video events semantics are correlated
- There are no abrupt changes in event semantics or motions

Misra et al. (2016) investigated temporal coherency in learning visual representations of video for an action recognition task [33]. Fernando and Gould (2016) suggest an end-to-end learning scheme that learns both the pooling operation and the classifier with back propagation [34].

iii) Multiple stream networks

This type of network is inspired by the human visual cortex. The visual cortex has two streams; Ventral and Dorsal. The Ventral stream identifies the object identity, color and appearance, while the Dorsal stream recognizes the motion of the object. Simonyan and Zisserman have devised an architecture that exploits both appearance and motion (spatial and temporal) information [35]. They have built a spatial stream network trained by video frames and a temporal stream network trained by optical flow fields.

Object detection and activity/action recognition are effective and appropriate techniques to semantically describe a video at a lower level of semantics, better representing the video content than high-level approaches. These approaches, on the other hand, do not use video contextual and temporal information to index multiple video frames simultaneously. Furthermore, they haven't considered the semantics at different scales, which might boost performance. Therefore, it is critical to consider both temporal and contextual information, as well as different scale representations of the semantics.

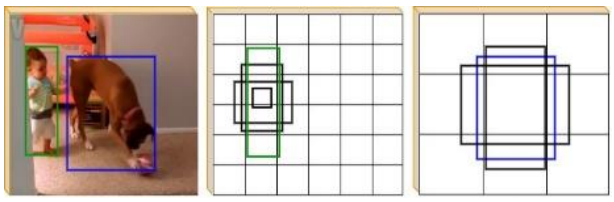
2.2.3. Extraction of low level semantics

In most cases, we need to determine a robust index which densely describes the content of the video (e.g., a man is riding a bicycle and a child playing a football) to densely describe a video segment. This would be more expressive and robust than the other two types. Jesus et al. review video captioning, which is the process of assigning a textual description to a video input [36]. Hemalatha et al. utilize a 2D and 3D CNN network to extract features to identify the domain, and the video domain beside a RNN network to generate the video captions [36]. Wanting et al. proposed an attention-based dual learning strategy (ADL) [37]. ADL is made up of two modules: a caption generation module that creates a reversible mapping between a video and its caption, and a video reconstruction module that uses the video captions to recreate the video frames. Vaidya et al. proposed a low-level semantic extraction architecture for caption generation. This architecture searched for the semantics as persons and objects, then the semantics information is aggregated over the video frames [38].

The lowest level of semantics extraction is to assign captions (such as a person is crossing the street) for video frames for indexing purposes. However, as video content becomes larger and the content changes rapidly, retrieval would be challenging when using such methods. Thus, these methods are ideal for small video clips datasets rather than the bigger ones because of the storage and retrieval constraints.

2.3 Video index representation

Video indexing ends by structuring the generated semantics in a type of structure for an efficient retrieval. It is crucial to have a predefined structure to



(a) (b) (c)

Figure. 2 MOSD Model: (a) Ground truth bounding boxes, (b) (4x4) Feature map, and (c) (3x3) Feature map. MOSD detects at different feature map scales (exactly, (3x3) and (4x4))

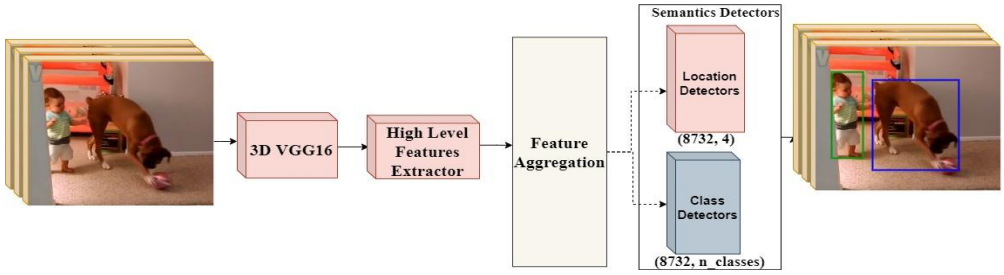


Figure. 3 Multi-object semantic video detection (MOSD). It comprises three main stages: (1) Base feature extractor (3D VGG16), (2) High-level features extractor, (3) Semantics detector

organize the generated semantics in effective way for ease retrieval. The video index is being structured and stored by using Structured Index Store or Semi-Structured Index Store. Structured Index Store: it is about structuring the semantics and storing it in a predefined static data model like relational database management systems, while Semi-Structured Index Store uses a semi-structured dynamic data model for storage such as NoSQL database (document-based, key-value, graphs, etc.).

Porter et al. represented a shot by a directed weighted graph in which its nodes represent the video semantic concepts and edges define the dissimilarities between each node semantic [39]. Podlesnaya et al. have – inspired by ImageNet – introduced a method of building an index by using graph databases, based on the WordNet lexical database [8]. Truong et al. structure and store the video index in a file [40].

3. Materials and methods

As the Deep Learning architectures evolve rapidly with time to tackle many challenges, this paper proposes the Multi-Object Semantic Video Detection (MOSD) model which is a semantic deep learning VID one. Unlike intermediate-level video indexing methods, MOSD considers both contextual and temporal information through a 3D convolution operation. In addition to, it considers a multi-scales representation of a semantic object in a video.

MOSD is characterized by extracting multiple semantic objects from different frames simultaneously to form a robust video index. It is an adapted 3D model from the 2D SSD object detection model [19]. Thus, it is much faster than the 2D object network since it can process a total of roughly 185 FPS.

The MOSD model is a CNN based network that outputs a collection of bounding boxes and confidence scores for the presence of semantic objects. These detection collections are followed by a non-max suppression to produce the final predictions. For each input frame, a set of feature maps is produced. Each one produces a set of detections using a set of convolutional filters. For a feature map of size *h, w,* and *c* channels per *i* input depth (e.g., 3 frames). A *3x3xc* kernel for each *i* input image is used to produce a score and 4 offsets relative to the default box. For each feature map, a set of anchor boxes of different scales (e.g. *4x4* and *3x3*) and aspect ratios are generated for each frame from the input ones. For each generated default box, offsets and semantic category scores are predicted. The default boxes are matched against the ground truth boxes. For each semantic object, some of the default boxes are identified as positive and the remaining are negative as showed in Fig. 2.

As shown in Fig. 3, MOSD model consists of three main phases:

1) Base feature extractor (3D VGG16)
2) High-level feature extractor: for higher-level features
3) Semantics detector: for multi-scale prediction

3.1 Base features extractor:

MOSD model is a 3D version of the 2D VGG16 to form the network backbone. A 3D convolution adds an extra dimension to the extraction process with a little increase of parameters. It utilizes a 3D convolution layer in the CNN to exploit the video temporal dimension. A stream of video frames passes the network to extract the features out of them. These features are then classified to semantic categories. Essentially, a 3D convolution is the same as the 2D one, but the kernel moves in the 3D convolution causing a better feature capture for multiple frames. The 3D convolution layer is more robust in detecting global/local features of consecutive frames simultaneously. The adapted 3D VGG16 architecture consists of 6 blocks, as shown in figure-4. Each block consists of multiple 3D convolutions and 3D pooling layers. These convolutions are Conv1, Conv2, Conv3, Conv4, Conv5, Conv6 and Conv7.

3.2 High-level features extractor:

More abstraction levels of the features are needed to form a robust feature map that reflects the features

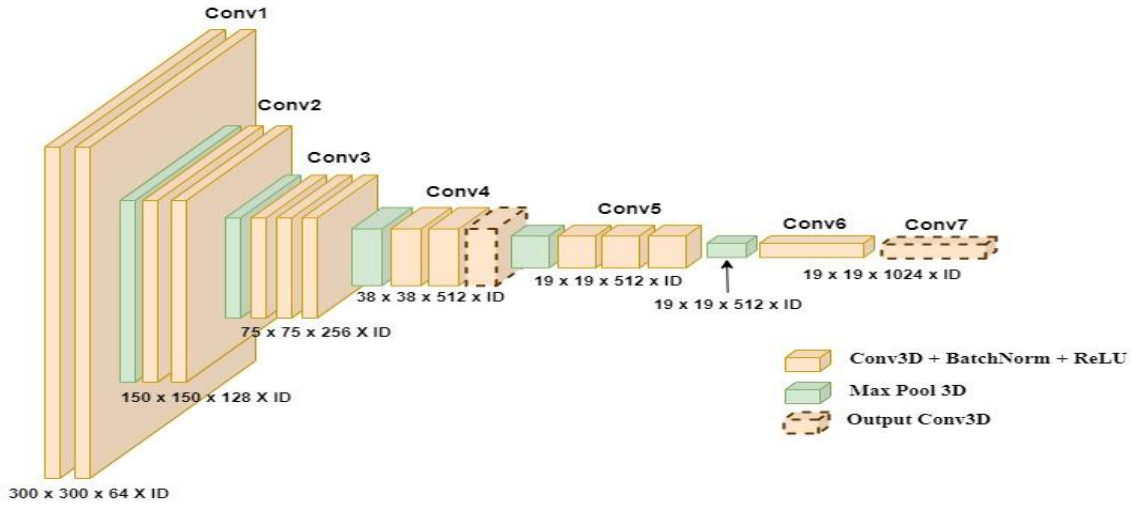


Figure. 4 MOSD base feature extractor (3D VGG16)

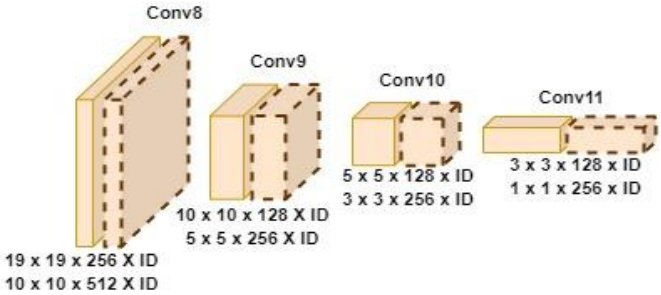


Figure. 2 MOSD high level features extractor

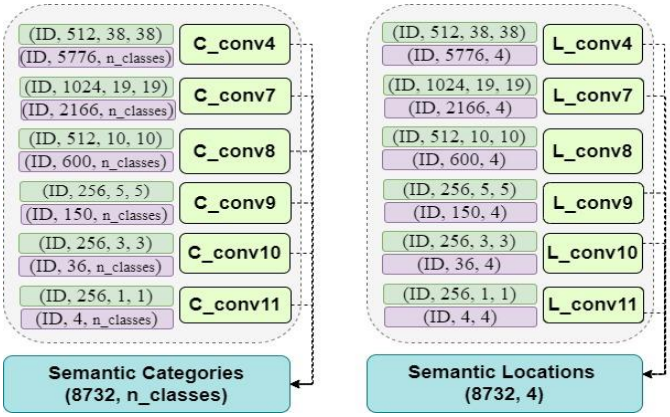


Figure. 3 MOSD semantic detector

in different scales. MOSD makes use of a high-level feature extractor that comprises 4 layers each and each has 2 convolutional operations. These layers use different kernel sizes to down-sample the base features into different smaller scales. All pairs of the convolution operations are applied with two different kernel sizes of (3x3 and 1x1) as shown in Fig. 5. These feature maps are useful for detecting tiny semantic objects as well as bigger ones. The generated feature maps are then passed to the prediction layers to detect the semantics' categories and locations.

3.3 Semantics detector:

The prediction layers are the MOSD final stages of semantic object detection. As illustrated in figure-6, the prediction is based on six distinct scales of the backbone's produced features and higher-level features. These six output feature map scales are produced by the dotted convolutions that are showed in Fig. 4 and 5. Two separate scales of base features, which are the result of Conv4 layer and Conv7's third convolution operation, are combined with four scales of high-level features. The second convolution of

Conv8, Conv9, Conv10, and Conv11 produces the four scales of high-level features. The six feature map scales are fed to six convolutions for semantic categories detection and another six for object localization. Each layer of the convolutional layer is specialized in detection of semantics of different sizes. The final output of the convolutional layers is the semantic categories and locations which are (8732 x number of classes) semantic categories and (8732 x 4) semantic locations.

3.4 Model training:

Regarding model training, MOSD utilizes the concept of default boxes and MultiBox loss function. During training, the default box that best matches the ground truth box needs to be determined. MOSD generates default boxes on top of the model structure which are varying in scale and aspect ratio. These default boxes are matched to the ground truth box to determine the best matching one. The best matching default box is the one that has the best Jaccard overlap. The default boxes that have a Jaccard overlap with a value higher than a threshold (0.5) are selected.

Jaccard Index: It is a metric that is used to quantify the overlap percentage between the default box and the ground truth box. It measures the intersection over union between them as presented in Eq. (1).

$$Jaccard\ Index\ (JI) = \frac{db_i \cap gt_j}{db_i \cup gt_j} \tag{1}$$

Where db_i is the i^{th} default box for $i \in$ set of default boxes, and gt_j is the j^{th} ground truth box for $j \in$ set of ground truth boxes.

Default Boxes: the MOSD uses multiple default boxes that vary in their location, scales, and aspect ratios. The used aspect ratios are $\{1, 2, 3, \frac{1}{3}, \frac{1}{2}\}$ and the scale for each feature map is calculated by Eq. (2).

$$s_i = s_{min} + \frac{s_{max} - s_{min}}{m-1}\ (i-1) \tag{2}$$

Where s_i represents the scale for the i^{th} feature map, $i \in [1, m]$, m is the number of feature maps used for prediction, $s_{min} = 0.2$ and $s_{max} = 0.9$, which means that the minimum scale is 0.2 and the maximum is 0.9. The width and height of the default box are calculated by Eqs. (3) and (4).

$$w = s_i . \sqrt{AR} \tag{3}$$

$$h = \frac{s_i}{\sqrt{AR}} \tag{4}$$

Where w and h are the default box width and height, respectively, and AR is the aspect ratio. We now have five default boxes for each location in the feature map. Another 6th default box with a scale $s'_i = \sqrt{s_i . s_{i+1}}$ for the aspect ratio = 1.

Loss Function: The objective function of the MOSD model is the weighted sum of both semantic confidence and localization loss. Let $p^c_{fij} = \{0,1\}$ be an indicator for matching the i^{th} element of ground truth with the j^{th} element of default box for category c in the frame f of the input frame. The sum of localization losses for all input frames. For each input frame, it is calculated by using the *smooth_l1* between the predicted box b and the ground truth g as defined by Eq. (5). The *smooth_l1* used quantifies the difference between the predicted box b and ground truth box g parameters *(cx, cy, h, w)* for each frame f in the input frames, for each of the positive default boxes.

The sum of semantic confidence losses for all input frames. For each input frame, it is defined to be the Softmax loss over multiple class confidences t which is defined by Eq. (6).

Where p^p_{fij} is an indicator for the i^{th} element of ground truth with the j^{th} element of default box for category p in the frame f of the input frame, $\hat{c}^p_{fi} = \frac{exp(c^p_{fi})}{\sum_p c^p_{fi}}$, g is the ground truth, b is predicted box, d is the default box and *ID* is the input depth of the model.

$$\boldsymbol{L_{localization}} = \sum_{f \in ID} \sum_{i \in Pos}^{N} \sum_{m \in \{cx,cy,w,h\}} p^k_{fij}\ SMOOTH\_L1\left(b^m_{fi} - \hat{g}^m_{fj}\right) \tag{5}$$

$$\boldsymbol{L_{confidence}} = \sum_{f \in ID} \left(-\sum_{i \in Pos}^{N} p^p_{fij}\ log\left(\hat{c}^p_{fi}\right) - \sum_{i \in Neg} log\left(\hat{c}^0_{fi}\right)\right) \tag{6}$$

The bounding box regression is conducted for the offsets of the center (cx, cy) of the default box and its width w and height h. Hence, the ground truth coordinates are transformed according to Eqs. (7.1) to (7.4).

$$\hat{g}^{cx}_j = log\left(\frac{g^{cx}_j - d^{cx}_i}{d^w_i}\right) \tag{7}$$

$$\hat{g}^{cy}_j = log\left(\frac{g^{cy}_j - d^{cy}_i}{d^h_i}\right)$$

$$\hat{g}_j^w = log(\frac{g_j^w}{d_i^w})$$

$$\hat{g}_j^h = log(\frac{g_j^h}{d_i^h})$$

Where $\hat{g}_j^{cx}, \hat{g}_j^{cy}$ are the j^{th} transformed ground truth coordinates, g_j^{cx}, g_j^{cy} are the j^{th}original ground truth coordinates, d_i^{cx}, d_i^{cy} are the i^{th} default box center coordinates and d_i^w, d_i^h are the i^{th} width and height of default box.

Let Nbe the number of matched (positive) default boxes of the input frames with ground truth. Therefore, the overall loss is given by Eq. (8).

$$L = \frac{1}{N}\left(L_{confidence} - \alpha L_{localization}\right) \quad (8)$$

Where $L_{confidence}$ is the confidence loss and $L_{localization}$ is the localization loss. *N* is not equal to zero and if it is, then the loss would be zero (there is no positive default box).

Building Video Index: The semantic objects generated by using MOSD are then structured and stored in a semi-structured index store (JSON). The final video index comprises the video metadata, the extracted semantic objects, and the semantic objects' occurrences.

4. Results and discussion

4.1 ILSVRC dataset

Before the results, it's crucial to define and conduct some statistics on the dataset used and work progress regarding VID ImageNet [4]. The VID ImageNet has been provided for public use since 2015 for use in the ImageNet Large Scale Visual Recognition (ILSVR) Challenge. There is a total of 3862 snippets for training. The number of snippets for each synset or category ranges from 56 to 458. There are 555 validation snippets and 937 test snippets. There are 30 basic-level categories in this dataset. The objects are chosen considering different factors such as movement type, level of video clutter, average number of object instances, and others.

4.2 Evaluation metrics:

The mean Average Precision (mAP) was used to measure the precision for the entire model. It was used to find the correct percentage predictions in the model. It is calculated using according to Eq. (9).

$$mAP = \frac{1}{n}\sum_{k=1}^{n} AP_k \quad (9)$$

Table 1. MOSD comparison among different configurations

	First frame	First-Last frames	First-Middle-Last frames
Frame Per Second (FPS)	78	126	185
1-Minute Video Number of Frames	60	120	180
Model Complexity	Normal	Interm ediate	High
Number of Parameters	77,176, 418	89,606 ,852	102,037, 286
mAP(%)	81.92%	83%	85.2 %

Where AP_k is the k^{th} class average precision and n is the total number of classes.

4.3 Implementation

The MOSD model has been implemented using Pytorch. The experiment is conducted with three different input depths. An input depth of 1,2, and 3 are used to train the model. Each frame of the input has a size of 3×300×300. The model training was conducted on a single NVIDIA GeForce GTX TITAN X GPU with 32 GB memory for 3 input depths. Based on these input depths, the video segment is represented by a different number of frames which are used to extract the semantic concepts.

For a 1-minute 24-FPS video, the total number of frames is roughly 1440 frames. For each one second of video we consider three cases, the first frame (i.e., 60 frames), first-last frames (i.e., 120 frames) and first-middle-last frames (i.e. 180 frames). Then the experiment has been conducted on the generated frames. The three cases are compared through different criterion as depicted in Table 1.

We have conducted our experiments on the training samples and we have observed the mAP accuracy, processed FPS, model parameters, and model complexity. It has been noticed that as we increase the input depth as we get more frames to be processed and the model ability to process FPS is increased. However, in the other side, the model gets more complexity as the input depth increases and the model parameters increases.

The three different cases achieved 81.92%, 83%, and 85.2 % respectively, for the first frame, first-last frames and first-middle-last frames. First-Middle-Last input depth case achieved the best mAP. It also has the highest FPS (185 FPS) but a quite large number of parameters. This large number of parameters slows down the processing time but

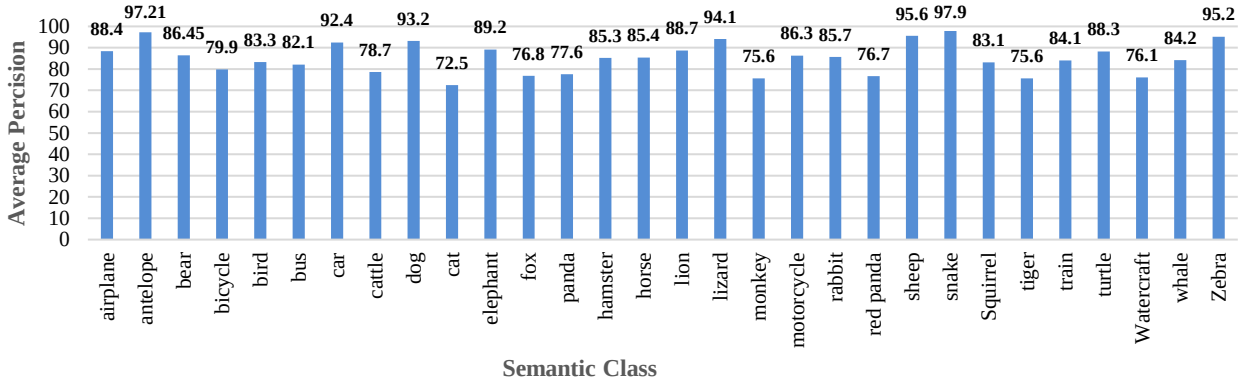


Figure. 7 Average percision per semantic class

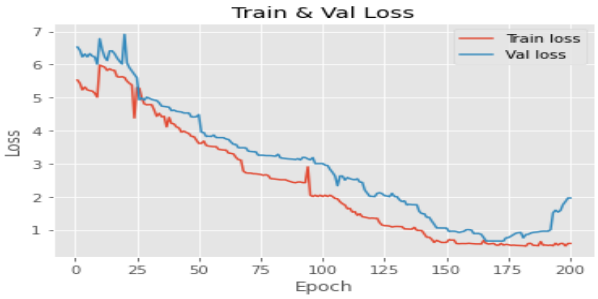


Figure 4. Training loss per epoch

increases the ability of the model to differentiate different cases. As a result, we concluded that the third case (First-Middle-Last) achieved the best case in terms of FPS and mAP accuracy.

4.4 Results analysis

In a training wise, the model shows efficiency in terms of the mAP in extracting the semantic objects out of the video. We have considered three frames for each second (first, middle, and last), so the input depth is three frames. The model training has been conducted on ImageNet VID training dataset [4] for 200 epochs and the model has achieved superior detection results on the ImageNet VID testing dataset. The loss decreases during training, which indicates the model fits the training data over time as depicted in figure-8. The model performance started to degrade on the validation set after 175 epochs. We thus used the early stopping approach to stop the training at this point. Along with applying normalization and data augmentation techniques to the dataset to avoid model over-fitting.

To prove efficiency of the proposed model, mAP has been conducted and compared with Detect-to-Track and Track-to-Detect (D&T) [25], T-CNN [20], Seq-NMS [21], Flow-Guided Feature Aggregation (FGFA) [24] and Temporal ROI Align [39]. They are all trained on ImageNet VID dataset with different backbones and detectors.

Table 2. MOSD vs Peer-Model Comparison. Mean Average Precision (mAP) comparison of D&T, TCNN, Seq-NMS, and FGFA vs MOSD

Method	**Backbone**	**Detector**	**mAP %**
D&T [25]	ResNet-101	R-FCN	79.8
	ResNet-101	Faster R-FNN	80.2
	Inception-v4	R-FCN	82.2
TCNN [20]	DeepID+Craft	R-CNN	73.8
Seq-NMS [21]	VGG16	Seq-NMS (max)	50.5
	VGG16	Seq-NMS (avg)	51.4
	VGG16	Seq-NMS (best)	53.6
FGFA [24]	ResNet	R-FCN	78.4
	Aligned Inception ResNet	R-FCN	80.1
Temporal ROI Align [29]	ResNet-101	RPN	84.3
MOSD	**3D-VGG16**	**MOSD-Detector**	**85.2**

MOSD performance was improved by combining video temporal information with multiple semantic representational scales. This allowed for simultaneous detection of multiple frames. Therefore, MOSD outperformed the previously described techniques. This was demonstrated by the results in Table 1; using the first, middle, and final video frames provided the best FPS as well as the best mAP. MOSD has achieved a mAP of 85.2 % on first-middle-last frames of each video second, which outperforms the aforementioned models as presented in Table 2 and Fig. 9. The average precision per each of the semantic objects has been calculated also as shown in Fig. 7.

As a result, The MOSD has superior results compared to the state-of-the-art video semantic detection models in the context of video semantic object detection.

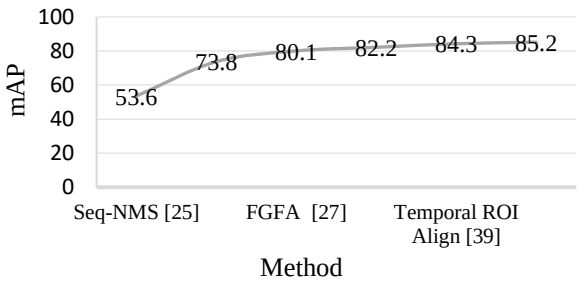


Figure. 5 mAP Comparison for MOSD and other methods

5. Conclusions

This paper presented a quick review on the VID extraction methods. Both of the features based and semantics based methods were carefully reviewed. It also proposed MOSD, a new intermediate-level semantic extraction approach that uses a 3D Convolutional Neural Network to perform many detections from multiple frames simultaneously time, rather than the usual intermediate-level video indexing methods. The key property of the MOSD model is that it utilizes the temporal context of the video for extracting an intermediate index level. The main idea of MOSD is multi-object semantic detection from video for indexing. It has the ability to detect 185 FPS. MOSD outperformed the state-of-the-art methods which used to extract intermediate semantic concepts of the video. As a future work, tinny mobile version of MOSD will be investigated to try to get various setups of MOSD. Increasing the FPS of the MOSD model will be further investigated.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contributions

"Conceptualization, Eslam Mofreh and Amr Abozaid; methodology, Kamal ElDahshan; software, Hesham Farouk; validation, Eslam Mofreh, Amr Abozaid, and Kamal ElDahshan; formal analysis, Eslam Mofreh; resources, Kamal ElDahshan; data curation, Amr Abozaid; writing—original draft preparation, Eslam Mofreh; writing—review and editing, Eslam Mofreh and Amr Abozaid; visualization, Eslam Mofreh; supervision, Amr Abozaid; project administration, Kamal ElDahshan; funding acquisition, Hesham Farouk".

Acknowledgments

This paper is based upon work supported by Science, Technology & Innovation Funding Authority (STDF) under grant (NCP 7 ID 33539).

References

[1] H. Farouk, A. Abozeid, K. Eldahshan, and M. H. Eissa, "GLOBAL DOMINAT SIFT FOR VIDEO INDEXING AND RETRIEVAL", *Journal of Theoretical and Applied Information Technology*, Vol. 97, No. 19, 2019.

[2] A. Abozaid, K. Eldahshan, H. Farouk, and E. Mofreh, "Video semantics exploration for indexing and retrieval", *Al-Azhar Bulletin of Science*, Vol. 31, No. 2-B, pp. 39-50, 2020.

[3] S. Wagenpfeil, F. Engel, P. M. Kevitt, and M. Hemmje, "AI-Based Semantic Multimedia Indexing and Retrieval for Social Media on Smartphones", *Information*, Vol. 12, No. 1, pp. 43, 2021.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge", *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252, 2015.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", In: *Proc. of IEEE Computer Society Conf. On Computer Vision and Pattern Recognition*, pp. 886-893, 2005.

[6] N. J. Janwe and K. K. Bhoyar, "Multi-label semantic concept detection in videos using fusion of asymmetrically trained deep convolutional neural networks and foreground driven concept co-occurrence matrix", *Applied Intelligence*, Vol. 48, No. 8, pp. 2047-2066, 2018.

[7] S. Ghatak and D. Bhattacharjee, "Video Indexing Through Human Face", In: *Proc. of International Conf. on Communication, Circuits, and Systems. Lecture Notes in Electrical Engineering*, pp. 99-107, 2021.

[8] A. Podlesnaya and S. Podlesnyy, "Deep Learning Based Semantic Video Indexing and Retrieval", *ArXiv*, Vol. abs/1601.07754, 2016.

[9] S. Protasov, A. M. Khan, K. Sozykin, and M. Ahmad, "Using deep features for video scene detection and annotation", *Signal, Image and Video Processing*, Vol. 12, No. 5, pp. 991-999, 2018.

[10] L. Wang, H. Zhang, and G. Yuan, "Big Data and Deep Learning-Based Video Classification Model for Sports", *Wireless Communications and Mobile Computing*, Vol. 2021, p. 1140611, 2021.

[11] R. S. Kızıltepe, J. Q. Gan, and J. J. Escobar, "A novel keyframe extraction method for video classification using deep neural networks",

Neural Computing and Applications, pp. 1-12, 2021.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Columbus*, pp. 580-587, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", In: *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 9, pp. 1904-1916, 2015.

[14] R. Girshick, "Fast R-CNN", In: *Proc. of IEEE International Conf. on Computer Vision*, pp. 1440-1448, 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks", In: *Proc. of the 28th International Conf. on Neural Information Processing Systems*, pp. 91-99, 2015.

[16] C. Szegedy, A. Toshev, and D. Erhan, "Deep Neural Networks for object detection", *Advances in Neural Information Processing Systems*, Vol. 26, 2013.

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks", In: *Proc. of International Conf. on Learning Representations*, 2014.

[18] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection", *ArXiv Preprint ArXiv:2004.10934*, 2020.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. Berg, "SSD: Single Shot MultiBox Detector", In: *Proc. of European Conf. on Computer Vision*, pp. 21-37, 2016.

[20] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, and X. Wang, "T-cnn: Tubelets with convolutional neural networks for object detection from videos", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 10, pp. 2896-2907, 2017.

[21] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for Video Object Detection", *ArXiv Preprint ArXiv:1602.08465*, 2016.

[22] B. Hatem, H. Zhang, V. Fresse, and E. B. Bourennane, "Improving Video Object Detection by Seq-Bbox Matching", In: *Proc. of the 14th International Joint Conf. on Computer Vision on Imaging and Computer Graphics Theory and Applications, Prague, Czech Republic*, pp. 226-233, 2019.

[23] K. Chen, Z. Wang, X. Wang, D. Gong, L. Yu, Y. Guo, and G. Ding, "Towards real-time object detection in GigaPixel-level video", *Neurocomputing*, Vol. 477, No. 7, pp. 14-24, 2022.

[24] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection", In: *Proc. of the IEEE International Conf. on Computer Vision*, pp. 408-417, 2017.

[25] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect", In: *Proc. of the IEEE International Conf. on Computer Vision*, pp. 3038-3046, 2017.

[26] Y. Chai, "Patchwork: A Patch-Wise Attention Network for Efficient Object Detection and Segmentation in Video Streams", In: *Proc. of IEEE/CVF International Conf. on Computer Vision*, pp. 3414-3423, 2019.

[27] M. Fujitake, "Video Representation Learning Through Prediction for Online Object Detection", *IEEE/CVF Winter Conf. on Applications of Computer Vision Workshops*, pp. 530-539, 2022.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", In: *Proc. of IEEE International Conf. on Computer Vision*, pp. 2980-2988, 2017.

[29] T. Gong, K. Chen, X. Wang, Q. Chu, F. Zhu, D. Lin, N. Yu, and H. Feng, "Temporal ROI align for video object recognition", In: *Proc. of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 1442-1450, 2021.

[30] J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification", *IEEE Conf. On Computer Vision and Pattern Recognition*, pp. 4694-4702, 2015.

[31] G. Varol, I. Laptev, and C. Schmid, "Long-Term Temporal Convolutions for Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 6, pp. 1510-1517, 2018.

[32] G. Yang and W. X. Zou, "Deep learning network model based on fusion of spatiotemporal features for action recognition", *Multimedia Tools and Applications*, Vol. 81, No. 5, pp. 1-22, 2022.

[33] I. Misra, C. L. Zitnick, and M. Hebert, "Unsupervised Learning using Sequential Verification for Action Recognition", *ArXiv*, Vol. abs/1603.08561, No. 7, p. 8, 2016.

[34] B. Fernando and S. Gould, "Learning End-to-end Video Classification with Rank-Pooling", In: *Proc. of the 33rd International Conf. on Machine Learning*, pp. 1187-1196, 2016.

[35] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", In: *Proc. of the 27th International Conf. on Neural Information Processing Systems, MIT Press*, pp. 568-576, 2014.

[36] M. Hemalatha and C. C. Sekhar, "Domain-Specific Semantics Guided Approach to Video Captioning", In: *Proc. of IEEE Winter Conf. on Applications of Computer Vision*, pp. 1576-1585, 2020.

[37] J. P. Martin, B. Bustos, S. J. F. Guimarães, I. Sipiran, J. Pérez, and G. C. Said, "A comprehensive review of the video-to-text problem", *Artificial Intelligence Review*, Vol. 55, No. 1, pp. 1-75, 2022.

[38] J. Vaidya, A. Subramaniam, and A. Mittal, "Co-Segmentation Aided Two-Stream Architecture for Video Captioning", In: *Proc. of IEEE/CVF Winter Conf. on Applications of Computer Vision*, pp. 2442-2452, 2022.

[39] S. V. Porter, M. Mirmehdi, and B. T. Thomas, "A shortest path representation for video summarisation", In: *Proc. of the 12th International Conf. on Image Analysis and Processing, IEEE Computer Society*, pp. 460-465, 2003.

[40] T. D. Truong, V. T. Nguyen, M. T. Tran, T. V. Trieu, T. Do, T. D. Ngo, and D. D. Le, "Video Search Based on Semantic Extraction and Locally Regional Object Proposal", In: *Proc. of International Conf. on MultiMedia Modeling*, pp. 451-456, 2018.