Al-Azhar Bulletin of Science Vol. 31, No. 2, (December), 2020, pp. 39-50

VIDEO SEMANTICS EXPLORATION FOR INDEXING AND RETRIEVAL

Kamal ElDahshan[1], Amr Abozeid[1], Hesham Farouk [2], Eslam Mofreh[1]

[1] *Department. of Mathematics, Computer Science Division, Faculty of Science, Al-Azhar University, Cairo, Egypt*
[2] *Computers and Systems Dept., Electronic Research Institute, Cairo, Egypt*
* **Corresponding author:** Heshamali68@hotmail.com

ABSTRACT

Video semantic concepts exploration is a fundamental problem in a video indexing and retrieval. It has a long history of investigation since early days till recent achievements. The challenges lie in bridging the gap between low level features and semantic level ones. To stand on the thoroughly situation, video semantic concepts exploration for indexing and retrieval purposes evolution from conventional methods to the state-of-the-art ones will be reviewed. The main contribution is to unify concepts involved and evolution in this interesting topic.

1. INTRODUCTION

There is a dramatic explosion in the amount of raw multimedia data produced including text, image, audio and video which are growing up rapidly since the last few years. This tremendous increase is due to social networks such as Facebook, Twitter, You-tube and also with the advancement of technology in smart handheld devices, cameras and storage devices. One of the most interesting data types is the video due to its huge amount of raw data it contains from both audio and visual channels. This led to an urgent need to semantically understand the video for indexing and retrieval purposes. Of course, this is a challenging task in computer vision and attracted researchers' attention around the world.

Perceptual understanding of the visual based data is one of the most interesting tasks due to the challenges it contains. The human visual system is confronted with a massive amount of visual information that need to be understood and interpreted. Our brains are capable of remembering from previous experiences then achieve excellently perception of visual surroundings [2].

To simulate the human visual system behavior, researchers work on making the computers interpret and understand the visual data from a human perspective.

The richest type of the multimedia is video. Video is rich of raw data that comes from visual and audio channels [1]. The video consists of a set of consecutive frames which are chronologically ordered with an audio, in addition to, speeches or user-defined metadata. Due to its richness of raw data and no prior structure, the video processing is considered a challengeable task. The challenge is to find efficient solutions with less processing and accurate results. The video will be addressed here in terms of its visual content.

One of the most beneficial applications of video is the indexing and retrieving. Indexing of video segments is about assigning it to a specific set of categories and labels. The video is indexed at different levels and contains many subtasks. Some of these tasks index the video at a high level manner while others are specialized and application-based. The most basic tasks of video indexing include object detection, action recognition, sequences classification, etc. Hence, organizing the resulted concepts in an effective way for ease of retrieval.

The investigation in this area has a long history and there is a tremendous progress in the number of research produced recently. These methods can be roughly divided into two approaches: 1) handcrafted (conventional) approaches 2) deep learning approaches.

Understanding the image and video contents lies in how the pixels interpreted. At early days, researchers interpreted image and video by extracting low-level features and tried to assign it specific semantics concepts. Therefore, there existed a semantic gap between these low-level features and semantic concepts. These

approaches are usually referred to as handcrafted or conventional. Weiming et al. [1], introduced one of the most considerable surveys that review this approach. Recently, deep learning models specifically Convolutional Neural Network (CNN) architectures achieved terrific results which has attracted researchers' attention world widely. Since 2010, deep learning has strongly emerged again after the Large Scale Visual Recognition Challenge (ILSVRC) has been launched. A considerable achievement has been introduced to CNN architectures with AlexNet (the winner in 2012) and since though there is an explosion in the research introduced in CNN architectures. CNN is deep learning model that preserves spatial information for processing visual imagery data. It will be reviewed in upcoming sections.

The motivation behind this study is the lack of studies introduced in this topic. Fortunately, this topic will be addressed in this paper from the handcrafted approaches to the state-of-the-art deep learning ones.

Contributions:

1- Introduce a review study to find out the directions in video indexing and retrieval.
2- Consolidate a basic structure of existing work in video indexing and retrieval for further work.
3- Reviewing of conventional methods and peers state-of-the-art deep learning ones.

2. Video Concepts:

Semantics concepts exploration is the essential step for many applications in the domain of video. There are several levels of representation of a video data. The lowest representation level of a video data is the pixel level which indicates color information. At a higher levels of abstractions, the features are being revealed to explore acquaintance behind. Semantic concepts are achieved at the final stage of the exploration levels.

Video semantic concepts exploration is about discovering hidden patterns and perceptual meaning of video hoping to realize a structured representation of semantic concepts.

The video concepts can be laid out in a set of layers such that in the first stage of semantic concepts there are a set of objects, motions and actions, and at successive layer there are objects relationships and a video events and other layers that are based on the application to finally extract semantic concepts which are useful for indexing.

During the era of research in computer vision, there are a going efforts in video semantic concepts exploration until today with achieved an immense success. Generally, this paradigms can be divided into two, the handcrafted (conventional) paradigms and the deep learning based ones. The handcrafted ones usually take multiple steps to achieve semantically understanding since it usually finds out low-level features from video pixels then goes up to a higher abstraction levels until realizing semantic concepts at the final stage which would be used in indexing. These low-level features include color, appearance, texture, shapes, etc.

The tasks which are reported in the literature for exploring video semantic concepts are basically the following:

1) Object Classification/Detection
2) Object Action Recognition
3) Event/Activity Recognition
4) Temporal Sequence Classification.

We'll review them in the upcoming sections. But before diving in the heart of the matter, there is a preliminary step which is "Video Structure Decomposition".

Some works considered this step as the first step in indexing of video before exploring the video semantic concepts. Therefore, it'll be reviewed it in the next section.

3. Video Structure Decomposition:

Video structure decomposition aims at dividing the problem of processing of video into sub ones. Instead of dealing with the whole video, it is efficient (time complexity) to deal with small manageable components. Most of the recent approaches neglected this step as it causes loss of the context and temporal information which are considered important for achieving better accuracy when detecting concepts of video. Therefore, some researchers considered the processing unit to be some of the representative frames. Usually the steps involved to decomposing the video structure are:

1- Shot/Scene boundary detection

2- Key-frame extraction

3.1 Shot Boundary Detection:

The video consists of a set consecutive frames which are chronologically ordered with an audio. Hence, the frame is the basic building unit of video. At a higher stage there is a shot which is a set of consecutive inter correlated set of frames. Shot boundaries are determined by a start and end of a camera operation such as camera angle change. Generally, boundaries due to cut or gradual transitions such fed, dissolve, wipe, etc. Cut transitions are easier to detect than the gradual transitions since the gradual transition distributed among multiple frames.

In shot boundary detection usually begin by extracting features from adjacent frames, then similarity distance between the frames are identified based on the extracted feature vectors and the closest frames are grouped together in one shot or one scene. The conventional methods used low level features as color histogram [42], SIFT [43] (Scale Invariant Feature Transform), motion vectors [44], bag of feature (BOF). Hence, researchers divided into those that uses threshold based approach, others treated shot boundary as a classification task and used statistical methods such as SVM for classification and others treated this as a clustering task in which frames in the same shot are clustered together in one cluster.

With deep learning success, CNN has been used as the source of features instead of low level features used by conventional methods [2], [3], [4].

3.2 Key-frame Extraction:

The redundant content nature of the video needs to be handled in an efficient manner for reduction of processing complexity. For these purpose, it is natural to process only some of the frames and neglect the duplicated ones. These frames should be selected efficiently to best reflect the shot content [66], [45], [46]. These descriptive and representative frames are referred to as Key-frames. We can categorize the methods used to extract key-frames into:

3.2.1 Reference Based Extraction: These methods work in extracting a frame which is used as a reference for the upcoming ones. The shot frames are compared to this reference frame to eventually extract the key-frames. This method is sub-divided into two sub categories, sequential comparison and global comparison.

i. **Sequential comparison**: In these methods, an initial key-frame is selected and sequentially compares it to the subsequent until it obtains one that is different from the previously selected one. Zhang [61] computes an energy function from two consecutive frames to measure the distance between these two frames. Hang et al. [60] used histogram difference between two consecutive frames to extract key-frames.

ii. **Global comparison**: In these methods, a reference frame is selected. All the shot frames are compared with this frame to extract key-frames. Sun et al [59], extracts the most occurred frame as a reference frame. The distance between the frames and this frame is measured. The frames at the peaks of the curve are selected as key-frames. Ferman and Teklap [58], extracts alpha trimmed average histogram to describe colors distribution in the shot frames. This alpha trimmed histogram is compared to the histogram of each of the frames.

3.2.2 Objective Based Extraction: The extraction process in these methods are done based on a predefined objective function. These objective function can be one of the following:

i. Temporal variance: These methods consider that the extracted key-frames are made from a set of segments with equal temporal variance. Divakaran et al. [41], extracts key-frames from shot segments which are having equal MPEG-7 motion activity. The frame at the midpoint is selected as a key-frame.

ii. Key-frame representativeness: These methods act to scaling up the number of frames that the key-frame represents. It can sub categorized into two cases, if the number of key-frames is fixed or not. If the number of key-frames is fixed, then these methods work to maximize the number of the number of the frames that represents the key-frame. In contrast if the number of key-frames is not fixed then these methods working to minimize the number of key-frames based on a predefined fidelity criterion [62], [63].

3.2.3 Structure Based Extraction: These methods aim initially at structuring the shot into structured representation. These methods can be

sub divided into those use graphs and others that use curve.

i. Graph: These methods assume that the key-frames should be less correlated. From this point, they aiming at reducing the correlation between the key-frames. Porter et al. [64], represents the shot by a directed weighted graph. The shortest path is the path that represents minimum correlation between frames that represented by the key-frame.

ii. **Curve**: These methods represent the shot frames into points in the feature domain and linking each of the frames to form a trajectory curve, then the set of points that best represent the curve are being searched. Calic and Izquierdo [65], use the macroblocks of MPEG compression algorithm as a source of features, then they statistically analysis these macroblocks to find out the dissimilarity between frames to be linked.

3.2.4 Features Based Extraction: These methods are based on extracted features and semantics of a key-frame. In these methods, the frames that lay in the same shot are considered to contain similar semantics. Calic and Thomus [56], use the regions resulted of segmentation to extract key-frames. Liu et al. [55], use color histogram to extract key-frames. Fan et al. [57], use an object segmentation to extract key-frames.

3.3 Object Classification/Detection:

Object detection is one of the most fundamental challenges of computer vision. The object detection revolves around recognizing the existence of specific objects and locates these objects in an image or a video. Object detection combines two subtasks, classification and localization. Image classification is to allocate object instances to a predefined object categories in digital image and video. Object detection is the quintessence of many applications like autonomous driving, identity recognition, video surveillance, medical, etc.

The lowest level of object features there are the object's color, brightness, appearance, texture, size, etc. In conventional methods usually features are extracted from the image and pre-crafted to the specific set of categories based on human perspectives.

Usually these methods doing well in video retrieval engines when queries are by an example, but it is more interesting to retrieve video by a text as well. This led to a gap between human perspective features and low-level feature.

The conventional computer vision paradigms utilize low level features and worked to find high level ones. Histogram of Oriented Gradients (HOG) detector [3] is an important one of the conventional methods that is bridging low-level features to high-level ones by counting the occurrence of gradient orientation of an image. HOG well describe the image object appearance and shape by intensity gradient and edge direction. Scale-invariant feature transform (SIFT) [4] describes high-level features of an object by object interesting points regardless image scale, noise and illumination. SIFT features are calculated for a reference image and stored and when there is a new example, SIFT is calculated and compared to the stored ones and then some key points in the example is identified to filters out best matches. ElDahshan et Al. [67], proposed a Global Dominant SIFT (GD-SIFT) descriptor for indexing and retrieval. Viola–Jones [5] introduced real-time object detector which motivated for face detection based on Haar features which consist of different types for different face locations.

In contrast, deep learning paradigms act as features extractors and classifiers simultaneously which shown high accurate results with superior performance. Precisely since 2012 a bang of researches in this area has been achieved in different CNN architectures thanks to deep learning. Alex Krizhevsky, proposed AlexNet (ILSVRC-2012 winner) which is the effectual re-emergence of CNN models. ZFNet (ILSVRC-2013 winner) improved AlexNet and achieved better accuracy. Visual Geometry Group from University of Oxford invented VGGNet (ILSVRC-2014 1st runner) which outperformed GoogleNet (ILSVRC-2014 winner) and won the localization task. The idea beyond VGGNet is that to reduce the number of parameters that the network uses. While GoogleNet introduced the term "Inception" which to use filters with different sizes and achieved significant performance than AlexNet and ZFNet. ResNet

(ILSVRC-2015 winner) introduced Residual Networks which solved vanishing gradient problem. Huang et al., introduced DenseNet which outperformed ResNet. Sara Sabour and Hinton proposed the idea of capsule networks (CapsNet) which radically changed CNN architecture by addressing spatial correlation of object parts. Continuously, a lot of CNN architectures are being introduced that realize considerable achievements until recent days. Subsequently, this motivated object detection progress and considerable achievements introduced to this task. Object detection architectures basically divided into region-based or two stage detectors and one stage one.

In contrast, deep learning paradigms act as features extractors and classifiers simultaneously which shown high accurate results with superior performance. Precisely since 2012 a bang of researches in this area has been achieved in different CNN architectures thanks to deep learning. There are a lot of considerable surveys in semantic object detection that discuss object detection in deep learning [51], [52], [53], [54]. It could be basically divided into region-based or two stage detectors and one stage one.

i. Two Stage Detector (Region Based): In these methods, category independent regions are being proposed, then each of the regions are fed to a CNN architecture to classify each of them to a specific-category. Girshick et al. [39], proposed RCNN which uses selective search as a regions proposal. Selective search proposes roughly 2000 of interest regions that may contain objects and then each of them is rescaled to 227x227, thereafter a pre-trained CNN architecture (e.g. VGG) is then fine-tuned to train a SVM to classifies the extracted regions and another regression model for tighten bounding boxes. RCCN yielded a boost performance to object detection with mean Average Precision (mAP) with 58.5%. But RCNN has the following drawbacks: 1) Highly consuming load on file system during training 2) Superfluous redundant computations because of extracting features for overlapped regions 3) Need of rescaling of regions before feeding to CNN.

K. He, X. Zhang proposed a Spatial Pyramid Pooling Networks (SPPNet) which mainly contributed with Spatial Pyramid Pooling (SPP) which generating of fixed length representation of regions without the need of rescaling the regions. SPPNet accelerates speed without sacrificing of detection accuracy, it has mAP of 59.2%.

Motivated by SPPNet and RCNN, Girshick proposes a Fast RCNN which unifies the three models used by RCNN into only one model. Instead of running CNN for each of the regions, Fast RCNN runs only one and share the extracted features across each of the region proposals by adding a Region of Interest Pooling RoI pooling layer between Conv layers and FC ones. RoI pooling is then used by bounding box regressor and object classifier. Although, the achieved improvement in the speed, it still not the dramatic solution because it uses a third party to propose regions of interest.

Instead of the methods that rely on selective search and edge box, S. Ren et al., proposed Faster RCNN which integrated Region Proposal Network (RPN) into CNN. Faster RCNN is a unified model that is composed of Region Proposal Network (RPN) and Fast RCNN that each share the same features. Therefore, Faster RCNN is working out to fine-tuning RPN in end-to-end to find regions by sliding window and use it with different scales and aspect ratios and the Fast RCNN one also is fine-tuned greatly improved both precision and detection efficiency. Faster R-CNN achieved mAP of 69.9% on PASCAL VOC 2007 compared to Fast R-CNN of 66.9% with shared convolutional computations and the running time of Faster R-CNN was roughly 10 times lower than Fast R-CNN with VGG backbone.

ii. One Stage Detector: These methods use one model for both detection and bounding box regression.

Szegedy et al.[30], was first of the contributors that contribute in single stage detectors and proposed a DetectorNet which treated an object detection as a regression problem. They used AlexNet and replaced the Softmax layer with a regression one.

Sermanet et al. [31], proposed OverFeat which is a single stage object detection that simultaneously do classification, localization and object detection. The main contribution of OverFeat is do multi-scale classification at

different regions, and predict bounding box with a regressor on top of features extractors.

Redmon et al. [32], proposed You Only Look Once (YOLO) network which is the first attempt to build a real-time object detector. It works by dividing the image into mxm cells, and each cell is responsible for detecting and object that centered in it. Each of the grid cells and predicts bounding boxes and their corresponding confidence scores. The confidence score is defined by Pr(object) * IoU(Predicted,Ground Truth) and the bounding box is identified with four values of the bounding box coordinates.

Liu et al., [33] proposed Single Shot MultiBox Detector which take a pyramidal hierarchy manner when extracting features with CNN. YOLO faced a difficulty while detecting small objects, instead SSD used a different manner by extracting features of each location with different sizes and aspect ratios. SSD uses VGG-16 as a backbone network for extracting features. On top of VGG-16, SSD added several Conv layers of decried sizes to extract features at multi-level scales in a pyramidal representation. This manner actually very useful when dealing with different object of different sizes since large-grained feature levels are good at capturing tinny objects while coarse-grained feature levels are good at large ones. Sabbeh et al. [50], uses VGG as a backbone for extracting semantic features from a video frame. They divide a frame into 224 x 224 overlapping segments then pass it through VGG to extract semantic objects.

A lot of efforts has been done till recent days by the researchers to investigate and further enhance this task. In the context of video, the context, and temporal information should be considered when during detection. Therefore, some researchers work on bounding-box level while some of them work on features level and others are utilized boundary box level and feature level together. The methods that work on bounding box level, such as, Kang et al. [34], who introduced a T-CNN model that uses precomputed optical flow fields and object tracking to propagate bounding boxes to nearby frames.

Han et al. [35], built Seq-NMS that boosted detection by utilizing high score detection from nearby frames. B. Hatem et al., proposed Seq-Bbox that build tubelets by linking bboxes across frames to infer missed and improve detection while the methods work on features level are achieved better improvement. Zhu et al. [36], introduced Flow-Guided Feature Aggregation (FGFA) model that used optical flow and features extracted from nearby frames for improving detection. Feichtenhoter et al. [37], proposed Detect-to-Track and Track-to-Detect (D&T) model that is jointly used for detection and tracking. It'd better to take a flavor of the bounding box and features level during detection. Yuning Chai [38], proposed PatchWork model for detecting objects from a video by utilizing a specialized memory that recovers lost context. Also, Patchwork adopted Q-learning based policy that intelligently selects sub windows to be processed in subsequent frame.

2. Action / Activity Recognition:

Action / Event Recognition: Video actions and events are an important concept that assist in building robust index. Action is the act which is been done by an object in a certain period of time, while the event is an occurrence of one or more actions in a consolidated. Action and event are two closely concepts and can be inferred from each other. Recognizing of actions and events of a video is a taugh task because of the temporal and context of video should be utilized in effective manner without loss of information as possible for best prediction. This task still challengeable and has a lot of investigation till now. We begin by reviewing conventional methods and peers of deep learning methods.

5.1 Conventional Action Recognition:

Our brain treat and action by first representing it with a symbolic system then treated it with a set of processors. Therefore, an action can be represented by an easy to compute structured system, representative of an action and differentiable of different actions [48]. The Earliest models of action recognition are make use of 3D of video to represent an action. Hogg et al. (1983), represents action by the Walker model for human action recognition. Rohr et al. (1994), make use of cylinders to represents pedestrian recognition. The use of 3D model to describe action is very expensive in the context of videos instead we'd discuss other methods that are more efficient.

Basically these methods divide an action recognition into two main parts: 1) Action Representation (Feature Extraction) 2) Action Classification [47].

Action Representation: By action representation, it meant to extract the action features. The extracted features need to be representative and discriminative for the action. The is no doubt being a challenging problem due to the variations in objects appearance and their motion speed, camera view (left, right, top, down or zooming), pose variations and many more.
An action features exist in two levels, holistic level and local level. Holistic representation is to represent body with a thorough representation while local representation is to reflect local features.

Holistic Representation: The action information contains both spatial and temporal information which form a 3D shape. These methods aim to represent the object motion entirely. It extracts information about the motion in a certain cubic region contains the object. One of the considerable work introduced by Bobick et. al, who suggests to represent the action by Motion Energy Image (MEI) and Motion History Image (MHI). MEI shows where the action happens and MHI shows how the motion changes. This work success in representing the action into good manner although there is a problem that MEI and MHI are sensitive to different views. Laptev and Lindeberg [10] addresses this problem by proposing 3D Motion History Volume (MHV) from different camera views. Alper Yilmaz and Mubarak Shah, encoded the action into multiple 2D images represent the contours of the object with time change and generating 3D Spatio-Temporal Volume (STV) and analyzing the STV by using the differential geometric surface properties, such as peaks, pits, valleys, and ridges. There also works [21], [22] that encoded motion information by optical flow which estimates the motion by taking the moving object pixels' intensities over time by assuming that the illumination does not change.

2) Local Representations: These methods take the local regions that have salient motion information. Space-Time Interest Point (STIP) [18], [19], [23], [24] is the first emerge and most important work of local representations. It shows their effectiveness detecting actions with different translations and appearance variations of the motion. These describe motion by local features. Laptev [11], [24] extends Harries corner detector [26] in space-time space. Bregonzio et al. [24] detected spatial-temporal interest points using Gabor filters. [26] detect interest points by using the spatiotemporal Hessian matrix. Other algorithms detect interest points by extending 2D detectors to 3D ones, 3D SIFT [20], HOG3D [19], local trinary patterns [27], etc. Other optical flow features based approaches [19], [28] are combined Histogram Optical Flow (HOF) with HOG features to represent human actions. The aforementioned approaches keep track of spatio-temporal information in short-term duration. Feature trajectory is an approach for detecting information in long-term durations [25], [100], [101]. To obtain features for trajectories interest points are extracted and tracked by KTL tracker [102]. [79], [25], [29] concatenated HOG, HOF and MBH features for describing trajectories.

5.1.2 Action Classification:

After representing an action with a feature representation, their features are learned to assign to specific class label. According to [47], action classification can be divided into:

1) Direct classification: In these methods, an action features are being extracted and represented in feature representation, and thereafter these features are trained to assign class label to it. Classification is being done by the one of the off-the-shelf classifiers, such as support vector machine [9], [10], [11], k-nearest neighbor (k-NN) [12], [13], etc.

2) Spatio-Temporal Approaches: These methods classify actions successively by considering context and temporal dimension of a video segment. Some of these methods use hidden Markov models (HMMs) [14], [15], [16], conditional random fields (CRFs), Gaussian mixture model (GMM) [17].

3) Feature Approaches: There is a spectral of researchers who use low-level features and bridged to high-level ones because of the semantic gap between low level and semantic features. Bag-of-words was used as low-level features, and thereafter bridged it to high level one by histogram of words to finally feed to the classifier. Another approach is to fuse features

from different extractors, it was discussed in [17], which learns global Gaussian mixture model (GMM), and uses multiple GMMs to characterize local regions at multiple scales.

5.2 Deep action recognition:

There are three categories of deep models used for action recognition:

- Spatiotemporal networks
- Multiple stream networks
- Temporal coherency networks

5.2.1 Spatio-Temporal Networks:

When it related to sequences and videos, then the temporal information should be considered (i.e. keeping track of both spatial and temporal information) for action recognition. From this point, Ji et al. (2013) suggests using temporal dimension alongside the 2D CNN convolution operation to be 3D operation, and this proved its benefice ability in action recognition task.

There are other researchers that investigated adding temporal information for action recognition, Ng et al. (2015) proposed the temporal pooling and found that max temporal pooling is more beneficial. Karpathy et al. (2014) proposed slow fusion. In slow fusion architecture, each consecutive set of frames are passed to the same set of layers of convolution, pooling, and fully connected layer and the responses of these parts are fused via fully connected layer to generate the action description.

Varol et al. (2016) found enhanced effect when increasing the temporal duration of the input and combining results of different inputs temporal durations of video.

By adding the temporal dimension, the parameters get increased and this will affect the 3D convolution operation performance. Sun et al. (2015) investigated this issue and suggests factorizing the 3D filter into 2D and 1D ones and this proved to be efficient.

Baccouche et al. (2011) suggests using two separated networks. 3D convolutional network for feature extraction and LSTM for action classification based on the extracted features.

Donahue et al. (2015), suggests using composite structure of both CNN and LSTM and the resulting network named Long-term Recurrent Convolutional Network (LRCN) and it proved success in action recognition and image and video annotation.

5.2.2 Temporal coherency networks:

The concept of temporal coherency of a video is that each consecutive set of frames are semantically and dynamically coherent.

A video is said to be coherent if:

1. The video frames are in its appropriate temporal order.
2. The video events semantics are correlated.
3. There are no abrupt changes in event semantics, or motions.

Goroshin et al. (2015) and Wang, and Gupta (2015), investigated unsupervised autoencoder architecture for video representation with the assumption that adjacent video frames are semantically correlated. Misra et al. (2016), investigated temporal coherency in learning visual representation of a video for action recognition task.

Jiwen Lu et al. (2016), investigated temporal coherency concept by proposing a deep architecture for visual tracking. Rahul et al. (2016), also investigated temporal coherency by proposing a deep architecture for human re-identification in surveillance videos.

Another related work investigated by Yingwei Li et al., who proposes an approach for action recognition for long-range dynamics (which is supposed to be usually inhomogeneous) by grouping short-term homogeneous ones.

5.2.3 Multiple Stream Networks:

This type of networks is inspired by our visual cortex. The visual cortex has two streams, Ventral stream and Dosaral stream. The Ventral stream identifies the object identity, color and appearance and the Dosaral stream recognizing the motion of the object.

Simonyan and Zisserman devised an architecture that exploits both appearance and motion (spatial and temporal) information. They built spatial stream network trained by video frames and temporal stream network trained by optical flow fields.

Feichtenhofer et al.(2016) , study that the early fusion of the spatial and temporal information at the convolution layer rather than the softmax layer. They achieved the same performance as in Simonyan and Zisserman with half of the parameters.

3. Video Indexing:

A video indexing is the final stage after the PRE aforementioned stages. After extracting a set of semantic concepts, thereafter indexing can be done into levels based on the extracted semantic concepts. At the first level, there is the video and sub scenes and shots classification results, at a higher level there are a set of semantic objects per each of the frames, motions, and events, and at the final level there is an inverted index that best describe the video. A video Indexing is about organizing and storing these extracted semantic concepts in a structured efficient manner for and ease of retrieval. There are different methods of indexing representation, and they can be divided into a hierarchical representation and graph representation.

Hierarchical representation: these methods represent video index at different levels of semantic levels. Gang et al., introduced a hierarchical representation of video semantics at two levels, labels, and semantics.

Graph representation: these methods depicted a video as a graph of semantic features, and the nodes represent the set of semantics concepts, and the edges are the dissimilarities between these between each node semantics, and the video shot frames are recognized by the path that offers a minimum value of weights. Porter et al. [64], depicted a shot by a directed weighted graph. Motivated by ImageNet, Podlesnaya et al.[8], introduced a way of building an index that is oriented by a graph that the nodes are the salient objects and edges are linked to other objects or the WordNet lexical database and used a Neo4j graph-oriented database.

4. Video Querying and Retrieval:

By obtaining a video index, we're ready to query and retrieve video segments. Video querying is to request about video or just a small segment of video. Once a query has been send, a measure of difference between query and the indices are being calculated for searching the database for best candidates.

7.1 Video Querying Types:

Querying a video can be of different types, and it should be handled in any video retrieval engine. The query may be by an example, set of specific keywords, or even by natural language.

Query by an example: this type can be done by an example segment or an image. It is somewhat a harsh retrieval one because it requires extracting of features from the provided example and then measuring the similarity difference between it and the database of video indices.

Query by semantics: This type of querying enables the user to provide a set of objects, motions or some event to retrieve based on. The retrieval engine must be intelligently retrieving the most relevant segments of video that are most likely contains the semantics provided in the query.

Query by keywords or labels: In this type the user can type some of the specific set of descriptive words to retrieve segments based on it. The retrieval model analysis these models to get a set of semantics and match them with the database of video indices. The most relevant segments are returned to the user. Truong et al. [49], provides ad-hoc querying in text format in a video search model.

Query by natural language: In this type, the user can write what in his mind with just a set of natural language words and the retrieval model intelligently selects the best candidates that fit the query semantics. It is the most practical type of video retrieval engines. Aytar et al. [40], used a method of matching semantic words to retrieve relevant segments.

5. CONCLUSION:

Semantic concepts exploration for indexing and retrieval was explained in each of the conventional methods and their analogues of deep learning methods. The different tasks involved in semantic concepts exploration for video indexing were linked and consolidated from its origin till recently achieved works. Research trends in each of the steps was analyzed to unify concepts in this topic for further work.

6. REFERENCES:

[1] W. Hu, N. Xie, L. Li, X. Zeng and S. Maybank, "A Survey on Visual Content-Based Video Indexing and Retrieval," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 797-819, Nov. 2011. doi: 10.1109/TSMCC.2011.2109710

[2] Chai, Yuning. "Patchwork: A Patch-wise Attention Network for Efficient Object Detection and Segmentation in Video Streams." *ArXiv* abs/ 1904.01784 (2019): n. pag.

[3] 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)*, San Diego, CA, USA, 2005, pp. 886-893 vol. 1. doi: 10.1109/CVPR.2005.177

[4] Lowe, D.G., "Distinctive Image Features from Scale-Invariant Keypoints". *International Journal of Computer Vision* **60,** 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94

[5] Viola, Paul & Jones, Michael. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. IEEE Conf Comput Vis Pattern Recognit. 1. I-511. 10.1109/CVPR.2001.990517.

[6] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 1879–1886.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual 29 recognition", in European conference on computer vision. Springer, 2014, pp. 346–361.

[8] Podlesnaya, Anna & Podlesnyy, Sergey. (2018). Deep Learning Based Semantic Video Indexing and Retrieval. Lecture Notes in Networks and Systems. 359-372. 10.1007/978-3-319-56991-8_27.

[9] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A ¨ local svm approach," in IEEE ICPR, 2004.

[10] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in IEEE Conference on Computer Vision & Pattern Recognition, 2009.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in CVPR, 2008.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Proc. ICCV, 2005.

[13] I. Laptev and P. Perez, "Retrieving actions in movies," in ICCV, 2007.

[14] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semimarkov model," in CVPR, 2005.

[15] S. Rajko, G. Qian, T. Ingalls, and J. James, "Real-time gesture recognition with minimal training requirements and on-line learning," in CVPR, 2007.

[16] N. Ikizler and D. Forsyth, "Searching video for complex activities with finite state models," in CVPR, 2007.

[17] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in CVPR, 2011.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in VS-PETS, 2005.

[19] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in BMVC, 2008.

[20] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in Proc. ACM Multimedia, 2007.

[21] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proceedings of Imaging Understanding Workshop, 1981.

[22] B. Horn and B. Schunck, "Determining optical flow," Artificial Intelligence, vol. 17, pp. 185–203, 1981. [23] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in CVPR, 2010.

[23] I. Laptev and T. Lindeberg, "Space-time interest points," in ICCV, 2003, pp. 432–439.

[25] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action Recognition ¨ by Dense Trajectories," in IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, United States, Jun. 2011, pp. 3169–3176.

[26] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scaleinvariant spatio-temporal interest poing detector," in ECCV, 2008.

[27] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in CVPR, 2009.

[28] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in CVPR, 2005.

[29] M. Jain, H. Jegou, and P. Bouthemy, “Better exploiting motion for better ´ action recognition,” in CVPR, 2013.

[30] Szegedy, C. & Toshev, Alexander & Erhan, Dumitru. (2013). Deep Neural Networks for object detection. Advances in Neural Information Processing Systems. 26.

[31] Sermanet, P. & Eigen, D. & Zhang, X. & Mathieu, M. & Fergus, R. & Lecun, Yann. (2013). OverFeat: Integrated Recognition. Localization and Detection using Convolutional Networks. 1-16.

[32] Redmon, Joseph et al. “You Only Look Once: Unified, Real-Time Object Detection.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 779-788.

[33] Liu W. et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham.

[34] Kang, Kai et al. “T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos.” IEEE Transactions on Circuits and Systems for Video Technology 28.10 (2018): 2896–2907. Crossref. Web.

[35] Han, W., Khorrami, P., Paine, T. L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S. & Huang, T. S. (2016). Seq-NMS for Video Object Detection.. *CoRR*, abs/1602.08465.

[36] Zhu, Xizhou et al. “Flow-Guided Feature Aggregation for Video Object Detection.” *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 408-417.

[37] C. Feichtenhofer, A. Pinz and A. Zisserman, "Detect to Track and Track to Detect," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3057-3065.

[38] Chai, Yuning. “Patchwork: A Patch-wise Attention Network for Efficient Object Detection and Segmentation in Video Streams.” *ArXiv* abs/1904.01784 (2019): n. pag.

[39] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587.

[40] Y. Aytar, M. Shah, and J. B. Luo, “Utilizing semantic word similarity measures for video retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[41] J. Rong, W. Jin, and L. Wu, “Key frame extraction using inter-shot information,” in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2004, pp. 571–574.

[42] C. H. Hoi, L. S. Wong, and A. Lyu, “Chinese university of Hong Kong at TRECVID 2006: Shot boundary detection and video search,” in *Proc. TREC Video Retrieval Eval.*, 2006.

[43] Y. Chang, D. J. Lee, Y. Hong, and J. Archibald, “Unsupervised video shot detection using clustering ensemble with a color global scale invariant feature transform descriptor,” *EURASIP J. Image Video Process.*,vol. 2008, pp. 1–10, 2008.

[44] Z.-C. Zhao and A.-N. Cai, “Shot boundary detection algorithm in compressed domain based on adaboost and fuzzy theory,” in *Proc. Int. Conf. Nat. Comput.*, 2006, pp. 617–626.

[45] B. T. Truong and S. Venkatesh, “Video abstraction: A systematic review and classification,” *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 3, no. 1, art. 3, pp. 1–37, Feb. 2007.

[46] K. W. Sze, K. M. Lam, and G. P. Qiu, “A new key frame representation for video segment retrieval,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 9, pp. 1148–1155, Sep. 2005.

[47] Kong, Yu & Fu, Yun. (2018). Human Action Recognition and Prediction: A Survey.

[48] Herath, Samitha et al. “Going deeper into action recognition: A survey.” *Image Vis. Comput.* 60 (2016): 4-21.

[49] Truong, Thanh-Dat & Nguyen, Vinh-Tiep & Tran, Minh-Triet & Trieu, Trang-Vinh & Do, Tien & Duc, Thanh & Le, Duy-Dinh. (2018). Video Search Based on Semantic Extraction and Locally Regional Object Proposal. 10.1007/978-3-319-73600-6_49.

[50] Sabbeh, Sahar. (2012). Multi feature content based video retrieval using high level semantic concept.

[51] Liu, Li & Ouyang, Wanli & Wang, Xiaogang & Fieguth, Paul & Chen, Jie & Liu, Xinwang & Pietikäinen, Matti. (2018). Deep Learning for Generic Object Detection: A Survey.

[52] Uçar, Ayşegül & demir, Yakup & Güzeliş, Cüneyt. (2017). Object recognition and detection with deep learning for autonomous driving applications. SIMULATION. 93. 003754971770993. 10.1177/0037549717709932.

[53] Zhao, Zhong-Qiu et al. "Object Detection With Deep Learning: A Review." *IEEE Transactions on Neural Networks and Learning Systems* 30 (2018): 3212-3232.

[54] Jiao, Licheng & Zhang, Fan & Liu, Fang & Yang, Shuyuan & Li, Lingling & Feng, Zhixi & Qu, Rong. (2019). A Survey of Deep Learning-based Object Detection. IEEE Access. PP. 1-1. 10.1109/ ACCESS.2019.2939201.

[55] L. J. Liu and G. L. Fan, "Combined key-frame extraction and object based video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 869–884, Jul. 2005.

[56] J. Calic and B. Thomas, "Spatial analysis in key-frame extraction using video segmentation," in *Proc. Workshop Image Anal. Multimedia Interactive Services*, Lisbon, Portugal, Apr. 2004.

[57] X. M. Song and G. L. Fan, "Joint key-frame extraction and object segmentation for content-based video analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 904–914, Jul. 2006.

[58] A. M. Ferman and A. M. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Trans.Multimedia*, vol. 5, no. 2, pp. 244–256, Jun. 2003.

[59] Z. H. Sun, K. B. Jia, and H. X. Chen, "Video key frame extraction based on spatial-temporal color distribution," in *Proc. Int. Conf. Intell. Inform. Hiding Multimedia Signal Process.*, 2008, p. 196-199.

[60] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognit.*, vol. 30, no. 4, pp. 643–658, 1997.

[61] X.-D. Zhang, T.-Y. Liu, K.-T. Lo, and J. Feng, "Dynamic selection and effective compression of key frames for video abstraction," *Pattern Recognit. Lett.*, vol. 24, no. 9–10, pp. 1523–1532, Jun. 2003.

[62] J. Rong, W. Jin, and L. Wu, "Key frame extraction using inter-shot information," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2004, pp. 571–574.

[63] M. Cooper and J. Foote, "Discriminative techniques for keyframe selection," in *Proc. IEEE Int. Conf.Multimedia Expo.*, Jul. 2005, pp. 502–505.

[64] S. V. Porter, M. Mirmehdi, and B. T. Thomas, "A shortest path representation for video summarization," in *Proc. Int. Conf. Image Anal.Process.*, Sep. 2003, pp. 460–465.

[65] J. Calic and E. Izquierdo, "Efficient key-frame extraction and video analysis," in *Proc. Int. Conf. Inf. Technol.: Coding Comput.*, Apr. 2002, pp. 28–33.

[66] A. Abozeid, H. Farouk, and K. ElDahshan. *2017. Scalable Video Summarization: A Comparative Study. In Proceedings of the International Conference on Compute and Data Analysis (ICCDA '17). Association for Computing Machinery, New York, NY, USA, 215–219. DOI:https://doi.org/ 10.1145/3093241.3093287*

[67] *Amr Abozeid, Hesham Farouk, Kamal ElDahshan, and M. Hamza. Eissa.* Global dominant sift for video indexing and retrieval. In Journal of Theoretical and Applied Information Technology. 15th October 2019. Vol.97. No 19

الملخص:

يعد استكشاف المفاهيم الدلالية للفيديو مشكلة أساسية في فهرسة واسترجاع الفيديو. لديها تاريخ طويل من البحث منذ الأيام الأولى حتى الإنجازات المحققة مؤخراً. تكمن التحديات في سد الفجوة بين الميزات ذات المستوى المنخفض والسمات الدلالية. للوقوف على الوضع تماماً ، سيتم استعراض تطور استكشاف المفاهيم الدلالية للفيديو لاغراض الفهرسة والاسترجاع من الأساليب القديمة حتى أحدث الأساليب. المساهمة الرئيسية هنا هي توحيد المفاهيم واستعراض التطور في هذا الموضوع المثير للاهتمام.