Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software department

Proposing a Recommendation System to Exhibit Product Advertising at Proper Time Stamps of Video

A Dissertation

Submitted to the Council of the College of Information Technology, University of Babylon in Partial Fulfillment of the Requirements for the Degree of Doctorate of Philosophy in Information Technology-Software

By

Heba Adnan Raheem Ali

Supervised by

Prof. Dr. Tawfiq A. al-assadi

2023 A.D **1444 A.H**

Supervisor Certification

I certify that the dissertation entitled **(Proposing a Recommendation System to Exhibit Product Advertising at Proper Time Stamps of Video**) was prepared under my supervision at the department of Software/ College of Information Technology/ University of Babylon in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Information Technology-Software.

Signature:

Date: / /2023

The Head of the Department Certification

In view of the available recommendations, I forward the dissertation entitled "**Proposing a Recommendation System to Exhibit Product Advertising at Proper Time Stamps of Video**" for debate by the examination committee.

Signature:
Prof. Dr. Ahmed Saleem Abbas
Head of Software Department
Date: / /2023

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled (**Proposing a Recommendation System to Exhibit Product Advertising at Proper Time Stamps of Video**) presented by the student (**Heba Adnan Raheem Ali**) and examined her in its content and what is related to it, and that, in our opinion, it is adequate with ( ) standing as a dissertation for the degree of Doctor of Philosophy in Information Technology-Software.

Signature:
Name: Dr.Yahya Mahdi Hadi AL- Mayali
Title: Professor
Date: / / 2023
(**Chairman**)

Signature:
Name: Dr. Majid Jabbar Jawad
Title: Professor
Date: / / 2023
(**Member**)

Signature:
Name: Dr. Alharith A. Abdullah
Title: Assistant Professor
Date: / / 2023
(**Member**)

Signature:
Name: Dr. Mehdi Ebady Manaa
Title: Assistant Professor
Date: / / 2023
(**Member**)

Signature:
Name: Dr. Enas Hamood AL-Saadi
Title: Assistant Professor
Date: / / 2023
(**Member**)

Signature:
Name: Dr. Tawfiq A. al-assadi
Title: Professor
Date: / / 2023
(**Member and Supervisor**)

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:
Name: **Dr. Hussein Atiya Lafta**
Title: Professor
Date: / / 2023
(**Dean of College of Information Technology**)

بسم الله الرحمن الرحيم

صدق الله العظيم

المجادلة / آية ١١

Dedication

I dedicate this work to my parents, my brothers, my husband, my children, my supervisor, and classmates.

I also dedicate it to all those who lit my path and to everyone who supported me.

Heba Adnan Raheem Ali

Acknowledgement

Thanks be to God, Lord of all the worlds, for helping me finish what I started and present this work in the best way.

I would like to express my deepest appreciation and gratitude to my supervisor, **Prof. Dr. Tawfiq A. al-assadi**. His dedication to this work has been limitless. His efforts to ensure the high quality of work content were invaluable. I have been fortunate enough to be taught and mentored by him.

My deepest gratitude goes to all my friends, and everyone who helps me in this work.

Finally, special thanks are due to my family for their help and encouragement while doing this work.

Heba Adnan Raheem Ali

Abstract

The advertisement is a notice or announcement in a public medium to promote a commodity. Advertisements play a substantial part in introducing new products into the market and urge people on to purchase them. They also offer e-commerce companies an opportunity to recommend their goods in videos.

This dissertation aims to implement an effectual video advertising system to display suitable video of product advertisements (ads) to certain user at proper time stamps by recommendation. This takes into consideration video content or semantics (the relationship between video and commodities) as well as current viewing behavior (the user and video relationship).

The system proposed consists of the following main stages. (1) the video structure is analyzed into its primary constituents. (2) the associations between users and videos will be detected through the ORB feature Extraction and FLANN matching methods, and then applying a summarization technique of the original video to identify the important shot in it. Summarization relies on an enhanced clustering technique employing entropy metric calculation and It yields a video summary which contains the key frame extraction of the shot and gets the most important shot with a representative Key Frame. (3) the similarity between the video and product will be detected relying on Yolo deep learning model with the YouTube search engine. (4) the recommendation algorithms based on an optimizer model are proposed to integrate the key frame-product association and the most important shot of the original video so as to recommend an appropriate video advertisement to the user. The GWO and the comprehensive search algorithms have been proposed to search for and find the video ad most similar to the original video and recommend it to the user.

As demonstrated in the experimental findings, the proposed system yields fine results for running efficiency, shot detection and key frame extraction,

video class specification or object detection, and recommendation for video advertising. When compared to other methods, the findings show that the (ORB) based algorithm has a speed advantage for feature extraction. The system results hit 95% in precision and 98% in recall. This makes our approach successfully detect the shots and provide a high recall rate. The results of determining the video category by employing the Yolo deep learning model (Yolov5s) reach 100% in precision. Thus, the proposed method is more accurate in getting the video content. The results of the recommendation and video advertising algorithm demonstrate that all the recommended videos are accurately classified and suggested to the actual type of viewing videos in both the GWO and comprehensive algorithms.

Publication Associated with this Dissertation

I. Published Articles

1) H. A. Raheem and T. A. al-assadi, "Recommender Systems for E-commerce in online video advertising: Survey," *2021 International Conference on Advanced Computer Applications (ACA)*, Maysan, Iraq, 2021, pp. 68-73, doi: 10.1109/ACA52198.2021.9626785.
https://ieeexplore.ieee.org/document/9626785/

2) H. A. Raheem and T. A. Al-Assadi, "Video Important Shot Detection Based on ORB Algorithm and FLANN Technique," *2022 8th International Engineering Conference on Sustainable Technology and Development (IEC)*, Erbil, Iraq, 2022, pp. 113-117, doi: 10.1109/IEC54822.2022.9807488.
https://ieeexplore.ieee.org/document/9807488/

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

Abbreviation	Description
A3P	Adaptive Privacy Policy Prediction
Ad	Advertising
CAVVA	Computational Affective Video-in-Video Advertising
CF	Collaborative Filtering
CIR	Content Information Retrieval
CLD	Color Layout Descriptor
CNN	Convolutional Neural Networks
CTR	click-through rate
DEAR	Deep Reinforcement Learning framework
DL	Deep Learning
DQN	Deep Q-Network
FLANN	Fast Library for Approximate Nearest Neighbors
FN	False negative
FP	False positive
FPR	False positive rate
GMV	Gross Merchandise Volume
GWO	Grey Wolf Optimization
HSV	Hue, Saturation and Value
ICRR	Iterative Co-Relation Regression
KF	Key Frame
KFESS	Key Frame Extraction in the Summary Space
KNN	K-Nearest-Neighbor
ORB	Oriented Fast and Rotated Brief

Poet	Product-Oriented Video Captioner
PSO	Particle Swarm Optimization
R-CNN	Region Convolutional Network
REDAR	REgularized DualfActor Regression
RGB	Red, Green and Blue
RMSE	**A-** Root Mean Squared Error
RPNs	Regions Proposal Networks
RS	Recommendation Systems
SBD	Shot Boundary Detection
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Feature
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TN	True negative
TP	True positive
VSUHCM	Video Summarization Using Higher Order Color Moments
YOLO	You Only Look Once

Chapter One

General Introduction

1.1. Introduction

Advertising is the act of announcing something by a notice or statement in a public setting. Advertising (ad) additionally enables a new product to enter the market. By recommending their products in video form, e-commerce companies may capitalize on this trend and increase product sales [1].

To efficiently mine the link between viewers, videos and items, a new type of e-commerce called video e-commerce is being introduced for widespread usage in online video advertising [2].

Existent online video adverts primarily focus on (ad) insertion at the beginning, during or at the end of video [3]. They spatially substitute a particular area with product advertisement in sports videos, for instance, and personalize ad insertion in an interactive TV setting. Multiple problems arise in these frameworks. First, the video advertisements that are added into video broadcasts are most of the time annoying and uninteresting. Second, typical video advertisements disrupt the flow of the video program as they are presented during the show or sports video highlights. Third, the relevance of rich content videos cannot be measured merely by the textual keywords employed in online video advertising [4,5].

This dissertation aims to transfer shopping attention of users from online stores to commercials seen on streaming video services. On the other hand, personalized product advertising may be shown to specific users at relevant moments in videos (appropriate video time stamps) with full consideration of how individual users, videos, and products are interconnected. The architecture of Video e-Commerce is illustrated in Figure 1.1.

Figure 1.1: Architecture of Video e-Commerce [25]

Video e-Commerce consists of the following main stages:

1- Videos are seen by users.

2- Modeling Key frame-product association.

3- Ranking the importance of video content.

4- Advertising video.

5- Recommending ads to users.

6- Users buy items suggested.

One way to get the most out of a huge video dataset is to have it summarized into a shorter video that gives the same meaning as the original [6]. Video summary may be used for many different purposes. If there is home security footage, for instance, we may edit it down to just a few minutes of relevant footage displaying unusual occurrences for better understand what happened. The same holds true in a sports video when the key moments (goals, penalties, etc.) are highlighted and summed up graphically [7]. The input for video summary is the original video in its entirety. The goal is to choose relevant key frames from the

original video in order to create a brief video that conveys the essentials without requiring the viewer to watch the whole original clip [8]. Due to its length, the spectator may not have time to see the whole film. It is possible that a user may want to watch anything related to the topic of interest. Extracting the most compelling visuals from a video and posting them on social media has become more popular as a means to concisely convey the meaning of a much longer original video [9,10].

In recent years, the popularity of recommendation systems (RS) has grown widely. When customers do not have the background knowledge to make an informed decision, RSs assist in gathering information and making a choice [11]. Their assistance in data organization might greatly benefit the client. They are methods used by software systems to help users discover fresh content that suits their preferences by sifting through vast amounts of data. Finding out exactly what a user loves and dislikes is essential for making relevant recommendations [12]. Because they play a key part in getting new items , recommendation algorithms have a significant role in online video advertising. Most recommendation systems will generate many suggestions using the traditional recommendation methods, as shown in Figure 1.2 .

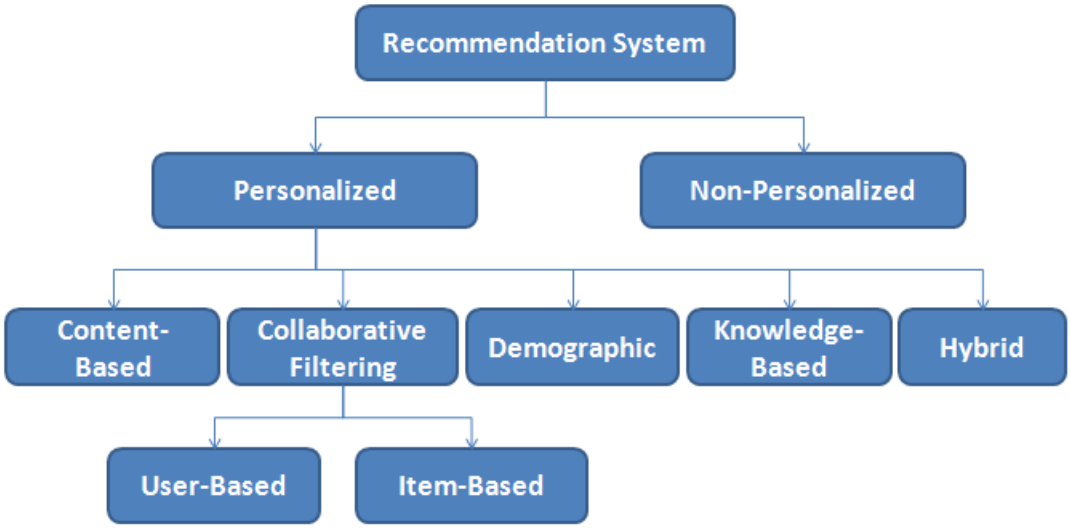


Figure 1.2: Types of Recommendation Systems [11]

1.2 Related Works

Recent studies on online video advertising by suggestion have taken individual viewers' preferences into account. This dissertation is closely related to shot detection and video summarization, as well as online advertising and product recommendation. These will be briefly reviewed below.

1.2.1 Shot Detection and Summarization Techniques

There are some related works that give insights on research that has come before regarding shot detection and video summarization.

In Priyanka S et al (2015) [13], the authors have attempted to study the work of some of the video shot detection techniques used such as ChiSquare, Intersection, ECR, and Scaled Metrics. They read the input video frame by frame, and then each frame is transformed into a gray model. Next, for this different video, shot detection techniques are applied. They proved that shot detection plays a crucial part in browsing, retrieving, classifying and detecting all events. Video shot detection offers simple jump points for browsing videos and determining their category.

Li, F.et al (2016) [14], proposed a scene mutation change detection algorithm combined with (SIFT) feature point extraction. First, using the (SIFT) technique, the feature points of two neighboring video frames are retrieved and their numbers are counted. Matching feature points are then counted up after doing image matching on the two adjacent video frames. In order to evaluate the scene change by this ratio, the ratio between the number of matching feature points in the present frame and the number of those in the antecedent frame is determined. When images are matched, the suggested method can detect scene change. As a result, the method cannot be widely applied, but also it ensures accurate scene change detection.

In Hannane, R et al (2016) [15], the authors used Local Double Thresholding to identify frame transitions after calculating (SIFT) visual discontinuity values. To determine whether a transition is present or not, this technique computes and compares the distances between the computed histograms of two adjacent frames using an adaptive threshold. In order to extract the key frame, they applied (SVD) to every frame in the shot before using entropy. Entropy is then calculated, and the frame with the highest level of entropy values is chosen as the key frame for the shot.

In S. P. Algur et al (2016) [16], the evaluation of entropy-based information theory is used to recognize shots within video sequences. Entropy value is utilized as a local and a global feature as well. In terms of global feature, entropy value is utilized to add a new frame into a cluster. The first cluster is initiated once the first frame entropy is obtained. However, new clusters are initiated whenever a different entropy value is obtained. In this regard, the proposed method has utilized a square of entropy value. On the other hand, entropy is also employed as a local feature. The results show that it cannot identify

transitions like fading or cutting without first estimating joint entropy and mutual information across frames. Entropy metrics have an impact on the degree of relevance of a video shot and video summarization.

In H.Z.N et al (2018) [17], a reliable approach for measuring shot importance has been offered by the authors. The strategy has taken into account elements that have an impact on how significant a video shot is. The shot importance equation takes into consideration length, activity, the average interest point, and camera motion. Weights were chosen depending on the results of the experiment. Shot length was given the least weight among these four parameters because it was discovered during the examination of several shots that some long shots exhibit little activity and are nearly stable. It was clear that among the other factors utilized in this work, shot activity was the most crucial one. It is, hence, heavily weighted.

Yue et al. (2018) [18] suggested a technique for the loop closure detection algorithm and key frame (KF) extraction using robot motion data. To extract the (KF), they used both linear and angular velocities. The time interval is based on linear velocity when it is small. The (ORB) descriptor was used by them as a feature. The results demonstrate that the proposed method can extract appropriate key frames at key points on robot trajectory and can be used in loop closure detection.

Based on (KF) extraction, Li et al. (2018) [19] developed the Key Frame Extraction in the Summary Space (KFESS) framework for video summarization. They started by mapping the video data to the summary space, a high-dimensional space. The representative (KF) is then chosen via a learned representation. A perceptual hash technique was then used to remove similar frames from the representative (KF) set. As a consequence, we can get the key frame set after we

filter out similar frames from the representative frame set. Lastly, we set the key frames in temporal order to form the video summary. In addition, the ground truth constituted by filtering out analogous frames from human-created summaries is employed to assess the quality of the summary.

Liang, B et al (2021) [20] proposes a technique for news video summarization based on (SURF) features and a clustering algorithm. Shot segmentation and key frame extraction make up the two parts of the procedure. Our first step is to pre-process the news video in accordance with its unique properties, extract the (SURF) feature from each frame as a local feature, and then match the (SURF) feature points between adjacent frames. After that, using the matched points as a base, we compute the similarity scores between neighbouring frames and use the double threshold method to identify shot boundaries. Following shot segmentation, the colour histograms of each frame in the shot are dynamically clustered. Smaller clusters are combined and the key frames are selected from each cluster to produce the news video summary. The extracted key frames reflect minimal redundancy and can accurately depict a news video description.

1.2.2 Online Advertising and Product Recommendation

There has been a many of recent researches on the effectiveness of recommending videos to viewers on the basis of their individual preferences and interests in internet advertising. Here is a sampling of these research directions.

Z. Wang et al (2013) [21] is a collaborative social content advertising framework presented to recommend users that films be imported and then re-shared inside the social media platform. First, a method for updating the user-content matrix is presented. Video submissions from users are kept up-to-date to serve as the basis for the suggestion. In order to discover the link between viewers and videos, a combined social-content model is developed. This possibly improved

precision in video re-sharing ads and online sales. Weibo traces and experimental results demonstrate the usefulness of a combined social and content recommendation method, which improves upon both the state-of-the-art content-based and collaborative filtering techniques in terms of suggestion accuracy.

In Yadati et al (2014) [22], Computational Affective Video-in-Video Advertising (CAVVA) was proposed. It considers the effective influence of videos alongside advertising. In the first step, the ads identify potential ad insertion sites. In the second step, they conjointly recognize the right adverts in accordance with notions from advertising and consumer research. The results demonstrated that (CAVVA) would improve the user experience and make it easier for advertisers to reach a wider audience. The framework's weaknesses lie in its inability to create targeted advertisements, investigate the semantic viewpoint, or make use of the user's focus on the video content to improve the advertisements.

Cui et al (2014) [23] presented the REgularized DualfActor Regression (REDAR) technique based on collaborative processing. The basic idea is to represent videos with the social qualities of users and represent people with the content attributes of videos. Thus, both may be defined in a shared space that is a union of the two. After that, we will reframe the video recommendation issue as a matching problem based on shared characteristics. By developing an incremental version of (REDAR), the benefits of the suggested technique are successfully used to reduce the sparsity issue and overcome the cold-start problem. When applied to a real-world social network dataset, the proposed approach for video recommendation achieved relative improvements of over 20% above the state-of-the-art techniques.

Yashar Deldjoo et al (2016) [24] developed a novel content-based approach to the video recommendation job. Without expecting any high-level

linguistic choices like genre, cast, or reviews, which are more expensive to acquire, the technique extracts and utilizes the low-level visual elements (lighting, color, and motion) from video material to furnish consumers with individualized suggestions. When they leverage the idea of low-level visual alternatives rather than high-level genre information, they get more accurate video suggestions, making this framework crucial in video advertising. The suggested method works much better on movie trailers than it does on full-length films for extracting obvious elements. The quality of this job may improve not only in terms of accuracy but also in recognizing novelty and diversity.

Z.-Q. Cheng et al (2016) [25] described Video e-commerce, a method for online video advertising that is able to display relevant product adverts to specific consumers based on video content. In order to accurately model the connections between people, videos, and goods, the authors suggest using Co-Relation Regression, a Heterogeneous Information Network, and Video Scene Importance. This guarantees product variety, solves data sparsity and cold-start issues, and highlights the significance of the video. Finally, product recommendations are made using Heterogeneous Relation Matrix Factorization. Currently, this method is being tested on a publicly available dataset. It aims to propose product adverts to specific clients at optimal times. However, the speed of the complete framework may be improved by investigating the potential of large-scale web advertising and performing real-time updates for new items.

Z.-Q. Cheng et al (2017) [26] expanded their aforementioned previous work (2016) using an iterative Co-Relation Regression (ICRR) model that can efficiently mine the associations between users, videos, and goods. To accommodate the need for massive amounts of internet advertising, the model effectively adjusts to new key frames or new things that decrease the complexity of

time. In this way, the whole architecture can be tested in real time using data available online. One potential drawback of this work is the loss of personal privacy that may arise from a viral film being widely distributed online. In my perspective, secure communication techniques are necessary for this project.

Hussain et al (2017) [27] presented automatically Understanding Advertisements in Images and Videos, which is a new challenge that authors set out to solve. Both a large dataset of annotated images of commercials and a companion collection of annotated videos of adverts have been suggested. They often provide analyses detailing the capabilities they want to include in visual systems in order to understand advertisements and demonstrate a preliminary solution for interpreting ad symbols. They also frequently demonstrate question-answering that successfully deciphers the nuanced content of advertisements via the use of subject, mood, comedy, and excitement forecasts. The advantages of the suggested framework stem from its ability to automatically comprehend the tactics used by advertisements.

Gunavathie. M.A et al (2018) [28] suggested a video e-commerce system that implements video advertising on a massive scale in accordance with a predetermined set of rules. The first step is to determine consumer preferences by studying how people interact with online stores. After gathering user preferences, they analyze the content relationship between videos and items, create a relationship between key frames and products, and then decide when and where the advertisement will be displayed inside the movie. A SURF detection technique is also employed for feature extraction. To determine the user's rating or preference matrix, they use a collaborative filtering technique. Specifically, this study improves privacy settings for users' data by proposing a method called Adaptive

Privacy Policy Prediction (A3P). The shortcomings of the work lie in its inability to deal with cold start and its reliance on a very small data set.

In James Hahn et al (2019) [29], a methodology for discovering the efficiency of video advertisements was suggested by the authors. The goal of this study was to use standard video processing methods and a database of human-annotated attributes to estimate the efficacy ratings of commercials scraped from the YouTube platform. It includes some new thoughts on the data collected from the ad. To achieve the best outcomes, they also use a number of learning techniques including support vector machines, decision trees, regression classifiers, and an ensemble of other classifiers. Accuracy in estimating the effectiveness of advertisements on this dataset is shown with an indicated value of 84%. The benefit of the framework is that businesses may save both time and money if they are certain that their advertisements are reaching their target demographics.

S. Zhang et al (2020) [30] proposed a method called Product-Oriented Video Captioner and abbreviated as (Poet) for accompanying a description that narrates the preferred highlights depicted in the product video for the purpose of attracting potential customers with similar interests or preferences. To do this, it first constructs a video graph to extract the product features from the videos and then makes recommendations to clients in the form of movies. Poet then models these features to identify their dynamic relationships both inside and across frames. Additionally, it determines the connection between items and videos by investigating product-scene interactions. The findings prove the efficacy of video graph modeling, with (poet) achieving the best outcomes on product-oriented video data sets and serving relevant product adverts to viewers.

Jun Ikeda et al (2020) [31] present a technique for predicting the click-through rate (CTR) of video advertising and evaluate the variables that affect the

(CTR). Through the use of the multimodal characteristics of online video advertisements, which include video, text, and metadata characteristics, we provide an optimal framework for precisely predicting the impacts in this study. The aim of this study is to predict the effects of online video advertising using data that was already accessible prior to its release, to determine the variables that affect CTR, and finally to offer the fundamental resources for creating video advertisements. In order to benefit from the multimodal character of online video ads, we create the method of feature extraction of metadata and the suppression of overfitting. Results indicate that compared to the baseline, the suggested strategy can significantly improve.

Xiangyu Zhao et al (2021) [32] provide a Deep Q-network architecture (DQN) and a Deep Reinforcement Learning framework (DEAR) for online advertising in recommender systems. It can make decisions about three interrelated tasks at once. These are whether to interpolate an ad into a recommendation list and, if so, which ad and placement are best for doing so, and simultaneously increase ad income and reduce the detrimental effects of ads on user experience. It is important to note that the suggested DQN architecture can utilize two established DQN architectures, each of which can simultaneously assess the Q-value of two or more associated action types. They conduct extensive studies based on a short video site to evaluate their suggested framework. The findings demonstrate that the suggested framework can greatly enhance the performance of online advertising in recommender systems. Table 1.1 shows a comparative analysis between algorithms of online video ads and product recommendation.

Table 1.1: Comparative Analysis between Algorithms of online Video ads and Product Recommendation

Ref.No	Problem addressed	Advantage	Limitation	Dataset used
[21]	A collaborative social-content advertising framework to suggest videos to import or re-share online. Authors offer user-content matrix updating mechanism.	Finding the relationship between users and videos, and significantly improving the accuracy of suggestion.	The algorithms should account for the individual preferences of consumers.	Weibo and Youku traces.
[22]	Computational Affective Video-in-Video Advertising (CAVVA), which considers the emotional influence of videos as well as advertising.	It improves the user's knowledge while also assisting in increasing the earning potential of the advertising content by reducing user disruption caused by ad insertion and increasing user engagement with the advertising content.	The ability to create customized advertising and investigate the semantic standpoint while also applying user attention to video material to improve adverts.	Datasets from prominent internet video collections (advertisements) and YouTube (videos).

[23]	Developing a common attribute-based representation for people and videos in social media. The REDAR approach based on matrix factorization is proposed to enhance the performance and interpretability of video recommendation systems.	By constructing an incremental version of REDAR, this strategy significantly alleviated the sparsity issue and solved the cold-start problem.	The equalization of social and content qualities, the sparsity of user-video interactions, and the cold-start issue.	Videos and a real social network dataset.
[25]	Video e-commerce for online video advertising, which is able to exhibit appropriate product ads to particular users according to video content.	It ensures the product diversity, alleviates the data-sparsity and cold-start problems, and portrays the importance of video content.	The time complexity.	Online dataset.
[27]	Proposing a method for an Automatic Understanding of Image and Video ads to help advertisers have their commercials understood automatically.	This strategy will help viewers know ads better and buy more products.	Scalability and Diversity.	Set of human annotations gathered using Amazon's Mechanical Turk.

1.3 Problem Statement

In the existing system, video based advertising system encountered several problems. It does not consider the user preference. Products and videos are increased dynamically, when the advertising systems face millions to billions of products and users, it cannot deal with large scale online advertising and conduct real time updating. Some of the most commonly used buttons online are the ones that allow users to skip the (ad) or simply close it altogether. If the advertisement system showing (ads) on the video while the user watching this is may be an irrelevant (ads) for user and also user will get disturbed. Sometimes when people want to watch video, unwanted (ads) are placed over and over again on particular key frame. Also it cannot handle the large amount of newly added videos, products and users.

1.4 Dissertation Objectives

The goals of this dissertation are:

1- To propose a method for extracting a set of KFs from each shot to give a good summarization of video and detecting the important shot for it in order to embed the proper advertisement in the most important shot.

2- To find the content association between videos and products and to build the link between key frames and products.

3- To propose a personalized system for recommending product ads to users based on video content and current viewing behavior of user.

1.5 The Challenges of Dissertation

These systems have a number of weaknesses. They are stated below:

1. Some of these systems are non-personalized. That is, they recommend the same product (ads) to all users without considering user preference and video content.
2. The textual keywords employed in online video advertising are not sufficient for gauging the relevance of rich content videos.
3. One of the important challenges is when (ads) are shown at an improper time. It further cannot deal with the large number of newly added videos, products, and users.
4. If the original viewing video is multi-topic, then there is a great challenge about accurately determining the type and content of the video, which mainly affects the recommendation of the appropriate advertising video for the content of the original viewing video.
5. Video content analysis and get high accuracy for object detection is an important challenge, in other words, determining the type or content of the video through the objects in each frame.
6. Dealing with YouTube, as it always gets updates that cause a problem in dealing due to YouTube policies.
7. Obtaining a dataset or videos that represent the views of the user so that it contains multiple topics and not a specific topic, and this was the biggest challenge because most sites contain videos for a specific topic. Also, advertising videos do not have free links to download them because they have a financial cost.

1.6 Contributions

The contributions of this dissertation are:

1. Proposing a summarization method in video file to extract an important shot video and represent the associations between users and original videos (viewing) based on the (ORB) feature Extraction and (FLANN) matching methods, then providing a video summary using the enhanced clustering technique and entropy metric calculation.
2. Obtaining the content of the original video to find the relation (similarity) between videos and products based on the Yolo deep learning model (Yolov5s) with the YouTube search engine.
3. Building an online video advertising system by recommendation. The (GWO) algorithm and comprehensive search algorithm have been proposed to suggest the video ad that is similar to the original video.

1.7 Dissertation Organization

In addition to Chapter 1, the dissertation contains the following chapters:

i. In chapter 2, fundamental concepts are discussed. These include video structure, key frame extraction, feature extraction and feature matching approaches, summarization, object detection and deep learning, online advertising and recommendation methods.
ii. In chapter 3, the proposed system and its major stages are presented.
iii. In chapter 4, the results obtained from implementing the proposed system are presented and discussed.
iv. In chapter 5, conclusions are drawn and future works are suggested.

Chapter Two

Theoretical Background

2.1 Introduction

Currently, customers are increasingly drawn to the convenience of online buying [1]. Advertisements (ads) based on user recommendations are starting to play a bigger role in the world of online video. Depending on historical data and the ever-increasing demand, they provide product recommendations [31]. So, with the proliferation of internet video, e-commerce businesses now have a platform from which they promote their items via in-video recommendations [32]. In order to provide product suggestions to customers, businesses often utilize recommendation algorithms that are trained automatically based on data collected from customers. Recently, recommendation systems have gained widespread use. When a customer does not have the background knowledge to make an informed decision, they might enlist their help in gathering information and making a choice [2]. The introduction of new items to the market is an important function of recommendation systems in the context of online video advertising. Accordingly, the goal of the recommendation-based online video advertising is to locate material that will pique the viewer's attention and be useful to the recipient. Predictive modeling and heuristics are used on the available data to generate user-specific lists of relevant and engaging content [25,26].

2.2. Video Structure Analysis and Preprocessing

Video summary, video browsing, video compression, analytics, online video advertisements, and so on all rely heavily on the results of video structural analyses [33]. Despite their large file sizes and low information sensitivity, videos often have a complicated structure made up of the scene, shot, and frame, as illustrated in Figure 2.1.

With the use of video structure analysis, we may break down a video into its component parts, each of which can be analyzed independently for their meaning. Shot boundary detection (SBD), key frame (KF) extraction, and segmentation (scene analysis) are all components of video structure [34]. The video shot is a series of frames captured by using a single camera in an unbroken run, as frames are the smallest elements of video. While shots are easily defined, scenes that have any kind of semantic significance are more difficult to pin down [35]. A scene is a sequence of photographs that thematically constitute a semantic unity, a single time and place. The scene segmentation issue is a big challenge due to the inclusion of semantic information in defining scenes. It may also be thought as video clips made up of groups of shots all related to the same theme [36].

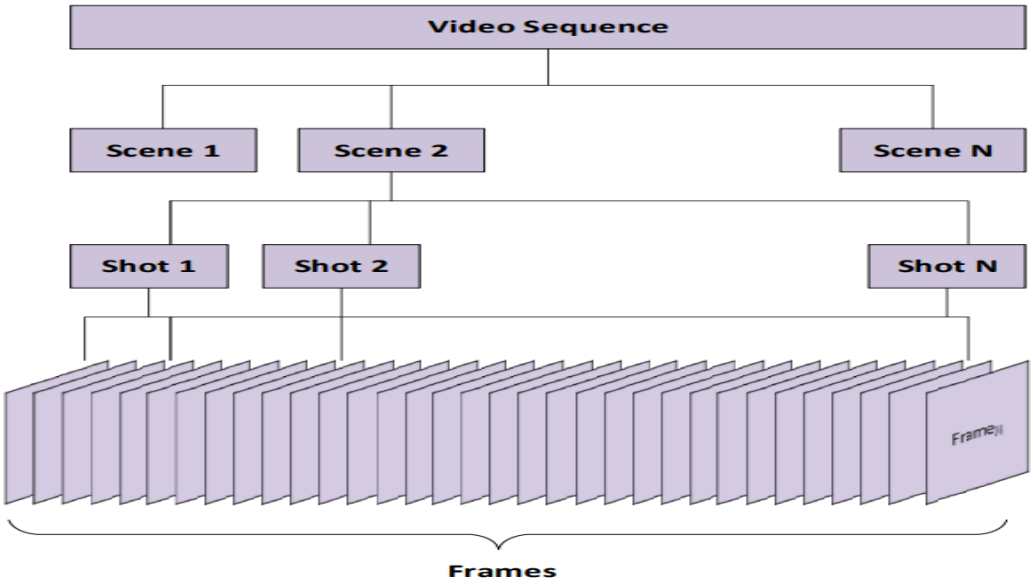


Figure 2.1: The Structure of Video [33].

Frames may be extracted from an input video, and any sounds in those frames can be eliminated using video preprocessing. In order to break a video down into its constituent parts, it is necessary to identify the cuts between shots. The beginning and finish frames of a shot are shown after the shot boundaries are located in order to pinpoint the exact vantage point from which the image was

taken [37]. It is possible to define the shooting limit in terms of a dividing line between two shots. By using shot boundary detection (SBD), the video may be cut up into individual shots. SBD is an important and relevant stage because it has far-reaching consequences for later stages. Because of fluctuations in both the surrounding environment and the quality of the acquiring network, noise might be present in the collected video and image. Applying an image filter to improve received signals is one solution to the noise problem [38].

2.2.1 Key Frame Extraction and Visual Similarity

The extraction of the key frame (KF) is discussed extensively motivated by the fact that the video itself is inherently redundant. This technique involves cutting out unnecessary frames to reduce the movie file size. The term “extraction” refers to the process of selecting frames or sequences of frames that accurately depict an image. The frames should keep what makes the shot memorable while cutting out most of the filler [39]. For this reason, the use of KFs is often regarded as a significant development in the fields of video summarization and shot detection. A video's most crucial shot may be quickly identified with the help of a condensed version of the footage.

The KF extraction process is always accompanied by the problem of determining the size of frames set. Several techniques and features are employed to extract KFs from video clips. For example, a sequential comparison between frames is used and the color histograms of the current frame are compared with the histograms of the previously extracted KF [40]. KF extraction is a major topic and it is inspired from the nature of video because redundancy is a primary characteristic of the video. The redundant frames can be eliminated to make the video more compact by KF extraction which is the process of extracting frames or

sets of frames that give a good representation of a shot. The frames must preserve the salient feature of the shot, while removing most of the repeated frames [41].

2.3 Feature Extraction and Video Matching Approaches

A feature is a single integer or a collection of numbers computed from a digital image. Detecting and matching features is crucial to several tasks including pattern recognition, image stitching, image clustering, content information retrieval (CIR), and object or class identification [42].

In order to help the user interpret and make connections between the vast amounts of data included in images, the extraction phase helps to transform that data into several statistical variables. In addition, feature extraction is most important since the success of the classification process depends on the particular features that are available for discrimination [43].

The images may be used to infer both local and global properties. Global features are often used in image similarity, pattern matching, and classification. Local features, on the other hand, are typically used for object recognition and image matching [44].

The extraction phase aids in translating the dense information of images into a plethora of statistical numbers that are difficult for humans to detect and connect. More importantly, feature extraction is the most critical step since the quality of the classification process depends critically on the features used for discrimination [45]. The size of the raw image is often substantially larger than the extracted components. A better performance and efficiency in image processing are achieved by dimensionality reduction. Feature selection comes first in this phase, and it plays a crucial part in gauging the level of difficulty of the analysis and its overall efficacy. It is the system requirements and the developer's expertise

that usually guide the selection of features. A collection of vectors computed from the original image is the output of the feature extraction method [46].

2.3.1 Global Feature Extraction

When applied to an input image or object, global features allow for generalization with a single vector. Perhaps the most well-known worldwide property of an image is its histogram [47]**.**

As a modeling of the probability density values of gray levels, the characteristics of a histogram may be categorized as statistically based features. The image's gray scale values and their characteristics may be gleaned from the image's statistical features. Depending on the pattern of data, the histogram may take on a wide variety of forms [48].

2.3.2 Color Spaces

The RGB color space is among the most popular options. The image's color distribution is represented by the image's R, G, and B layers. In addition, most modern display systems use the RGB color model. However, the human visual system does not map onto the RGB color space Thus, RGB is often converted to other color spaces, for instance, Hue, Saturation and Value (HSV) [20,49].

The HSV color space is regarded as superior to the RGB one as HSV reflects the perception of the human visual system. It is often used in many software applications such as Photoshop. Furthermore, it is preferred for the image contrast enhancement due to its sturdiness to color deformation. Therefore, it is used in many video applications such as computing shot activity for shot importance measurement [20,49]. The color similarities between each pair of consecutive frames in the shot determine the shot activity (see Equations 2.1 and 2.2) [17].

$$ColorSim(x,y) = \sum_{h \in bins} \min(H_x(h), H_y(h)) \quad \ldots (2.1)$$

where H_x is the HSV histogram for frames x and H_y represents the HSV color feature for frame y.

$$ShotActivity = \frac{1}{b-a} \sum_{f=a}^{b-1} (1 - ColorSim(f, f+1)) \quad \ldots (2.2)$$

In which a and b represent the initial and end frames of the image or scene [17].

2.3.3 Local Feature Extraction

Several image points are used to calculate local image characteristics. The image processing difficulties of occlusion and clutter are therefore better handled [50]. The local characteristics of an image sometimes have connections to many properties. This category includes everything from blobs and areas to edges and corners. These characteristics define the core elements that may summarize the meaning of the supplied image [50]. Local features offer a powerful means that may be employed in a lot of various computer recognition, pattern recognition, and robotics applications, like real-time visual monitoring, image extraction, and video summarization [51]. Several extraction methods have been proposed to extract features from an image. These include Scale Invariant Feature Transform (SIFT), Speeded Up Robust Feature (SURF), Oriented Fast and Rotated Brief (ORB), etc. The (SIFT) features are robust against transformations in viewpoint, light, noise, affine distortions, and occlusions and clutter [15] .The (SURF) algorithm is both quicker and more resistant to variations in lighting conditions [52]. The (ORB) method is noise resistant and rotation invariant [18]. Image scaling, rotation, and illumination modifications have little effect on these descriptors.

2.8.3.1 Oriented FAST and Rotated Brief (ORB)

The (ORB) method was proposed by [53] for extracting distinctive image features to build an extremely fast binary descriptor. These characteristics

are unaffected by image scale, rotation, or lighting changes. The (ORB) is two orders of magnitude faster than the (SIFT) and (SURF) crucial point detectors as well as descriptors, while doing similarly well in many studies, according to the researchers. As a result of its great efficiency and low cost, the ORB approach is desirable. Binary Robust Independent Elementary Features (BRIEF) is the latest alternative for (SIFT) and it needs a computation complexity less than that of the (SIFT) with nearly identical matching performance. A rotation matrix is computed in (ORB) by utilizing the patch orientation and then the (BRIEF) descriptors are directed to the orientation (see Figure 2.2).

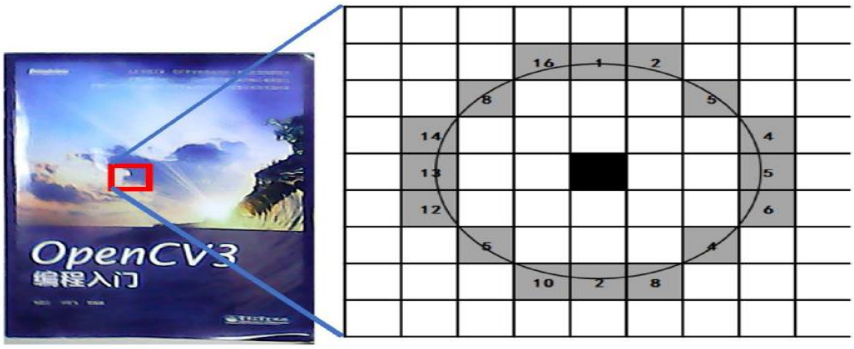


Figure 2.2: Feature Extraction and Description Using ORB [54]

The video matching technique is pretty straightforward given the image alignment approach. The image alignment technique is used to analyze prospective frame matches in the search for feasible pairings between images in the primary and the secondary movies. Once a matched second frame is located for each time delivery, the secondary image is warped to align with the time delivery. The program produces a new edition of the secondary image that is temporally and spatially aligned with the original video. The ORB feature extraction and matching steps processes are sorted as follows [53]:

1. Extracting feature points and finding the location of the key points by applying the fast accelerated segment test to identify the corners (FAST).
2. Selecting the N best points by Harris.

3. Scale-pyramid transformation and adding the points' direction in Intensity Centroid.
4. Extracting the binary descriptor by BRIFF and getting steered brief.
5. Finding low correlative pixel blocks in greedy algorithm.
6. Receiving a descriptor of features.

The BRIEF algorithm is defined by ORB as follows:

a. Perform each exam on all training patches.

b. Form the vector T by ordering the experiments by the separation from the mean of 0.5.

c. Greedy search:

- Put the first test into the result vector R and remove it from T.
- Take the next test from T and compare it against all tests in R. If its absolute correlation is greater than a threshold, discard it; otherwise add it to R.
- Repeat the previous step until there are 256 tests in R. If there are fewer than 256, raise the threshold and try again.

BRIEF shows significant improvement in the variance and correlation over steered BRIEF.

2.3.3.2 Fast Library for Approximate Nearest Neighbors (FLANN)

FLANN is the abbreviation of Fast Library for Approximate Nearest Neighbors in the Euclidian space . It includes a series of techniques designed for rapid closest neighbor search in huge datasets with high dimensional attributes. For big datasets, it outperforms the Brute-Force (BF) Matcher. To incorporate a more complicated parameter drawing process, the FLANN approach was employed rather than a (BF) matcher. However, it will not locate the finest potential matches. Instead, it is just going to find a good matching candidate [55].

A k-dimensional tree will be utilized to create the FLANN parameters, which is an alternative technique to arrange data structures. A k-dimensional tree is a spatially data structure in computer science that is used to organize points in a k-dimensional space [56].

The closest neighbor search issue is important in many applications including image recognition, data reduction, pattern identification and classification, M.L, document information retrieval, analytics, and data analysis. However, in high-dimensional spaces, this problem seems to be complicated, and no solutions appear to exist that outperform the typical brute-force search. Thus, the FLANN approach is a library that performs nearest neighbor search in high-dimensional spaces using quick approximation. Many computer vision issues are solved with it. In other terms, the key to this approach is to seek for the nearest point by Distance measure from the instance point's neighborhood (see Equation 2.3).

$$D(x, y) = \|x, y\| = \sqrt{\sum_{i-1}^{d}(x_i - y_i)^2} \qquad \dots (2.3)$$

See the (X, Y) If the D value generated is less than one, the range between the current local feature pairs is close. This means that the similarity level of the local feature pairs in the image is higher. The KD-Tree strategy is used in this method to partition the data points in the n-dimensional spatial query brush into multiple unique sections. Its goal is to discover the nearest Distance measure to the input location in the area. The KD-tree structure maintains all Euclidean distances in the n-dimensional space brush, enabling the closest point to the reference point to be found simply and effectively. The top-to-bottom search technique of the KD-tree structure is recursive.

In order to remove the mismatching, the R-ratio between both the distances of the closest and second-nearest neighbors is calculated using the KNN matching method (see Equation 2.4) [54].

$$R = \frac{D(V_p, V_{mq})}{D(V_p, V_{nq})} \quad \ldots(2.4)$$

where Vp denotes the feature map for the characteristic point p, Vmq denotes the classification model for the characteristic point's nearest neighbor q in the image, Vnq denotes the feature representation for the featured point that is actually the second-closest neighbor in the image, and D is the vector separation (see Figure 2.3).

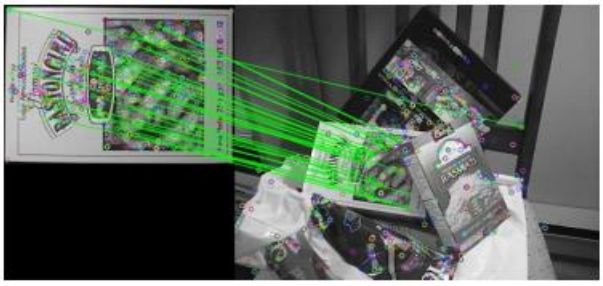


Figure 2.3: Feature Matching Using FLANN [55]

2.4 The Summarization Approach

Video summarization is a means to create a short version of a video to supply the user with a synthetic and useful visual abstract of video sequence in order to obtain the more informative video segments [6].

Although a video file is quite large, it contains redundancy that is worth noting. For this reason, it is necessary in the area of video processing to eliminate as much superfluous info as possible without missing essential details. An abstract for a video is created by studying the structure of the input video and cutting out any unnecessary scenes or segments. An effective video synopsis has to be brief in comparison to the original (see Figure 2.4).

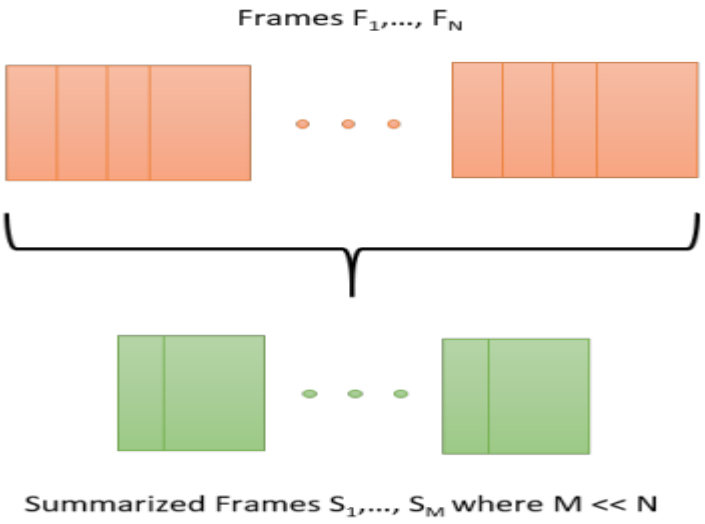


Figure 2.4: Video Summarization Process[57].

There are essentially two distinct types of video summarization approaches used. The first is static, whereas the second is dynamic. A collection of essential frames from the original film are used to create a static video storyboard summary, and there are no limitations on timing or sequencing. However, in order to create the video summary, it picks the most important, brief, and dynamic audio and video clips [7]. The majority of the techniques use visual attributes that are calculated from video frames. Additionally, there are techniques that evaluate the semantic implications displayed in the video to generate a summary that is more useful [8].

In terms of surfing and navigation, a good video overview will allow the user to acquire the essential data on the target video sequence in the shortest amount of time or the most information about just the target video frames in the time limit specified. Summaries that have been manually created can help viewers navigate vast video archives and make judgments about choosing, watching, sharing, or removing content more quickly [10].

Shot boundary identification and key frame extraction are the first phases in the process of video summarization. The video will be divided into shots

using a procedure called shot boundary detection. Choose the key frame that contains the most information possible from each image. Video Summarization Using Higher Order Color Moments (VSUHCM) for Shot boundary detection and key frame retrieval based on image histograms with higher order color moments and skewness that transforms a long dull movie into a brief exciting video. Techniques for video summary include a wide range of topics including e-learning, sports, news, home videos, and movies [9,20].

There are several methods for summarizing videos. Among them are feature, object-based, event-based, and content-based summaries. There is a clear hierarchy present in the various video summarizing methods (see Figure 2.5).

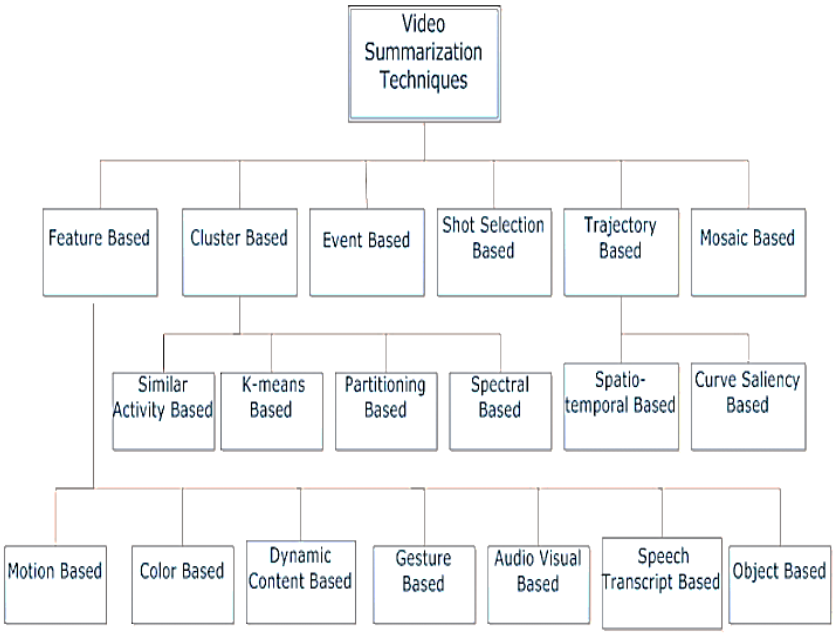


Figure 2.5: Hierarchical Structure of Video Summarization Techniques [58].

As can be seen from the figure 2.5, video summarization techniques will be categorize in 6 major categories based on mechanism and overall process. video summarization techniques are conducted that provides an overview and classification in different ways as video skims, highlights, still image and moving image abstracts, and video summarization techniques for mobile applications.

Techniques are based on features, clustering, event, shot selection, mosaic, and trajectory analysis [58].

2.5 Distance Measures

Standardized distance metrics for use in a metric space are enumerated below. If we define S to be a collection of points, we say that S is a space of points. There are several distance measurements that may be used to determine the distance between any pair of locations in S. As a rule, a distance measure is a real-valued function d (x, y) on some space S that operates on two locations in S. To be classified as a distance measure, a function must adhere to the following criteria, even if it otherwise matches the aforementioned definition [17,59]:

1. *d(x, y)* is larger than zero. (No distance measure is negative).
2. *d(x, y)* is zero if and only if x and y refer to the same point in S.
3. *d(x, y)* equals *d(y, x)* , (Distance should be symmetric).
4. $d(x, y) \leq d(x, z) + d(z, y)$, (inequality property).

2.5.1 Euclidean Distance

The Euclidean distance is the most popular and widely used distance in many areas of computer science and math. If each of the endpoints in the space is represented by a vector made up of n real numbers, then the space is said to be of dimension n in the Euclidean system. The standard representation of the Euclidean measure is as mentioned in Equation (2.3) within section (2.3.3.2).

2.5.2 Cosine Distance

Two vectors, either of which might include Boolean or integer components, are compared using cosine distance. In this context, a pair of points is seen as directions by which an angle might be formed between them. Therefore, the angle formed by any two points equals the cosine distance. The angle always

lies between zero and 180 degrees, regardless of the number of dimensions in the vectors involved [17].

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula (see Equation 2.5) [49].

$$cosine\ similarity = S_c(A,B) \coloneqq \cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i-1}^{n} \boldsymbol{A_i B_i}}{\sqrt{\sum_{i-1}^{n} \boldsymbol{A^2}_i}\sqrt{\sum_{i-1}^{n} \boldsymbol{B^2}_i}} \quad \ldots(2.5)$$

Where A.B is the dot product of the A and B vectors.

2.6 Similarity Measures

This section contrasts with the preceding one by introducing numerous distance or similarity measurements that are often used to evaluate the degree of similarity between two images. Let us assume that (C and R) are two matrices of size (nm). The current image denoted by C is compared to the reference image denoted by R. The next subsections detail many methods for doing this [17,49].

2.6.1 Mutual Information (MI)

As a fundamental idea in information theory, mutual information is hard to overstate. Using MI, you may evaluate the degree of mutual reliance between two variables (see Equations 2.6, 2.7, and 2.8).

$$MI(x,y) = H(x) + H(y) - H(x,y) \quad \ldots(2.6)$$

$$H(x) = -\sum p(x) \log p(x) \quad \ldots(2.7)$$

$$H(x, y) = -\sum pxy(x,y) \log pxy(x,y) \quad \ldots(2.8)$$

Where H(x), H(y) and H(x, y) are entropies of variable x, variable y, and joint entropy of variables x and y respectively. The entropy of the random variable x measures the amount of information that can be learned from an instance of x. p(x) is the probability distribution of x joint entropy. It is a measure of the amount of uncertainty existing in both variables x and y [16,17].

2.6.2 Image Correlation

When comparing two images, correlation is often used. In order to determine the degree of similarity between two images, it uses a correlational analysis. The correlation value is calculated by correlating two images, namely, the test image and the reference image. Correlation values are typically -1 when comparing images that are dissimilar. When the value is close to 1, for example, there is a high probability that the two images are identical. Finding a connection between two images requires a reasonable amount of processing time. In order to determine an image's correlation with another, Equation 2.9 can be used [17].

$$corr(a,b) = \frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(\boldsymbol{a(i,j)}-\boldsymbol{\mu_a})(\boldsymbol{b(i,j)}-\boldsymbol{\mu_b})}{\sqrt{(\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(\boldsymbol{a(i,j)}-\boldsymbol{\mu_a})^2)(\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(\boldsymbol{b(i,j)}-\boldsymbol{\mu_b})^2)}} \quad \ldots (2.9)$$

Where μa and μb are means of a and b respectively.

2.6.3 Relative Entropy (RE)

Distance may be measured using relative entropy, commonly known as the Kullback-Leibler divergence. The program determines the gap between two sets of probabilities. Due to its high processing speed, (RE) is commonly used in engineering. To determine (RE), just plug the numbers into Equation 2.10 [17,60].

$$RE(P,Q) = \sum_{x\epsilon X} p(x)\log(\frac{p(x)}{Q(x)}) \quad \ldots(2.10)$$

Where p and q, are two probability distributions.

2.7 Object Detection and Deep Learning Techniques

The process of detecting and classifying things in still images, movies, and even real-time video is known as object detection. It is one of the most advanced computer vision techniques currently available [61]. In order to do this with new data, object identification models are trained using a large amount of

annotated images. As input images are fed in, an annotated output image is automatically generated. The object detection bounding box is crucial as it pinpoints the item's borders by locating a distinct quadrilateral, usually a square or rectangle. Each one is accompanied with a descriptive label that indicates what kind of thing it is meant to depict, such as a human, an automobile, or a dog. As long as the model has some kind of previous information about the things it is labeling, its bounding boxes may overlap to show off more than one in a single image [62].

The use of deep learning for object recognition allows for a quick and precise method of predicting where an item will be found in an image. With the help of deep learning, an object detector can be trained to recognize objects in images without any human input [61].

Two methods, namely, classical image processing and deep learning networks, are available for object identification. In computer vision tasks, supervised learning is the norm although unsupervised learning is becoming more popular. The growth in GPUs' computational capability from year to year is the limiting factor in performance. Deep learning object recognition is far more resistant to occlusion, intricate sceneries, and difficult lighting, which is a major advantage. Yet, there are some drawbacks. For example, it is necessary to have huge amounts of training data available. In addition, annotating images is a time-consuming and costly operation. To train a unique deep learning (DL) object identification system, classifying 500,000 images is an example of a tiny dataset. However, labeled data is available from numerous benchmark datasets (MS COCO, Caltech, KITTI, PASCAL VOC, V5) [62].

2.7.1 Yolo Deep Learning Model (Yolov5s)

YOLO first appeared in the machine vision scene in 2015 [90]. YOLO is a contraction of (You Only Look Once). It is a tool for Integrated Real-Time Object Recognition, and it instantly drew the attention of other computer vision experts. Before YOLO, CNN such as R-CNN employed Regions Proposal Networks (RPNs). It first generates proposed coordinates on the input image, and then runs a classification on the bounding boxes. Next, it uses post-processing to eliminate duplicate detections and improve the bounding boxes. It was unsuitable for training various stages of the R-CNN system individually. The R-CNN network proved both difficult and slow to optimize [64].

The author's objective is to use a neural network to create a unified representation of all stages. After going through a single network of several convolutional networks, with the input image including or not including the objects, predictive vectors are generated for each object in the image by the system. Instead of iteratively categorizing separate portions of an image, the YOLO system calculates all the characteristics of the image and produces predictions for all items at the same time. That is the concept of (You Only Look Once) [64].

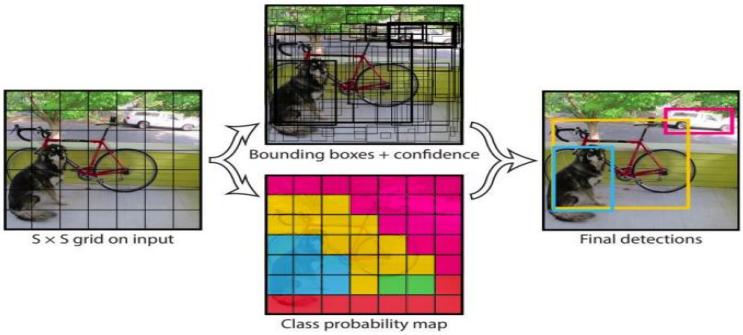


Figure 2.6: YOLO Model with 7x7 Grid Cell [64].

The YOLO algorithm was improved to five versions (including that of the original) during the following five years, with many of the most inventive ideas coming from the computer vision research community.

As a result, the model contains S S grid cells for an image. Each cell predicts B bounding boxes using 5 parameters and shares C class prediction probability. The overall YOLO output of design variables will be S*S (5*B + C). For example, if the YOLO model is evaluated on the famous COCO dataset, which comprises 80 classes, and each cell predicts two bounding boxes, the total production variables are $7 \times 7 \times (5*2 + 80)$ [60].

Yet, YOLOv5 had an edge in engineering. YOLOv5 is written using Python instead of C, as former versions. The work in [63] established the anchor box architecture in YOLOv2 in addition to a procedure for selecting boundary boxes of a size and form which closely reflect the training set's ground truth bounding boxes. The researchers selected the 5 best-fit bounding boxes for the COCO dataset (actually contain 80 classes) and employed them as the default, utilizing the k-mean clustering technique with variegated k values. This both reduces the training time and promotes the network accuracy [63,64].

2.8. Online E-Commerce Recommendation System

The term "e-commerce business" refers to an industry in which products and services are bought and sold online. User's preferences and habits may be deduced with the use of a recommendation system, which also aids in the discovery of new relevant information. An explicit recommendation rule is used to create a suggestion for the user [2]. When utilized properly, recommendation systems may provide a large number of useful ideas using any of the following methods:

- Users' item descriptions and filtering choices are employed in content-based filtering. It is true that the goods are tagged with keywords according to their descriptions, and that the user's past purchases might inform which related items should be shown as suggestions. For the purpose of describing the qualities of an object, we give it a set of keywords [65].

  - Collaborative Filtering (CF) is the most popular technique currently in use. The system is built using collaborative filtering methods, which take into account the user's prior actions (a rating is assigned to items that have already been chosen) and the decisions of other users who have made choices that are similar to the user's. Next, the method is utilized to determine the objects' ratings as a replacement for the user's [5]. Similarity between users or products is used to provide suggestions. The system offers new products to users on the basis of similarities between users or objects in this manner [11].

- With demographic filtering, only information about the user, such as age, gender, work status, location, and possessions, is used to generate recommendations. Recommendations are created by comparing user profiles for shared demographic characteristics [66].

- A knowledge-based recommender system is one that makes suggestions for items to the user on the basis of deductions on the user's requirements and preferences. All approaches for making recommendations may be seen as forms of inference [12]. However, knowledge-based strategies stand out because they make use of actual domain expertise [1].

- Hybrid recommender systems are built by combining several recommender systems to provide a stronger overall structure. In other words, combining several recommender systems may mitigate the weaknesses of one method by leveraging the strengths of another, resulting in a more robust overall system

[31]. The following are some examples of problems that recommendation techniques often encounter:

- Cold-Start: This is an issue that arises if there are any new users or resources introduced to the system. Thus, it is impossible to forecast the tastes of the new users, and the new goods will not be evaluated or purchased by the users, resulting in fewer accurate suggestions [11]. In order to avoid the cold-start issue, you may do things like have the user score certain elements at the outset or tell you what their overall preference is for the design [69].
- Since the majority of users have evaluated a very small sample of goods, this phenomenon is known as data sparsity [70]. "When using (CF) systems, it is difficult to generate accurate predictions regarding things due to the limited number of ratings available. Several methods, such as multidimensional recommendation models, (SVD) methods, demographic filtering, and a content-boosted (CF) algorithm, may be used to address this issue." [71].
- The exponential expansion of online marketplaces has brought with it a new set of challenges related to scalability. Rapid findings from (RS) methods are a need for today's large-scale applications, so their use has become more common. In a short period of time, RSs may look for a number of potential neighbors. However, the requirements of modern e-commerce companies demand that they seek for even more nearby neighbors. One dimensionality reduction is often used to address scalability difficulties. These may be solved by using clustering methods [72].
- Inclusivity: Recommendation algorithms might propose products that are either very close to the one being searched for or completely unrelated. We call this predicament the diversity problem, since it occurs when suggestions are made on the basis of similarities rather than differences. Users are able to learn

information they would not have been able to discover on their own because of the variety of suggestions provided. There seems to be some concern that accuracy might suffer if an algorithm prioritizes diversity improvement above everything else [73].

- When you provide your personal information to a recommender system, you get better recommendation services, but you also run the risk of having problems with data privacy and security. Cryptographic methods may be used to protect sensitive data and prevent unwanted tracking [74].

2.8.1 Video E-Commerce and System Resources

Because of developments in multimedia systems and high-speed networks, a massive amount of data from these systems, including digital video, is becoming more dominant. To put it simply, the movie is "a massive volume data item, it includes high redundancy and dense information." The structure is complex, including several scenarios, camera angles, and individual frames [33].

Both online marketplaces and video content providers must eventually merge. This is due to the fact that video content providers expect to traffic monetization by product display, while "e-commerce platforms introduce their products to videos in order to increase GMV" [25]. Video advertising on the internet is often known as video marketing. YouTube clip content-based product placement remains one of the least-used digital commerce methods. Despite possessing a very positive impact on shoppers, e-Commerce is used to display relevant product ads to the correct individuals by suggestion in order to efficiently extract the relationship between users, videos, and products. The objective is to adapt consumers' typical online buying habits for usage in commercially produced videos. On the other hand, its stated objective is to provide relevant product adverts to appropriate consumers (based on the users' individual preferences).

For mass-market video advertising on the web, video e-Commerce has been provided. It may show specific viewers adverts for products that may be of interest to them, so that user data, video data, and product data can all be mined efficiently [28]. Organizations like Amazon.com and Netflix have made their suggested systems prominent features of their online stores and streaming platforms respectively. Fewer efforts, however, focus on recommending and analyzing video material for internet advertising [26].

2.8.2 Online Advertising

One of the best methods for companies of any size to reach more users, build a client base, and generate more cash is via online advertising. Adults all around the globe spend most of the time on social networking sites like Facebook and Twitter, and advertising have learned to meet them where they spend their time online [1].

The popularity of websites, social media, and mobile applications has led to a surge in the demand for online advertising. Nonetheless, tailored advertising is still a source of debate. We can be sure that our ads are reaching the right user by the success of a campaign in online advertising. This may be measured in a variety of ways, including the number of users who clicked on the ad, the number of users who subscribed to the product, the number of users who bought something, the percentage of users who converted, etc. Furthermore, real time bidding and prediction systems are crucial for the success of the online advertising environment since the number of publishers and advertisers has largely expanded [12,76].

Existing methods for advertising and recommending videos confront several obstacles. One difficulty is that there is an extraordinarily low correlation between consumers and product adverts on e-commerce platforms because of the sheer variety and volume of items available. Current recommendation systems are

plagued by age-old issues like cold-start, data sparsity, and so on. This is especially problematic considering the restricted nature of purchase data for items shown in videos. In addition, many of these systems ignore non-personalized user choices in favor of recommending the same product adverts to everyone [25,26].

The security of shared movies may be at risk (including the user's privacy). The fact that certain advertising appears at inopportune times is also a significant obstacle. Furthermore, it cannot deal with the flow of new films, items, and users [28,77].

2.8.3 Product Recommendation

The topic of recommendation systems is quickly growing in importance for E-commerce and online services. It takes a lot of time to search for different things while purchasing online. A buyer interested in several different items may locate them more quickly with the help of a recommendation system. Because it allows customers to purchase online with confidence and ease, this effective suggestion system is gaining popularity rapidly [28,78].

Providers and consumers alike may win from the deployment of recommendation systems [1]. Previous research shows that recommendation engines aid customers in making more informed purchasing choices, saving time and money [79]. A product recommendation system is a piece of software that is meant to help users find and explore new products, services, and information that they would like. We narrow down on fashion items and create a technique that just takes one input image to provide a ranked list of similar-style suggestions as part of product recommendation's analysis of the current state of affairs [80].

2.8.4 Recommendation Based on Meta-Heuristic Methods

To find, create or choose a heuristic (partial search algorithm) that can offer a sufficiently good solution to an optimization problem issue while dealing

with limited knowledge or computational resources is the goal of a metaheuristic, which is a higher-level method or heuristic [81]. Rather than trying to exhaustively enumerate or otherwise examine every possible solution, metaheuristics instead sample a selection of the options. Certain metaheuristics may be applicable to a wide range of issues without requiring extensive preconditioning of the optimization issue being tackled [82]. Metaheuristics, in contrast to optimization algorithms and repetitive approaches, do not ensure that a globally optimum solution can be discovered for a given class of problems [83]. Stochastic optimization is used in many popular metaheuristics, making the obtained solution conditional on a given collection of random variables [84]. Metaheuristics may typically identify excellent solutions in combinatorial optimization with less computing work than optimizers, iterative approaches, or basic heuristics by looking through a vast range of reasonable options. Therefore, they are effectual methods for tackling optimization issues [85,86].

In the E-commerce industry, promotional tactics are used to drive traffic to the selling platform/online portal. This traffic is converted into customers who pay and efforts are made to retain those customers. The predictive performance of many E-commerce sites is upgrading using algorithms like Support Vector Machine (SVM), Particle Swarm Optimization (PSO), Neural Network, Grey Wolf Optimization (GWO), Machine learning (ML), etc. [87] The main goal of these algorithms is providing the best results for a variety of application areas like predicting customer behavior, generating product titles for E-commerce and predicting quality, predicting relevant products and average ratings for those product lines, and numerous other existent E-commerce scenarios. The goal is to conduct a comparative analysis of several accessible E-commerce approaches [88,89]. The aim is to make a comparison analysis of a variety of available

methods in E-commerce market predictions in addition to some diverse applications.

2.8.4.1 Grey Wolf Optimizer Algorithm (GWO)

Grey Wolf Optimizer (GWO) is an innovative heuristic swarm intelligence optimization technique first suggested by work in [90]. The wolf's superior hunting skills make it the top predator in the food chain. Wolves have a rigorous social structure within their pack, and they like spending time with others. They are categorized into four subtypes (alpha, beta, delta, and omega) to reflect the natural order of the pack's leadership. The top three wolves are given the respective titles of alpha, beta, and delta, while the remaining members are given the designation of omega. Hunting (optimization) in the GWO is directed by alpha, beta, and delta [90]. The top three wolves lead their pack mates to the most fruitful hunting grounds by pointing out where to look. They evaluate the potential location of their prey at each iteration of the seeking process. Wolves' positions are revised according to following formula throughout the optimization process [91]:

$$\vec{D} = \left|\vec{C} \,.\, \overrightarrow{X_p}(t) - \vec{X}(t)\right| \quad \ldots(2.11)$$

$$\vec{X}\,(t+1) = \overrightarrow{X_p}(t) - \vec{A}.\vec{D} \quad \ldots(2.12)$$

where A and C are coefficient vectors, XP is a position vector for the prey, and X is the wolf's current location at the current iteration (where t is the iteration number). A and C may be written as a vector using the notation and calculated as follows:

$$\vec{A} = 2\overrightarrow{a}.\overrightarrow{r_1} - \vec{a} \quad \ldots(2.13)$$

$$\vec{C} = 2.\overrightarrow{r_2} \quad \ldots(2.14)$$

where, the coefficient $\overrightarrow{a}$ linearly decreases from 2 to 0 with the increasing of iteration number, $\overrightarrow{r_1}$ and $\overrightarrow{r_{12}}$ are random vector located in the scope [0, 1].

Grey wolves have the ability to recognize the location of prey and encircle them. The hunt is usually guided by the alpha. The beta and delta might also participate in hunting occasionally. However, in an abstract search space we have no idea about the location of the optimum (prey). In order to mathematically simulate the hunting behavior of grey wolves, we suppose that the alpha (best candidate solution) beta, and delta have better knowledge about the potential location of prey. Therefore, we save the first three best solutions obtained so far and oblige the other search agents (including the omegas) to update their positions according to the position of the best search agent. The following formula is proposed in this regard:

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3} \quad \ldots(2.15)$$

The pseudo code of the (GWO) is as shown in algorithm 2.1[92].

Algorithm 2.1 : Grey Wolf Optimizer (GWO) Algorithm

1. Initialize the grey wolf positions Xi (i = 1, 2, ..., n)
2. Initialize a, A, and C
3. Calculate the fitness of each search agent
4. Xα=the best search agent
5. Xβ=the second best search agent
6. Xδ=the third best search agent
7. while (t < Max number of iterations)
8. for each search agent
9. Update the position of the current search agent by Equation (2.15)
10. end for
11. Update a, A, and C as in Equations (2.13 and 2.14) respectively
12. Calculate the fitness of all search agents
13. Update Xα, Xβ, and Xδ
14. t=t+1
15. end while
16. return Xα (the best search agent)

2.9. Evaluation Criteria

The evaluation metrics were defined on the basis of the confusion matrix, as demonstrated in the equations bellow:

B- Precision

It is computed by dividing the number of TP by that of TP and FP as illustrated in Equation (2.16) [93].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \ldots (2.16)$$

C- Accuracy

It is calculated by dividing the number of right guesses by the total number of forecasts as shown in Equation (2.17) below [94].

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP+TN+FP + FN}} \quad \ldots(2.17)$$

D- Recall

It is computed by dividing the amount of TP by the total amount of TP and FN, as displayed in Equation (2.18) [95].

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \ldots(2.18)$$

E- F1-score

It is the result of 2*((precision*recall)/ (precision + recall)). It is called the f1- measure too. It can be calculated by Equation (2.19) [96].

$$\text{F1} - \text{score} = \frac{(2*\text{TP})}{(2*\text{TP+FN+FP})} \quad \ldots(2.19)$$

F-False Alert Rate (FAR)

The fraction of negative prediction, also known as fall-out or false positive rate (FPR), is incorrectly interpreted as positive (anomaly) for all negative predictions. The lower the value, the better. Equation (2.18) below can be used to calculate this measure [97].

$$\text{False Alert Rate(FAR)} = \frac{\text{FP}}{\text{FP+ TN}} \quad \ldots(2.20)$$

G- Error Rate

It can be computed by dividing the whole number of erroneous predictions by the total number of dataset predictions, as shown in Equation (2.21).

$$ERR = \frac{b+c}{a+b+c+d} \quad \ldots(2.21)$$

Besides, in some cases, it is calculated as follows:

Error Rate = Incorrect Predictions / Total Predictions

Error Rate = 1 – Accuracy

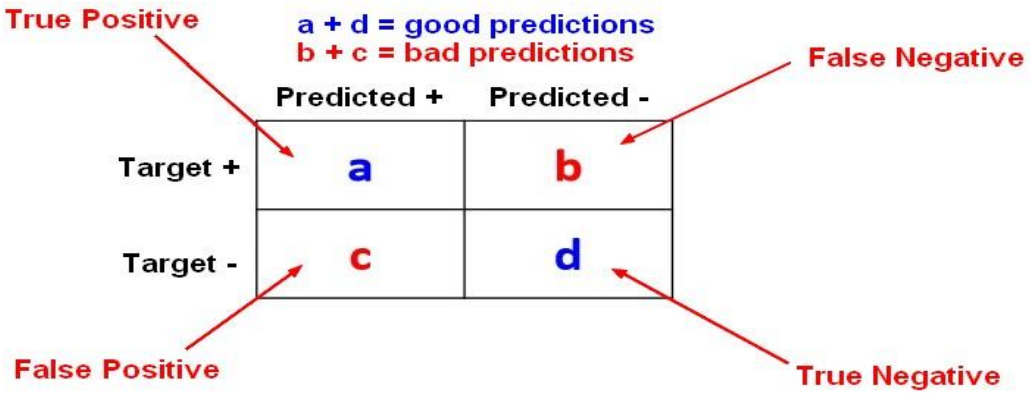


Figure 2.7: Example of Error Rate Calculation [80].

H- Mean Absolute Error

It is a metric used to assess how near the projections or deviations are to the final results. The average absolute error is provided by Equation 2.22 [98]:

$$MAE = \frac{SAE}{N} = \frac{\sum_{i=1}^{n} |x_i - \hat{x}_i|}{N} \quad \ldots(2.22)$$

MAE is the collection of absolute errors (or deviations), N is the number of non-missing data points, $\mathbf{x_i}$ is the time series of real observations and $\hat{\mathbf{x}}_\mathbf{i}$ is the evaluated or forecasted time set.

I- Root Mean Squared Error

This is the square root of the mean squares of the values. The errors are squared before the calculation of the mean. RMSE gives big errors a comparatively

high weight. The RMSE of an algorithm exception with respect to the evaluated parameter $\mathbf{x_{model}}$ is referred to the square root of the mean squared error [99]:

$$\text{RMSE} = \sqrt{\frac{\sum(x_{obs,i-x_{model,i}})}{N}} \quad \ldots(2.23)$$

Where $\mathbf{X_{obs}}$ is the observed values and $\mathbf{X_{model}}$ is the modeled values at time/place i. Relative and absolute values are different as the absolute error is the measure of result deviation from the actual value, whereas relative error is the percentage measure compared to the actual value [83].

2.10 The Confusion Matrix

The confusion matrix is used to evaluate accuracy, recall, and f-measure for assessing the model's performance. The confusion matrix describes a model's performance on a set of test data. It provides the classifier with two types of right predictions and two types of wrong ones. The confusion matrix is displayed in Table (2.1). The predicted output as true positive is denoted by TP, the expected output as true negative is denoted by TN, the predicted output as false positive is denoted by FP, and the predicted output as false negative is denoted by FN [100].

For the assessment of the classification approach, various procedures can be used. The confusion matrix was chosen to get the main performance measures for assessment purposes. It is employed to characterize the performance of a classifier or a classification technique by test data [100].

- **True Positive (TP)**: It denotes the positive cases which have been appropriately categorized.
- **False Negative (FN)**: It points to the positive cases which have been misclassified.
- **False Positive (FP)**: It denotes the negative cases which have been mistakenly predicted and categorized.

- **True Negative (TN)**: It denotes the negative cases correctly forecasted by the classification model.

Table 2.1: Confusion Matrices

Confusion Matrix		**Predicated Class**	
		Positive +	**Negative -**
Actual Class	**Positive +**	**TP**	**FN**
	Negative -	**FP**	**TN**

Chapter Three

The Proposed Approach

3.1 Introduction

After presenting the essential background in the previous chapter, the proposed system is introduced in this chapter. It is composed of major stages, each containing several stages. In general, the contribution of the proposed system is divided into main parts. The first part develops a new scheme to find a shot important and key-frame relatively based on video summarization. The second illustrates deep earning and video class detection. The third part develops a method for an online video advertising system by recommendation.

3.2 The Proposed Methodology

The proposed methodology is based on a video advertising system to exhibit appropriate video product ads to particular users at proper time stamps. This takes into account video content (semantics or relation between video and products) and viewing behavior (user and video). In the proposed framework, the first stage will be the analysis of video structure into its primary components. The next one will be detecting the associations between user-video and video-product. Finally, the recommendation system is proposed to integrate key frame-product association and the important shot of original video to recommend appropriate video advertisement (ad) to the user. The proposed system consists of the following main stages:

a) Preprocessing: splitting the video into its primary components.
b) Video summarization and important shot detection in order to find proper time stamps of videos that represent the associations between users and original videos.
c) Obtaining the content of the original video with the YouTube search engine to find the relation between videos and products.

d) Building an online video advertising system by recommendation. Figure 3.1 shows the block diagram which demonstrates the methodology of the proposed framework:

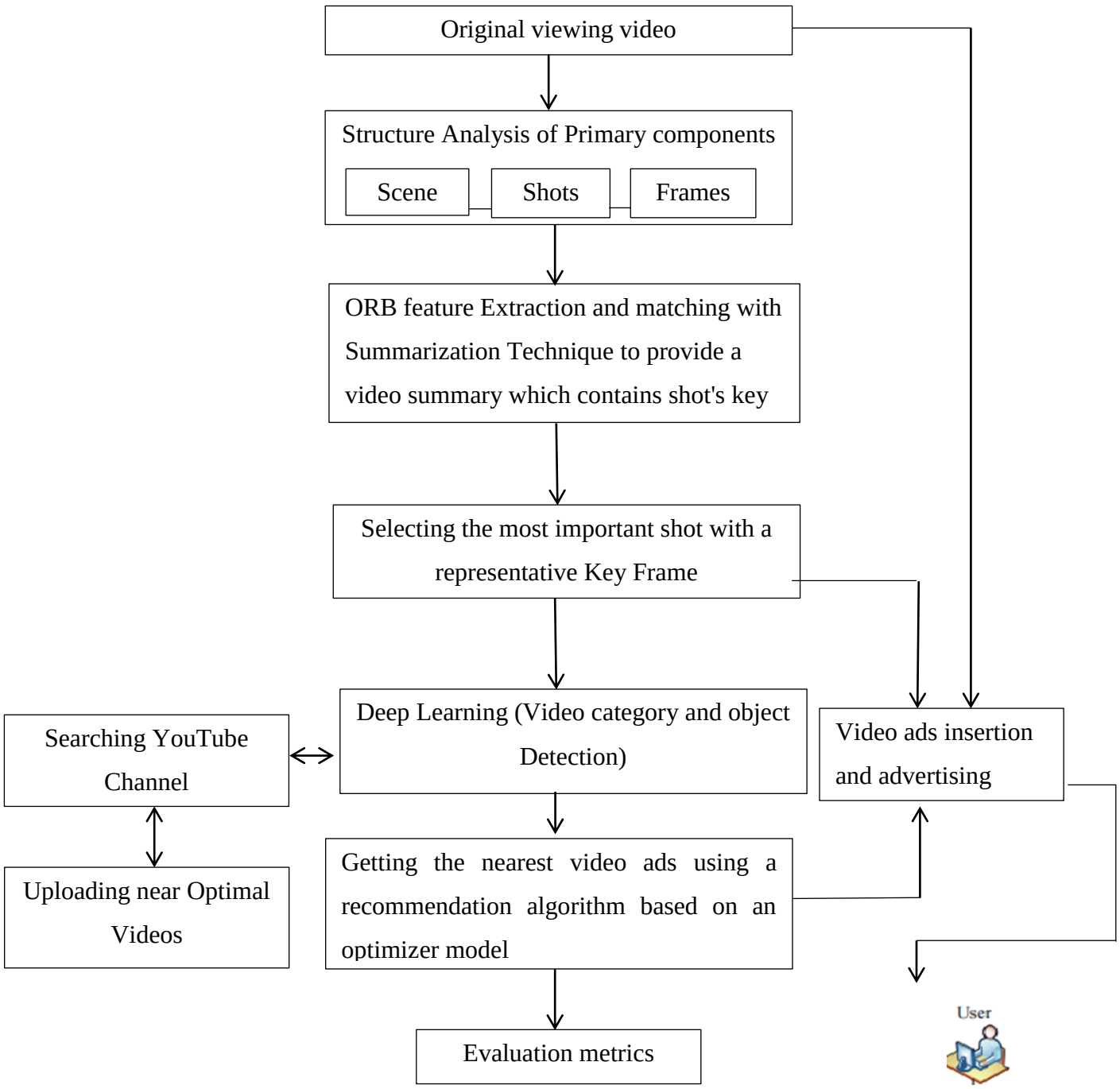


Figure 3.1: The Block Diagram of the Proposed System.

As shown in Figure 3.1, the proposed system is implemented on the basis of the following stages:

- Stage 1: Using the YouTube data API v 3.0 to connect python with YouTube servers.
- Stage 2: Selecting the original video using GUI and analyzing it into its important components.
- Stage 3: Extracting points (features) of interest from each frame in the shot by using the ORB algorithm.
- Stage 4: Computing the similarity between features using the FLANN method.
- Stage 5: Applying summarization for each scene using the entropy metric to find the important shot and its representative key-frame in the video.
- Stage 6: Object detection to obtain the content of the original video based on the Yolo deep learning model (Yolov5s).
- Stage 7: Searching the YouTube channel for the advertisement video to bring the ads videos similar to the uploaded video.
- Stage 8: The YouTube search engine will reply with a list of videos. We analyze the YouTube search results and compare the time of the suggested advertisement video with that of the main video.
- Stage 9: Applying the recommendation process using the recommendation algorithm based on an optimizer model for the purpose of returning the appropriate advertising video to the original video.
- Stage 10: Playing the viewing (main) video and embedding the advertisement video before the important shot.
- Stage 11: Playing the recommended advertisement video, and then the rest of the main video.

3.3 Video Structure Analysis

This stage involves splitting the video into its major components. The structure of the viewing video file will be analyzed in order to obtain its primary components (scenes, shots, frames) following the stages below:

a) Converting the video to a sequence of frames.
b) Transforming frames (RGB (24 bits)) to 8-bits gray scale.

The color layout descriptor (CLD) is used to determine the color spatial distribution. The frames are first taken from the video. The retrieved frames are color space transformed from Red, Green, and Blue (RGB) to a gray scale. Each gray scale frame is divided into 8x8 non-overlapping blocks. Therefore, the template images and the image compared in the video feed are gray scale treated before matching to remove noise. On this basis, histogram equalization is applied to the input image. This results in enhancing not only the amount of feature points in the source images and the template image, but also the amount of correctly corresponding feature points.

One of the motivations for choosing gray scale photographs rather than color images is to reduce the (RGB) color space to a single dimension description, i.e. monochromatic. This is required to reduce calculation costs and simplify processing. Some applications do not require color images, and utilizing them may increase the number of training data required to get effective results.

The shot boundary detection (SBD) technique attempts to divide the video into temporal parts (shots). This stage seeks to identify the sequence of frames acquired by a single camera operation. Frames in the same image frequently feature significant contrast. The Correlation and the shot boundary detection (SBD) stage determines the boundaries between shots.

The goal of extracting (KF) sets is to provide a compact and efficient representation of the shot. One of the most notable properties of video is its significant redundancy in video frames. This stage attempts to reduce the majority of the duplication while retaining the majority of the shot's key qualities. Key frames detection and feature extraction are shown in the next section for getting important shots to find the time stamp.

3.4 Important Shots and Video Summarization

There are some components to the video summarizing technique in this suggested study. They are as follows:

A- Using shot boundary identification based on (ORB) characteristics, the entire video is first divided into separate video segments (scenes and shots).

B- A summary for each scene is applied so that each shot's key frame extraction is realized using the enhanced clustering technique based on entropy metric calculation.

C- A summary of the video is produced by organizing key frames sequentially.

D- The entropy metrics based clustering for video summary, shot length, and shot activity factors have been used to determine shot relevance. After that it makes a matching between the frames using the (FLANN) algorithm for the purpose of finding the similarity ratio between them to detect the shot boundaries. Next, a summary will be made for each shot or scene to find important shots based on the entropy metric.

E- A comparison is made between shots that are obtained in each scene to determine the most important shot with its relative FK for all the video. The suggested method is displayed in Figure 3.2.

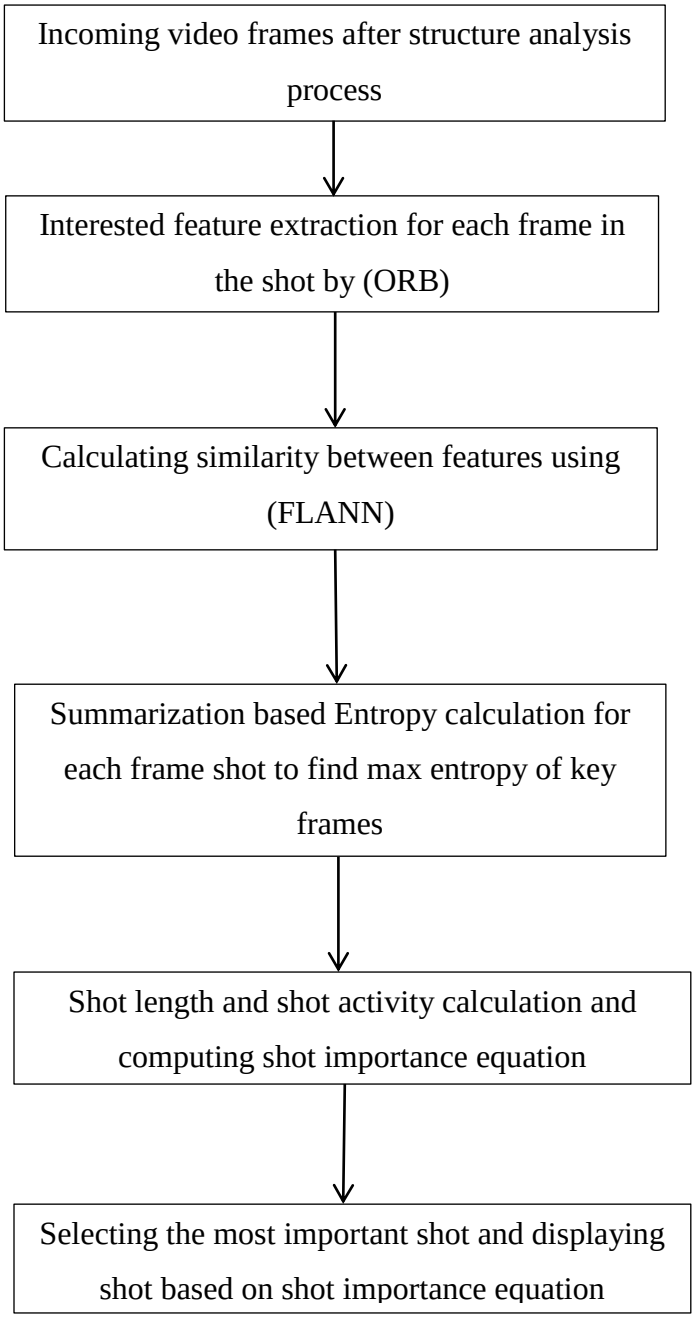


Figure 3.2: Framework of Detecting the Important Shot Based on Video Summarization

The main stages for detecting important shots based on video summarization are as follows:

3.4.1 Extracting Points of Interest Using (ORB) and (FLANN)

As mentioned in section 3.4, a video is represented as a set of frames, and then the (ORB) method is used to extract and describe the distinctive image features from the video. The feature points are pre-matched using the (FLANN) method. After that, the incorrect matching features are eliminated using the KNN algorithm.

The template image and each pixel of the video feed are initially processed in gray scale. And to improve feature point quality with matching efficiency, the notion of integrating histogram equalization is offered. Later, the (ORB) descriptor and (FLANN) matching method will be applied. The features sort for each frame and its following frame in the form of a tree, putting an index for them. Next, (FLANN) will do the comparison and compute the ratio of similarity using the Euclidian distance. In other words, the key is to search the nearest point from the nearest neighborhood of the instance point by Euclidean distance as mentioned in Equation 2.3.

So in our proposed method, if the distance ratio is less than 0.7, this means we have a good match. After that, by using a threshold T=3, for example, which means we must have at least three features that are similar, if the number of extracted features is more than or equal to 3, this means that it is in the same scene. Otherwise, it means we have moved to another scene. Subsequently, in order to eliminate the mismatching, the KNN matching described in Equation 2.4 is used to determine the R-ratio between the ranges of the closest and second-nearest neighbors. According to the above, the best and worst feature matching are discovered for each feature point and the index of each scene has been detected. Therefore, (FLANN) provides a hierarchical clustering index that can be used for matching features.

Based on the foregoing, the proposed method proposes merging the image enhancing technology with (ORB) method and utilizing them to process the incoming image frames in advance using histogram equalization. As a result, when used in conjunction with image enhancement technologies, the (ORB) method increases extraction of features and matching in order to determine shot or scene boundaries. Algorithm 3.1 illustrates the stages of extracting points of interest from each frame in the shot using the (ORB) and (FLANN) methods.

Algorithm 3.1: ORB Feature Extraction and Shot Detection Algorithm

Input: video frames
Output: video shots

Begin

1. For i=1 to number of video frames do
2. Convert each frame to gray scale
3. Eliminating noise and improving contrast by a histogram equalization
4. Feature points extraction by ORB algorithm and put the index for each feature
5. For j= 1 to total number of extracted feature point do
6. Sort the feature points based on FLANN
7. Computing the ratio of similarity between the features using the Euclidian distance as in Equation 2.3.
8. Eliminate feature mismatching using KNN algorithm as in Equation 2.4
9. End for j
10. End for i
11. Return the video shots and scene boundaries.

End

3.4.2 Video Summarization and Important Shot Calculation

After using the ORB and FLANN methods for extracting the features in order to obtain the similarity between them by comparing each frame with the next and detecting the shot or scene boundaries, a method of video summarization for extracting KFs using clustering is described.

Important frames need to be further extracted within each shot once the video has been divided into distinct shots. Using a clustering approach, a representative frame may be chosen from each shot in the scene as the KF.

The computation of an HSV color histogram, histogram clustering, merging tiny clusters, and selection of the KF are the four essential stages in the clustering-based KF extraction in shots. They are detailed below:

Stage 1: The generated vector is normalized by basing each bin on the number of pixels in the image after transforming each frame to an HSV color space and calculating the histogram to the number V.

Stage 2: Between the initial frame and each subsequent frame in the shot, the cosine distance (and angle) is calculated and added to the same group by calculating the cosine distance between two vectors in the current frame y and the first frame x. After converting each frame to the color space HSV and computing the histogram towards the number V, the resultant vector is standardized by basing every bin on the amount of pixels present in the image.

Stage 3: The histogram (for V value) is calculated from each group. The resulting vector can be viewed as the center of a cluster. The frame with the smallest difference from the average histogram is found for each group.

Stage 4: Computing the entropy metric for each frame in each scene to find the frame having the max entropy.

Stage 5: A collection of these representative frames for all the recognized scenes in the video sequences results in a video summarization at the end of the process. Algorithm 3.2 explains the proposed method of video summarization using clustering for KF extraction:

Algorithm 3.2: Key Frame Extraction with Video Summarization Using Clustering

Input: video shots
Output: set of key frames (video summary)
Begin

1. Initialize MaxEntropy= 0, MaxSceneEntropy =0
2. For each frame in the input shot do
3. Convert it to HSV color space
4. Finding V histogram
5. End for
6. Consider the 1st frame as reference frame R1
7. For the remaining frames do the following
8. Compute the cosine distance with R1
9. Segment or hash the frames into specific cluster based on angle R1
10. End for
11. Merge the small clusters with the previous cluster
12. For each cluster C do
13. For each frame G in cluster
14. entropy = shannon_entropy(G)
15. if entropy > MaxEntropy
16. MaxEntropy =entropy
17. IndEntropy=G
18. End If
19. End for
20. if entropy > MaxSceneEntropy
21. MaxSceneEntropy =entropy
22. IndSceneEntropy=C
23. End if
24. End for
25. Return the video summary with its representative key frames

End

In order to calculate the important shot with its representative (KF), shot length and shot activity measures have been used. At first, the number of frames in the shot being processed is represented by the shot length. Next, the shot activity is determined based on the color similarity between each pair of adjacent frames in the shot (see Equations 2.1 and 2.2).

After obtaining the length, activity and entropy metrics of the shot, the shot importance is calculated as in Equation 3.1:

$$Shot\ importance = w1.L + W2.shotact + W3.entropy \quad \ldots \quad (3.1)$$

Where **L** length is the length of the shot, shotact is the activity of the shot, and entropy is the metric of summary for the entire video. w1,w2, and w3 are empirically specified weights given to each one of the shot features, such that the summation of all weights equal to 1. The shot importance detection algorithm is as follows:

Algorithm 3.3: The Shot Importance Detection Algorithm

Input: Set of video shots.
Output: Shot importance scores.
begin
1. For i = 1 to length of shots do
   i. Compute the number of frames in each shot (shot length).
   ii. Calculate shot activity using Equations 2.1 and 2.2.
   iii. Entropy calculation for each frame in each shot to find representative key frame within max entropy as in Algorithm 3.2.
   iv. Apply shot importance calculation as in Equation 3.1, and then select and return the most important shot.
2. **End for**
End

3.5 Deep Learning and Video Class Detection

In this section as mentioned above, the class and object detection need to be determined to obtain the content of the original video based on the yolo deep learning model (Yolov5s). It is clearly shown that the online deep learning model consists of many predefined models.

To make the video analysis process more comprehensive, it is necessary to collect the information of each shot in a single 3D matrix. This matrix could be

analyzed in many ways to obtain the highest possible accuracy to determine the category of the video by using a pre-trained model.

Suppose we have the video (Vd) and the important shot (D), then we must apply the weighted mean measure as in Equation 3.2 to calculate and find the category average in video.

$$M_i = \frac{\sum_{J=1}^{N} \sum_{k=1}^{O} R_{k,i}^{J} * D_k^I}{\sum_{J=1}^{N} \sum_{k=1}^{O} D_K^J} \quad \dots (3.2)$$

Where:

N: Number of shots

O: Number of Objects in shot

$R_{k,i}^{J}$: Rate of object k in shot J that belongs to category i

D_k^J : Number of duplicated object K in shot J

This equation will help to find out the percentage of the object's belonging to any of four categories. The standard ratio is also needed to know for this category in the video, as in Equation 3.3.

$$SD = \sqrt{\frac{\sum_{J=1}^{N} \left[\frac{\sum_{J=1}^{N} \sum_{k=1}^{O} R_{k,i}^{J} * D_k^I}{\sum_{k=1}^{O} D_k^J} - M_i \right]^2}{N-1}} \quad \dots (3.3)$$

The standard measures are the percentage of the object spread on the whole video. If the object is spread over the entire video, it is the real candidate and not the object being concentrated in one place. The frequency of appearance of this object within the video also needs to be taken into account. All of these metrics will help to determine the real category for the original video by analyzing its components. Algorithm 3.4 explains the proposed method of the video class detection algorithm based on the yolo deep learning model.

Algorithm 3.4: Video Class Detection Algorithm

***Input**: Set of video shots.*
***Output**: The category of the video.*
Begin

1. For each shot do
2. For i=1 to number of frames do
3. Apply Yolo to detect all possible objects
4. Categorize each object into one of the specified categories
5. For j=1 to number of Categories do
6. Calculate mean and standard deviation
7. Compute the fitness (F) or score value (F= 0.9 mean + 0.1 S.d)
9. End for j
10. End for i
11. End for
12. Find the max fitness value (max f = max (fitness vector))
14. Decide the category of the video

End

3.6 API YouTube Channels and Similar Videos

After detecting the type of the entire video based on the video class detection in Algorithm 3.4, the YouTube channel is then searched for the advertisement video. The YouTube search results will be analyzed and the time of the suggested advertisement video will be compared with that of the main video. YouTube videos and their associated information descriptions, such as user comments/opinions, have become a data treasure trove. This information might be useful in a variety of study areas, including online video advertising. The proposed framework consists of the following four steps:

(a) Searching the YouTube channel for the advertisement video.

(b) Collecting the similar ads videos.

(c) Collecting video metadata.

(d) Replaying the list of videos with its metadata. The proposed framework consists of the following sub-stages:

1. Utilizing search Feature: To get a search request corresponding to a user's search on the YouTube website, the YouTube analytics API includes a resource called (Search). In addition to keywords, users may give additional parameters such as stream identification. So the categories of the gathered videos are controlled with publish after/before and sort to return the results. The gathered metadata includes:
   a) Metadata Structure: As a result, the acquired metadata may be classified into two types: Invariant data (a) and dynamic data (b). Invariant data includes:
      i. Video which consists of (Video ID, Video publish time, Duration, Video title, and Video description)
      ii. Video Category (the type of video)
      iii. Channel which consists of (Channel ID, Channel publish date, Channel title, and channel description)
   b) Dynamic data includes:
      i. Number of comments.
      ii. Number of likes.
      iii. Number of views

2. Metadata collection process.

After bringing the ads videos similar to the uploaded video and retrieving their metadata including comments count, viewing count, likes count, and video ID, the list of advertising videos will be entered on the next stage to recommend the appropriate ads video to the user. Therefore, the primary problem for collecting a big volume of YouTube schema on a continual basis is to find a huge number of video IDs in a timely and automated manner. Algorithm 3.5 shows the stages of the proposed framework of continuous YouTube videos collection with related metadata.

Algorithm 3.5: YouTube Video Ads Collection within Related Metadata

Input: video category, channel ID, API key.

Output: list of similar ads.

Begin

1. Fetch a list of video ID and indexes
2. Temporary buffer storage for saving the list of videos
3. For Index=1 to total number of video list length do
4. Fetch the video metadata with:
    i. Videos Title
    ii. Videos Views
    iii. Videos Likes
    iv. Videos Comments
5. End for
6. Return list of similar ads videos and related metadata description
7. Pass the returned list to recommendation algorithm

End

3.7 User Recommendation and Optimizer Model

After video ads collection within related metadata will be returned as in Algorithm 3.5, the personalized online video advertising is achieved using a recommendation algorithm based on an optimizer model for the purpose of returning the appropriate advertising video to the original video.

This study depends on the current user views only and analyzes the video content to find similarities between it and the advertising video. After that, it selects the closest or best advertising video for the offer. It is a modern trend in the subject of recommendation as it depends on finding the best and most appropriate element to display among the proposed elements (items). So we used the same idea to find the best proposed advertisement among all proposed advertisements to

display and include it within the original viewing video. Two algorithms have been used to search for and find the best video ads similar to the original video.

3.7.1 Grey Wolf Optimizer Recommendation Algorithm (GWO)

The concept of this algorithm depends on filtering videos obtained through the two previous processes: object detection and YouTube search engine, to find the closest ad video that is similar or appropriate to the original video content, in addition to having suitable parameters such as (likes, comments, index, and views). In other words, the (GWO) algorithm classifies advertising videos in a hierarchical manner, where the most appropriate video is at the top of the pyramid.

The (GWO) algorithm works hierarchically to find the ad closest to the watched video. Then it continues to work until the stop condition is met, so that the best video is at the top of the pyramid. There are several stopping conditions for this algorithm. They are as follows:

1. Reaching a fixed number of iterations (i).
2. Access to determined specifications such as (Fitness<= specific number) .
3. The difference between one iteration (i) and another is very small.

Evaluating all ads in each iteration depends on computing fitness or objective function based on getting information about each ad video to get the best results (see Equation 3.4).

$$\text{Fitness} = \alpha * \frac{1}{\text{no.of likes}} + \beta * \frac{1}{\text{no.of Views}} + \gamma * \frac{1}{\text{no.of comments}} \quad \dots (3.4)$$

Where α, β, and γ are coefficients which represent the weights that will be multiplied by each of the corresponding parameters (likes, views, comments) to get the best value for fitness (score value or objective function).

In the next process, the algorithm will do the top ranking of the most suitable ads based on their fitness values. This procedure will be continued based

on updating (GWO) parameters to get the best or the most similar video ad to the original video from the last iteration.

The similarity between the best video ad and viewing video is calculated based on computing the Euclidean distance between the best video ad and other video ads (see Equation 3.5).

$$\text{Eqlid}=\sqrt{(I-I_b)^2+(L-L_b)^2+(V-V_b)^2+(C-C_b)^2} \quad \ldots(3.5)$$

Where:

I, L, V, C represent index, likes, views, and comments of the ad video, whereas I_b, L_b, V_b, and C_b represent index, likes, views, and comments of the optimal (best) ad video. The proposed framework for getting the similar video ad uses the (GWO) algorithm as shown in Algorithm 3.6.

Algorithm 3.6: Selecting the Best Advertisement Video Using the (GWO) Algorithm

Input: list of similar ads videos and related meta data.
Output: the best ads video
Begin

1. Fetch all possible suitable ads
2. Get the info. about each ads video
3. For i=1 to max number of iterations do
4. Calculate all fitness Values
5. Evaluate all ads based on fitness using Equation 3.4
6. Top ranking of the most suitable ads
7. Update the GWO parameters.
8. End for
9. Choose the most suitable ads from the last iteration using Equation 3.5.

End

3.7.2 **The Proposed Comprehensive Search Recommendation Algorithm**

In this work, this algorithm relies on a comprehensive search for all the entered and suggested advertising videos from the YouTube channel, depending on the user query. Each of these advertisements with their information (index, likes, views, and comments) will have a differential score calculated according to the fitness equation. Next, the best ad among them will be selected by choosing the advertisement with the highest rating (min fitness value).

As mentioned in Algorithm 3.6, the evaluation of all ads depends on computing fitness or objective function as in equation (3.4) and is based on getting information about each ad video in order to get the best results. Thus, the proposed framework for getting the similar video ad is shown in algorithm 3.7.

Algorithm 3.7: Selecting the Best Advertisement Video Using a Proposed Comprehensive Search Algorithm
Input: list of similar ad videos and related meta data. **Output:** the best ad video **Begin** 1. Get all possible suitable ads. 2. Get the info. about each ad video. 3. for i=1 to total number of ad video do 4. Calculate the fitness value using Equation 3.4 5. End for 6. Find the max fitness value. 7. Decide the best ad video. **End**

.

3.8 Online Video Advertising

Advertising Videos and displaying appropriate product ads to users by embedding them at a proper timestamp in the important shot of the video is considered the final stage of the recommended process.

As shown in Figure 3.1, personalized video advertising has been illustrated and achieved by a few procedures. The first stage is the user-watching video relation, which indicates the commonalities between viewers and key frames. It is used to determine the best insert product adverts position using the summarizing approach. The second one, known as key frame-product relation, is the relationship between key frames and products. It combines user preference with video semantics based on deep learning and the YouTube search engine. After getting the nearest video ads using any of the proposed recommendation algorithms, the advertisement video is displayed, whose input parameters are the original video, advertisement video, and the most important shot index. Based on the specified condition, the algorithm will play the main video and embed the advertisement video before the important shot. After that, the recommended advertisement video will be played with the rest of the main video as shown in Algorithm 3.8.

Algorithm 3.8: Online Video Advertising Display Process

Input: *The original video, the advertisement video, the most important shot index*.
Output: *Displaying the recommended ad video*.
begin

1. Initialize shot index = 0
2. While (Index <> shot index) do
3. Display the original video
4. Index = Index +1
5. End while
6. Stop playing original video
7. Display advertisement video.
8. Playing the rest of the original video.

End

Chapter Four

Experimental Results

4.1 Introduction

This chapter illustrates the proposed system's implementation and experimental results, as well as comparisons with some related works. This chapter is composed of major sections. Section 4.2 describes the datasets used. Section 4.3 shows the implementation results. Section 4.4 demonstrates the experimental and analysis results of the proposed system.

4.2 Datasets Description

In this part, the datasets that were utilized to test the suggested system are described. We employed the widely used TVSum50 Benchmark Dataset [101] in the suggested research for video summarization and shot extraction. This dataset contains 50 video sequences collected from YouTube using 10 categories from the TRECVid MED task. These videos cover a wide range of topics, including flash mobs, news, video blogs, and more. A self-built video dataset that was gathered from the internet was used to verify the method's viability. Tables 4.1, 4.2 and 4.3 show the details of the datasets.

Table 4.1: The Details of the TVSum50 Dataset

Cat.	Descriptive Statatistics		
	#vid	#frm	length
VT	5	39,841	25m25s
VU	5	35,014	19m28s
GA	5	30,920	18m7s
MS	5	37,095	24m58s
PK	5	41,634	24m50s
PR	5	44,042	25m3s
FM	5	30,747	18m37s
BK	5	30,489	17m30s
BT	5	25,747	14m39s
DS	5	36,827	20m59s
Total	50	352,356	3h29m42s

We utilized the SumMe [102] dataset for additional testing, which was constructed as a standard for video summarizing. The 25 videos in the collection

cover sports, events, holidays, and other topics. The videos range in length from 1 to 6 minutes.

Further, a collected video dataset that was gathered from the internet was used to verify the method's viability. The popular online video collections were utilized for getting advertisements and YouTube videos dataset from the most popular video sharing site, as shown in Table 4.2:

Table 4.2: The Description of the Collected Video Dataset

Dataset Name	URL	Dataset Type
Open Video Project	https://open-video.org/	Mp4 and avi video file
Google Video	http://video.google.com/	Mp4 and avi video file
TRECVID	http://www-nlpir.nist.gov/projects/trecvid/.	Mp4 and avi video file
youtube8m	http://research.google.com/youtube8m/download.html	Mp4 and avi video file
Metacafe	http://www.metacafe.com/	Mp4 and avi video file
ucf101	https://www.crcv.ucf.edu/research/data-sets/ucf101/	Mp4 and avi video file

Table 4.3: The Description of All Datasets

Dataset	# of Videos	Duration (min)	Content
TVSum	50	2-10	flash mobs, news and video blogs
SumMe	25	1-6	holidays, events, cooking and sports
Collected Dataset	50	1-10	Cartoon, sports, cooking, transport and animals

Figures (4.1, 4.2, 4.3, 4.4 and 4.5) demonstrate a very small description of some of the dataset used

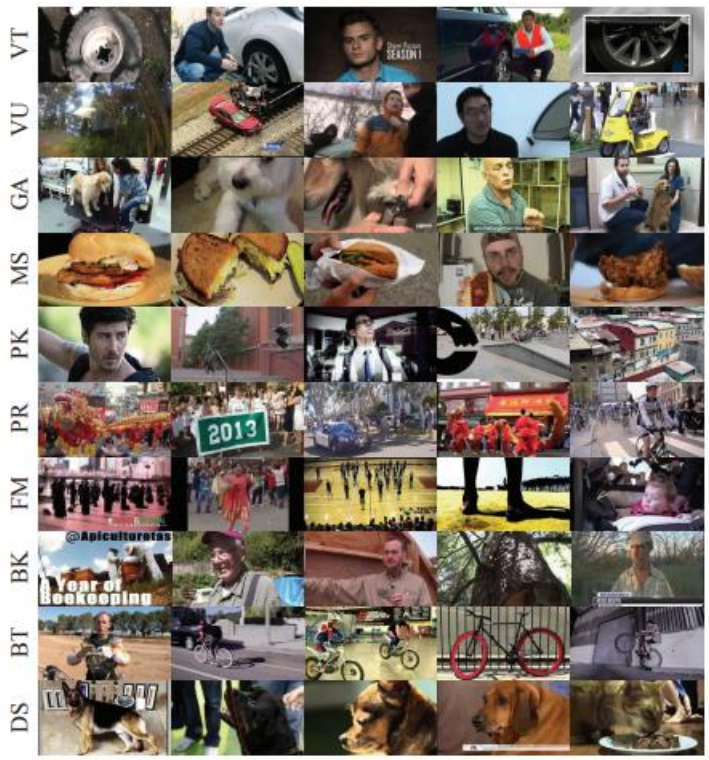


Figure 4.1: TVSum50 Dataset [101]

Figure 4.2: TRECVid MED Task Dataset [103]

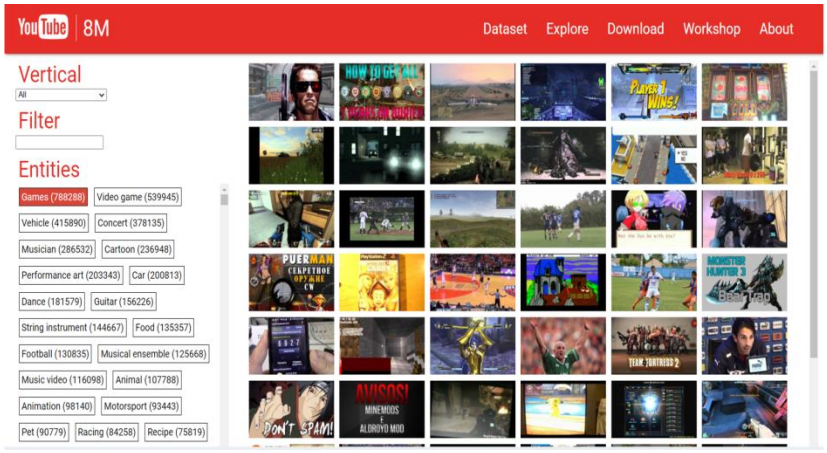


Figure 4.3: YouTube8m Dataset [104]

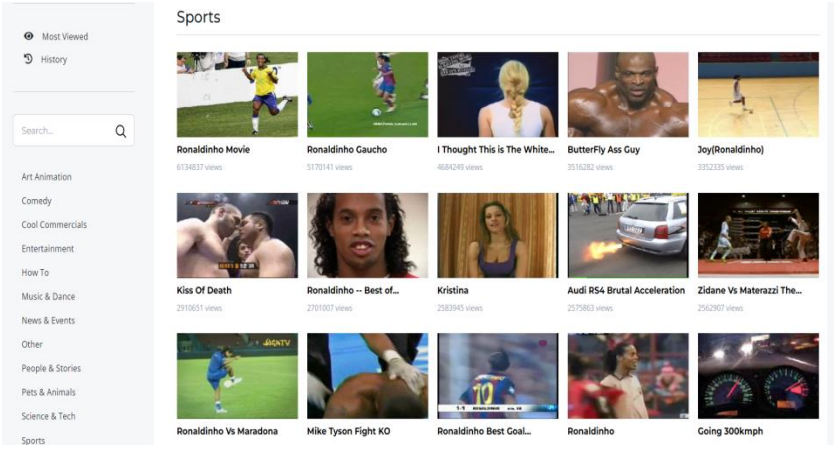


Figure 4.4: Metacafe Dataset [105]

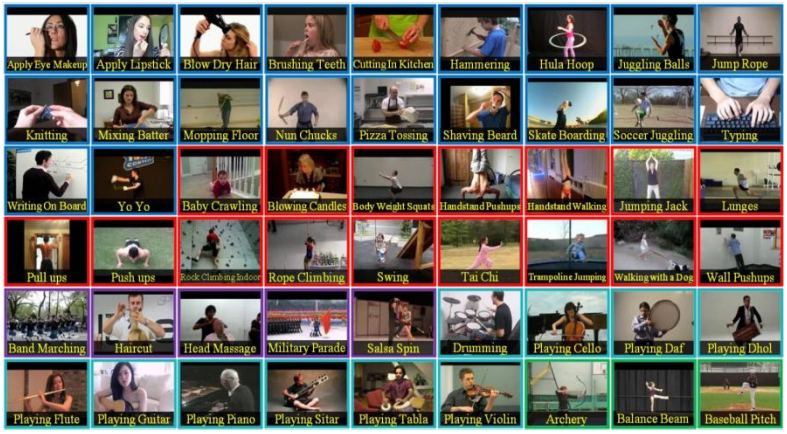


Figure 4.5: ucf101 Dataset [106]

4.3 The Implementation Results

This section describes the results of running the stages of the proposed system. This system is constructed in Python connected with YouTube data API v 3.0 and runs on a personal computer with the following specifications: Windows 10 with Intel Core i7, 1.50 GHz CPU and 8 GB RAM.

4.3.1 Video Preprocessing

As mentioned in chapter 3, each frame of the video stream and the template image are first processed in gray scale as shown in Figures (4.6 and 4.7).

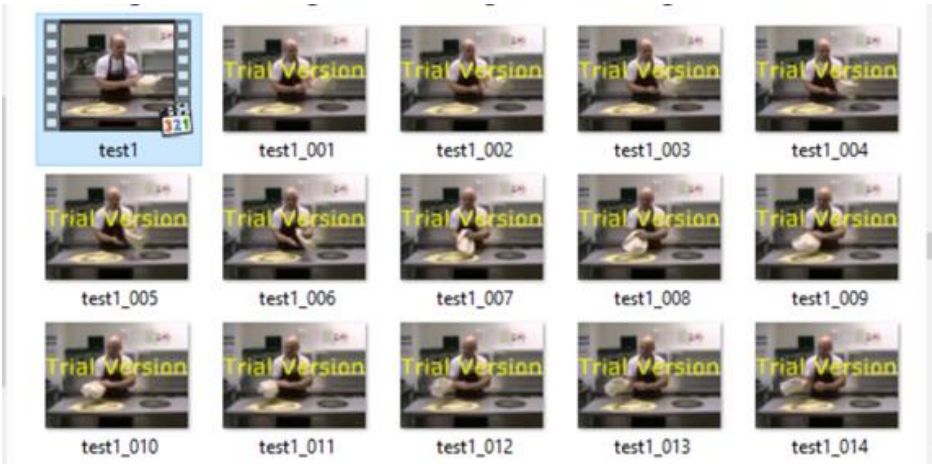


Figure 4.6: Converting a Video to Sequence of True Color Image *(Frames)*

Figure 4.7: Gray Scale Image Frames after Converting

Next, a histogram equalization is used to process the image to be entered into the video stream because it can make the image's gray probability distribution uniform and its histogram as accurate as possible.

The equalized image is produced after multiple transformations by computing the local histogram of the image and then redistributing the brightness. The histogram equalization modifies the contrast of the image. As a result, the algorithm can enhance the image's contrast and get more image details.

4.3.2 Implementation of Video Summarization and Important Shot Detection

The findings of implementing the suggested method to many examples of videos to produce a video summary and get the important shot based on the ORB algorithm as shown in Figure 4.8:

Figure 4.8: The Implementation Results of Summarizing a Sport Video

The chosen video has 298 frames according to the implementation findings on a sports video. The video under processing should then be divided into its primary shots or scenes in order to be further studied. The shot boundary is generated by measuring the amount of change between key frames. The chosen

video's scene borders are [1, 28,139] as a result. This indicates that the frames from 1 to 28 are similar, and the first move begins from frames 28 to 139. Later, each scene is summarized using entropy measurements to determine the maximum entropy for each scene or shot in order to identify the most important frame based on equation 3.1. We also show another example of the implementation results of a cooking video in Figure 4.9:

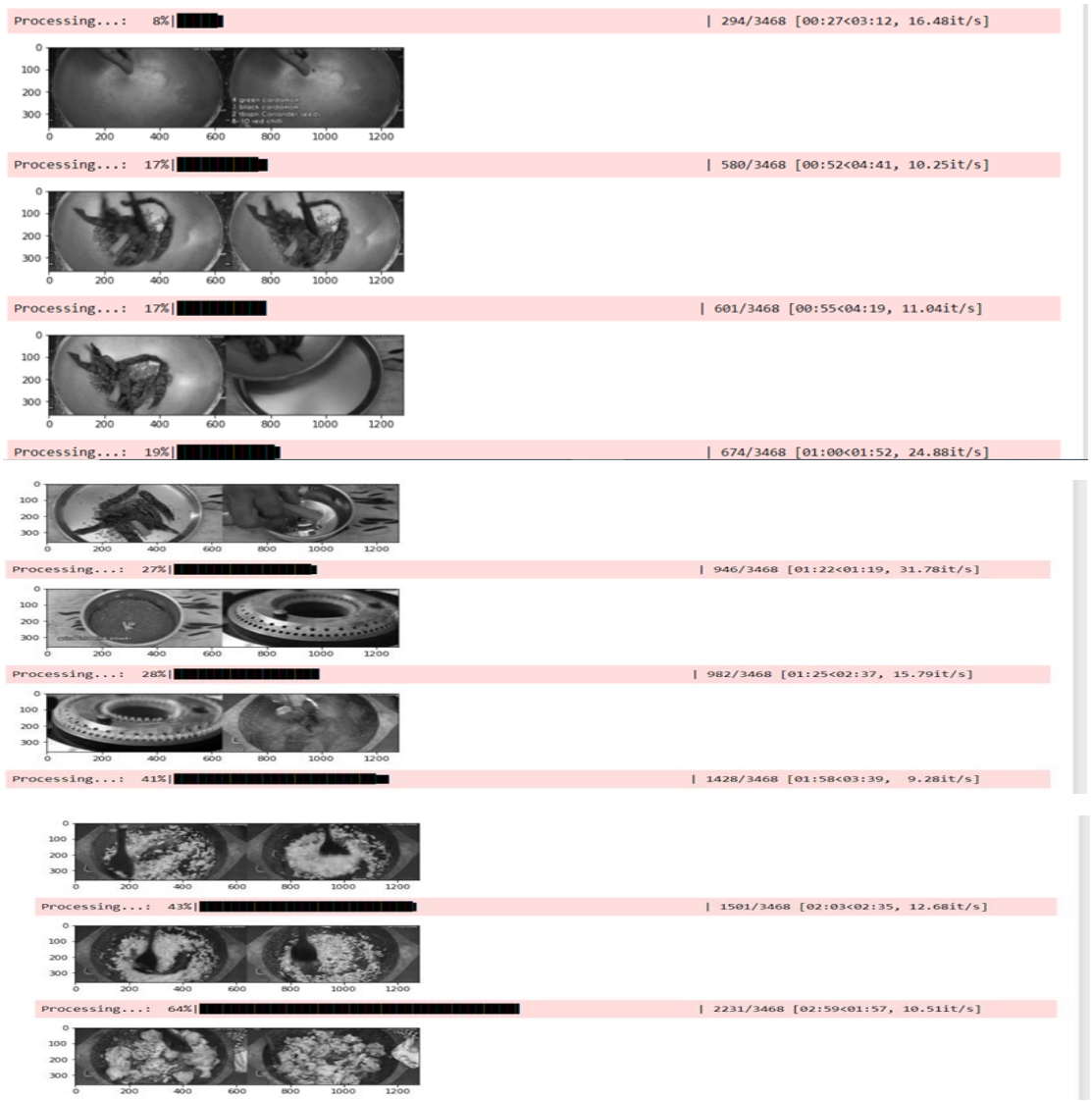


Figure 4.9: The Implementation Results of Summarizing a Cooking Video

The implantation results of Figure 4.9 show that a cooking video consists of 3468 frames, which are summarized into 11 scenes. The results of the important shots for three different videos are shown in Table 4.4.

Table 4.4: Shot Importance Metric for Video Summarization

Video	No. of frames	No. of shots	shot	Length of shot	Shot activity	entropy	Important shot
V1	298	3	0	28	0.136244	6.72409	**0.543955**
			1	110	1.079743	7.16624	**0.920573**
			2	160	1.04667	7.16959	0.992343
V2	3468	11	0	294	2.025776	7.676914	**0.690151**
			1	287	0.763116	7.677914	**0.614612**
			2	21	0.305103	7.676924	**0.522384**
			3	74	0.676018	7.676814	**0.557109**
			4	270	1.311491	7.676914	**0.642488**
			5	36	1.678162	7.675914	**0.606345**
			6	446	2.173106	7.676913	**0.736165**
			7	74	1.094283	7.666914	**0.581561**
			8	730	1.592474	7.676912	**0.772103**
			9	1016	4.276390	7.686434	1.0
			10	220	0.292053	7.686424	**0.571453**

V3	430	2	0	39	1.221937	0.024296	**0.441522**
			1	391	1.026632	7.145896	0.796163

For the purpose of excluding several shots, importance score is computed for each shot after the video is temporally segmented. In Table 4.4, rows consist of three different videos in terms of length, length of shot, activity of shot, and entropy in each shot; while its columns contain the results of each factor. The variety of shot importance can be noticed in the Table .The important shot will be obtained in scene 2 for v1,in scene 9 for v2, and in scene 1 for V3. The obtained results are subjectively evaluated by observing and analyzing shot features. The results are considerably related to the shots importance. Figure 4.10 bellow shows the relation between frame number and shot importance score.

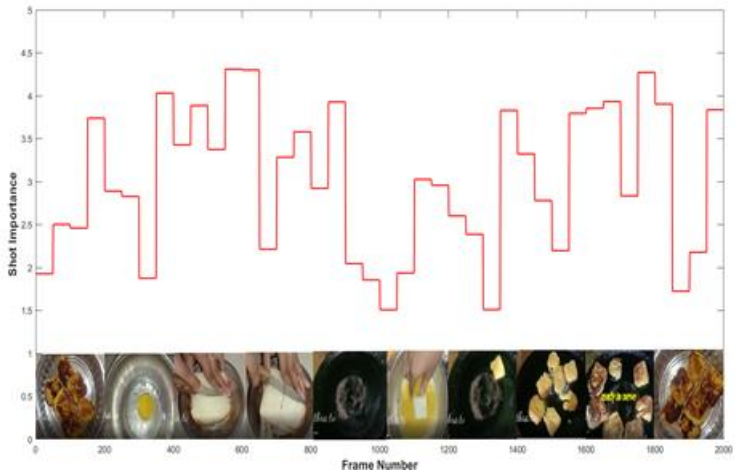


Figure 4.10: The Relation between Frame Number and Shot Importance Score for Any Chosen Video

The curve in Figure 4.10 represents the relationship between the frame number and the important shot. Each video contains a set of shots, and each shot contains a set of frames. These shots have a score of importance that is calculated

based on Equation 3.1. This importance varies from one video to another. The final important shot is the one that represents the highest point in the figure.

4.3.3 Implementation Results of Object Detection

As mentioned in chapter 3, the online deep learning model (Yolov5s) consists of many predefined models. It can predict the class type for the uploaded video and Search the YouTube channel for the advertisement video to bring the ad videos similar to the uploaded video as shown in Figures 4.11 and 4.12.

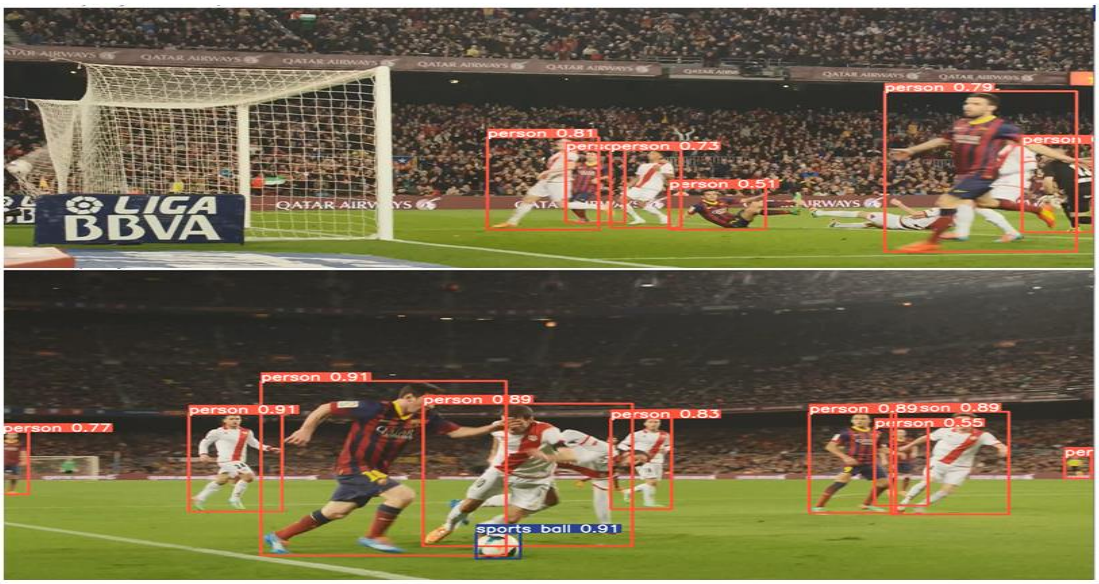


Figure 4.11: Object Detection of Sport Video

Figure 4.12: Object Detection of Cooking Video

4.3.4 Implementation Results of the API YouTube Channels and Similar Videos

After obtaining the type of the entire video based on video class detection, the YouTube channel is then searched for the ads video. The YouTube channel will return a list of similar ads videos and related metadata list including comments count, viewing count, likes count, and video ID, as shown in Figure 4.13.

List of the Returned (ad) Videos Similar to Sport Video
Video Id: c7WsbYDy4Hs Title: Original Unused Adidas Football Shoes in Pakistan View Count: 56863 Like Count: 542 Comment Count: 37 index:1 ---------- Video Id: yke7INSeDgI Title: Original Preloved Sneakers and Football shoes View Count: 890 Like Count: 9 Comment Count: 0 index: 2

Video Id: DtU12E-Pn3M

Title: It's About Time | Athletic Corner

View Count: 193

Like Count: 4

Comment Count: 0

index: 3

Video Id: DYKNSKJxXFE

Title: Top Football Shoes In Pakistan | Limited Stock | Athletic Corner

View Count: 210

Like Count: 3

Comment Count: 4

index: 4

Video Id: lEy8-0l7178

Title: Original Football shoes in Pakistan

View Count: 1474

Like Count: 18

Comment Count: 7

index: 5

Video Id: 8mJuKRxGuVI

Title: Puma NRGY Star Slip-On Mesh Running Shoes in Pakistan | Athletic Corner

View Count: 181

Like Count: 3

Comment Count: 0

index: 6

Video Id: YpkCf55R_G8

Title: Used Original Sneakers in Pakistan | Athleticorner

View Count: 267

Like Count: 2

Comment Count: 0

index: 7

Video Id: rTRvcqIVX90

Title: New adidas Football shoes and Sneakers now at a good price | 100% Original

View Count: 501

Like Count: 20

Comment Count: 0

index: 8

Video Id: Is_F5j6aA14

Title: New Balance Original Shoes in Lahore | Johar Town | Pakistan

View Count: 36

Like Count: 2

Comment Count: 0

index: 9

Video Id: QyX_eq_s6C8

Title: Puma Fierce for the Sneakerheads | The Casual Cruiser

View Count: 275

Like Count: 5

Comment Count: 0

index: 10

Video Id: qCAlZXD379s

Title: adidas Brand New Original Football Shoes

View Count: 191

Like Count: 9

Comment Count: 0

index: 11

Video Id: pxk24K6u0YM

Title: adidas Brand New Original Football Shoes Nemeziz

View Count: 267

Like Count: 14

Comment Count: 5

index: 12

Video Id: rgKWm4c4AYU

Title: Best Sneakers Collection

View Count: 113

Like Count: 4

Comment Count: 2

index: 13

Video Id: X_iGSOFskHE

Title: adidas Stan Smith Originals 'The Timeless sneakers' being packed in Pakistan

View Count: 570

Like Count: 4

Comment Count: 1

index: 14

Video Id: YvI4Ih8q8_A

Title: Original Sneakers in Pakistan

View Count: 147

Like Count: 3

Comment Count: 0

index: 15

Figure 4.13: List of the Returned (ad) Videos Similar to Sport Video

Next, the recommendation algorithms have been used to search for and find the best video ad similar to the original video. Figure 4.14 shows the results of returning the best ad video similar to the sport video using the GWO algorithm.

['c7WsbYDy4Hs', 'Original Unused adidas Football shoes in Pakistan', 1, 58202, 557, 41]
c7WsbYDy4Hs *//Video Id*
[28, 110, 160] *// length of each shot in video*
[0.13624417584004245, 1.0797430326186928, 1.0466766652256116] *// shot activity*
[6.724095693871937, 7.166246051013966, 7.169594495136871] *//entropy of each shot*
[0.5439551511238282, 0.9205739364351555, 0.992343926657974] // Important Shot score
Max Important Shot score is :
0.992343926657974
Important Shot number is:
2
advertisement starting frame is:
138

Figure 4.14: Returning the Best (ad)Video Using (GWO)

From the implementation results, the most suitable (ad) will be returned by (GWO). It has an index equal to 1 among all the returned YouTube channel advertisements list as in Figure 4.13. The important shot will be obtained in shot 2 and the advertisement video will be displayed in frame 138.

4.4 The Experimental Results

This section illustrates the experimental and analysis results of running the stages of the proposed system. A comparison is also made with some related works as shown in the following:

4.4.1 Implementation Time

The experimental results are shown in Table 4.5. Table 4.6 compares the extraction of feature speed with (SIFT) and (SURF) algorithms by [14] and [20].

Table 4.5: Running Effectiveness of the Proposed Approach

Detection of Shots		**Extraction of Key Frames**
Extraction of Features	**Matching Features**	1045.18 frames per sec.
43.25 frames per sec	30.62 frames per sec	

Table 4.6: Running Efficiency Comparison

Methods	**Frames**	**Feature Extraction Speed**
Proposed	1357	43.25 frames per sec
[20]	1351	41.20 frames per sec
[14]	1920	26.40 frames per sec

The findings show that our approach extracts ORB feature points at a rate of 43.25 frames per second, which is faster than [14] and [20]. As a result, the ORB-based algorithm outperforms SIFT and SURF-based algorithms in terms of speed.

4.4.2 Shot Detection Comparison

The evaluation parameters employed were recall and precision. The recall rate is determined by dividing the total number of video shots by the

proportion of true positive shots. The percentage of accurately identified shots is referred to as precision.

With regard to various types of movies, we assessed the recall and precision, and we made comparisons with the techniques in [14, 15, 16 and 20] to validate the proposed method's impact on shot detection as shown in Table 4.7.

Table 4.7: Comparing Shot Detection Techniques

Methods	**Collected Dataset**		**Methods**	**TVSum50**		**SumMe**	
proposed	Recall	Precision	**proposed**	Recall	Precision	Recall	Precision
	98.20%	95.30%		97.81%	90.87%	96.33%	93.20%
[14]	94.45%	95.79%					
[15]	96.40%	95.81%					
[16]	97%	95%					
[20]	97.20%	93.30%		95%	87%		

From Table 4.7 on the collected datasets TVSum50 and SumMe, we can see that the recall rate of our suggested method exhibits advantages. It makes use of the process for detecting and identifying local attributes in frame data, which can successfully detect the shots and offer a high recall rate. When the extracted images have a low precision index, the frames may appear repetitive. Yet when the recall rate is low, some video scenes that ought to have been recalled have been lost. Therefore, in terms of indicators for the video summarizing process, the recall rate is more significant. This makes our approach more useful, independent of effectiveness or recall rate.

4.4.3 Results of Video Class Specification and Shot Detection

To measure the efficacy of our suggested technique in determining the category of the video using the Yolo deep learning model (Yolov5s), as indicated in Tables 4.8 and 4.9, we examined many metrics, including (TS,FP,MS, recall and precision) using various types of videos and compared the outcomes with [13].

Table 4.8: Analysis of Video Class Specification Method and Shot Detection.

Sport			**cooking**			**transport**			**animals**		
TS	**FP**	**MS**	**TS**	**FP**	**MS**	**TS**	**FP**	**MS**	**TS**	**FP**	**MS**
32	0	36	6	0	0	6	0	0	11	0	4
27	0	41	4	0	2	4	2	0	13	0	2
35	0	33	4	0	2	2	0	4	8	0	7
30	0	38	18	12	0	6	0	0	4	0	11
40	0	28	6	0	0	5	0	1	12	0	3

According to the Yolo deep learning model (Yolov5s), the number of shots in videos of the four different classifications (sport, cooking, transportation, and animals) is shown in the table above, so that:

TS: The overall number of shots found.

FP: False Positive.

MS: Missing Shots.

Table 4.9: Comparison Analysis of our Method Performance Using Recall and Precision with Methods in [13]

method	**Sport**		**cooking**		**Transport**		**animals**	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Chi-Square	47.05882	100	100	100	100	100	73.33333	100
Intersection	39.70588	100	66.66667	100	66	66.66667	86.66667	100
ECR	51.47059	100	66.66667	100	33.33333	100	53.33333	100
Scaled Metrics	44.11765	100	100	60	100	100	40	100

proposed method	58.82353	100	100	100	83.33333	100	80	100

As mentioned above, recall is the ratio of real positive shots to all video shots. The rate of accurately identified shots is referred to as precision. Therefore, in comparison with the shot detection methods in [13], we can see that our proposed method is more accurate in shot detection and video class specification.

4.4.4 Experimental Results of The Recommendation and Video Advertising Algorithm

The evaluation measure (MSE and RMSE) for recommendation algorithm to suggest the appropriate recommender type of video advertisement for the actual type of viewing video is illustrated in Table 4.10.

Table 4.10: The Evaluation Measure for Recommendation Using (GWO) Algorithm

video no.	actual type	Recommended type	error	sqr error	mean sqrt error	root mean sqrt error
1	sporty	sporty	0.0	0.0	0.0	0.0
2	sporty	sporty	0.0	0.0		
3	sporty	sporty	0.0	0.0		
4	sporty	sporty	0.0	0.0		
5	sporty	sporty	0.0	0.0		
6	sporty	sporty	0.0	0.0		
7	trans	trans	0	0	0..67	0.818
8	trans	trans	0	0		
9	trans	sporty	1	1		
10	trans	trans	0	0		
11	trans	trans	0	0		
12	trans	trans	0	0		
13	cooking	cooking	0	0	0	0

14	cooking	cooking	0	0		
15	cooking	cooking	0	0		
16	cooking	cooking	0	0		
17	cooking	cooking	0	0		
18	cooking	cooking	0	0,67		
19	animal	animal	0	0	0	0
20	animal	animal	0	0		
21	animal	animal	0	0		
22	animal	animal	0	0		
23	animal	animal	0	0		
24	animal	animal	0	0		

As shown in table 4.10, we can see that all the recommended videos are accurately classified and suggested to the actual type of viewing videos. There is one error in recommended appropriate video ad for viewing video 9. On the other hand, when applying the same measures above to the comprehensive search algorithm, we have found that the error rate is completely zero, because it depends on searching and choosing between each ad and another in a sequential manner. So it is more appropriate with a few ad videos and takes more time complexity. This proves that the GWO results are realistic and more accurate, as it shown in Figure 4.11.

Table 4.11: The Evaluation Measure for Recommendation Using the comprehensive Algorithm

video no.	actual type	Recommended type	error	sqr error	mean sqrt error	root mean sqrt error
1	sporty	Sporty	0.0	0.0	0.0	0.0
2	sporty	Sporty	0.0	0.0		
3	sporty	Sporty	0.0	0.0		

4	sporty	Sporty	0.0	0.0		
5	sporty	Sporty	0.0	0.0		
6	sporty	Sporty	0.0	0.0		
7	trans	Trans	0	0	0..0	0.0
8	trans	Trans	0	0		
9	trans	Sporty	0	0		
10	trans	Trans	0	0		
11	trans	Trans	0	0		
12	trans	Trans	0	0		
13	cooking	Cooking	0	0	0	0
14	cooking	Cooking	0	0		
15	cooking	Cooking	0	0		
16	cooking	Cooking	0	0		
17	cooking	Cooking	0	0		
18	cooking	Cooking	0	0		
19	animal	Animal	0	0	0	0
20	animal	Animal	0	0		
21	animal	Animal	0	0		
22	animal	Animal	0	0		
23	animal	Animal	0	0		
24	animal	Animal	0	0		

Chapter Five

Conclusions and Future Work

5.1 Introduction

This chapter explains the proposed system's conclusions and outlines the main recommendations for future work. The first section presents the conclusions related to this dissertation. The second section suggests some future work related to this dissertation.

5.2 Conclusions

In this dissertation, a system is proposed that uses video advertising in order to show the appropriate video product advertisements to specific users at the proper time. There are several conclusions. They are as follows:

1. (ORB) and (FLANN) local feature descriptors and matching can provide a good representation of the frame image. Despite the fact that these features have some limitations, their combination gives additional robustness for feature extraction. In addition, combining (ORB) and (FLANN) with image enhancement improves feature extraction and matching in order to detect the shot or scene boundaries. The system results hit 95% in precision and 98% in recall.
2. The (ORB) based algorithm has a speed advantage over the (SIFT) and (SURF) based algorithm in terms of feature extraction and matching.
3. In order to reduce the size of shots while keeping the majority of the representative elements, a video summary technique has been used to represent the associations between users and original videos (viewing). It can provide a good representation of the entire video and detect the important shot by avoiding repetition. This process is very useful in important shot detection, so there is a need for reducing the size of shots while preserving most of the representative features.

4. Shot importance measure is calculated in order to ensure that the video summary only contains the most significant shots and to detect the appropriate time stamp to insert and display the (ad).
5. The real category of the original video can be detected using the Yolo deep learning model (Yolov5s) with YouTube search engine and the relation (similarity) between videos and products is found to obtain the content of the original video. The results of determining the category of the video by employing the Yolo deep learning model (Yolov5s) reach 100% in precision.
6. Recommendation methods with optimality algorithms (GWO and comprehensive) have been proposed. They depend on finding the best and most appropriate element to insert and display in the important scene for the original video.
7. The suggested system is able to display suitable items (ad) to specific users at the appropriate time.
8. The (GWO) results are realistic and more accurate in case the number of advertising videos is very large. The comprehensive recommendation algorithm, on the other hand, is more accurate with a few videos. Yet, it takes more time complexity because it will actually get one by one for each (ad) video and others. Therefore, it has no error rate in determining the appropriate ad according to the type of the original video.

5.3 Suggestions for Future Work

For further development of the current research, the following suggestions are important to implement in the future:

1- A combination of local and global feature extractor and descriptor methods can be used in order to increase the accuracy of frame image

representation and shot detection. These methods may include the histogram feature algorithm, Histogram of Gradients (HOG) and Expectation Maximization (EM). The combination of these features can avoid the weaknesses of each feature in isolation.

2- Another method, such as graph mining, can be utilized to give a good video summary and detect the most important shot.
3- Exploring online advertisements at the large scale and performing real-time updates for new objects will enable the framework as a whole to perform effectively in real time.
4- Designing and implementing a security method in order to guarantee the user's privacy.
5- A combination of a traditional recommendation algorithm with optimality can be used to select and suggest the best ad video for a particular user.

REFERENCES

REFERENCES

[1] Mohammad Hossein Bateni, Jon Feldman, Vahab Mirrokni, and Sam Chiu-wai Wong. "*Multiplicative bidding in online advertising*,"In Proceedings of the 15th ACM Conference on Economics and Computation (EC '14), pages 715–732, 2014.

[2] S. Sivapalan , Alireza Sadeghian and Hossein Rahanam," *Recommender Systems in E-Commerce*," Conference: World Automation CongressAt: Kona, HawaiiVolume: Proceedings, 2014.

[3] W. Kar, V. Swaminathan, and P. Albuquerque. "*Selection and ordering of linear online video ads*," In RecSys, pages 203–210, 2015.

[4] A. Veit et al., "*Learning visual clothing style with heterogeneous dyadic co-occurrences*," in Proc. Int. Conf. Comput. Vis., 2015, pp. 4642–4650.

[5] P. Zhou, Y. Zhou, D. Wu, and H. Jin, "*Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks*," IEEE Trans. Multimedia, vol. 18, no. 6, pp. 1217–1229, Jun. 2016.

[6] S. Angadi and V. Naik, "*Dynamic Summarization of Video Using Minimum Edge Weight Matching in Bipartite Graphs*," 139 International Journal of Image, Graphics and Signal Processing, vol. 8, p. 9, 2016.

[7] D.-j. Jeong, et al., "*A static video summarization method based on the sparse coding of features and representativeness of frames*," EURASIP Journal on Image and Video Processing, 2016.

[8] Mingyang Ma, Shaohui Mei1, Junhui Hou, Shuai Wan, and Zhiyong Wang, "*Video Summarization via Temporal Collaborative Representation of Adjacent Frames*", International Symposium on Intelligent Signal Processing and Communication Systems, Xiamen, China, November 6-9, 2017.

[9] S. Sarmadi, "*New Approach in Video Summarization Based on Color Feature*," Bulletin de la Société Royale des Sciences de Liège, vol. 86, pp. 535-542, 2017.

[10] Mohaiminul Al Nahian , A. S. M. Iftekhar , Mohammad Tariqul Islam S. M. Mahbubur Rahman† , and Dimitrios Hatzinakos," *CNN-Based Prediction of Frame-Level Shot*

Importance for Video Summarization ," International Conference on New Trends in Computing Sciences (ICTCS),‡ IEEE Xplore, 2017.

[11] D. Das, L. Sahoo and S. Datta," *A Survey on Recommendation System* , International Journal of Computer Applications (0975 – 8887) Volume 160 – No 7, February 2017.

[12] S. Sidana," *Recommendation systems for online advertising*," Computers and Society [cs.CY]. Université Grenoble Alpes, 2018.

[13] Priyanka S , Dr.Jharna Majumdar , and Santhosh Kumar K L," *Video Shot Detection on Embedded System*," International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015.

[14] Li, F.; Zhao, Y.; Wang, S.G.; Chen, H.-X, "*Video scene mutation change detection combined with SIFT algorithm*," Chin. Opt. 2016, 9, 74–80.

[15] Hannane, R.; Elboushaki, A.; Afdel, K.; Naghabhushan, P.; Javed, M ,"An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram," Int. J. Multimed. Inf. Retr. 2016, 5, 89–104.

[16] S. P. Algur, "Video key frame extraction using entropy value as global and local feature," arXiv preprint arXiv:1605.08857, 2016.

[17] H.Z.N. a.I.Hadi, a." *Robust Video Shot Importance Measurement Based on SIFT and Optical Flow*," International Journal of Pure and Applied Mathematics, Volume 119 No. 15, 1015- 1023.2018.

[18] Haosong Yue, Yue Yu, Xingming Wu, and Weihai Chen," *Keyframe extraction and loop closure detection considering robot motion*," 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018.

[19] X. Li , B. Zhao, and X. Lu," *Key Frame Extraction in the Summary Space*," IEEE Transactions on Cybernetics,Vo.48, Issue. 6, June 2018.

[20] Liang, B.; Li, N.; He, Z.; Wang, Z.; Fu, Y.; Lu, T ," *News Video Summarization Combining SURF and Color Histogram Features*," Entropy 2021, 23, 982. https://doi.org/10.3390/e23080982.

[21] Z. Wang et al., "*Joint social and content recommendation for user generated videos in online social network*," IEEE Trans. Multimedia, Apr, vol. 15, no. 3, pp. 698–709, 2013.

[22] K. Yadati, H. Katti, and M. Kankanhalli, " *CAVVA:Computational affective video-in-video advertising* ," IEEE Trans. Multimedia, vol. 16, no. 1, pp. 15–23 , Jan 2014 .

[23] P. Cui, Zhiyu Wang and Zhou Su," *What Videos Are Similar with You? Learning a Common Attributed Representation for Video Recommendation*," ACM, 2014.

[24] Y. Deldjoo , M. Elahi1 , and P. Cremonesi," *Content-Based Video Recommendation System Based on Stylistic Visual Features*," International Journal on Data Semantics Springer, 5:99–113, 2016 .

[25] Z.-Q. Cheng, Y. Liu, X. Wu, and X.-S. Hua, "*Video ecommerce: Towards online video advertising*," in Proc. ACM Int. Conf. Multimedia, 2016, pp. 1365–1374.

[26] Z.-Q. Cheng, Y. Liu, X. Wu, and X.-S. Hua, "*Video ecommerce++: Towards online video advertising*," in IEEE Transactions on Multimedia , June 2017, vol. 19, No. 6.

[27] Z. Hussain, M. Zhang , and X. Zhang," *Automatic Understanding of Image and Video Advertisements*," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[28] Gunavathie. M.Aa, ,and K. Baavi.P," *Video e-commerce: Large Scale Online Video Advertising based on user Preference*," I.J. Education and Management Engineering, September 2018.

[29] James Hahn and Dr. Adriana Kovashka," *Measuring Effectiveness of Video Advertisements*," DBLP - CS Bibliography, Jan 2019.

[30] S. Zhang , Z. Tan , and Jin Yu," *Poet: Product-oriented Video Captioner for E-commerce*," on Multimedia (MM '20), October 12–16, Seattle, WA, USA. ACM, NewYork, NY, USA, 2020 .

[31] Jun Ikeda, Hiroyuki Seshime, Xueting Wang, and Toshihiko Yamasaki ," *Predicting Online Video Advertising Effects with Multimodal Deep Learning*," International Conference on Pattern Recognition 2020 (ICPR).

[32] Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang ,Xiwang Yang , Xiaobing Liu , Jiliang Tang , and Hui Liu ," *DEAR: Deep Reinforcement Learning for Online Advertising Impression in Recommender Systems* ," The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), 2021.

[33] Y. Li and C. J. Kuo, "*Video content analysis using multimodal information: For movie content extraction*," indexing and representation: Springer Science & Business Media, 2013.

[34] C. V. Sheena and N. Narayanan, "*Key-frame extraction by analysis of histograms of video frames using statistical methods*," Procedia Computer Science, vol. 70, pp. 36-40, 2015.

[35] Gao, G. and Liu, C.H.,"*Video Cataloguing: Structure Parsing and Content Extraction* ,"Taylor & Francis Group, LLC, 2016.

[36] S. Tippaya, et al., "*Multi-modal visual features-based video shot boundary detection*," IEEE Access, vol. 5, pp. 12563-12575, 2017.

[37] Y. Bendraou, "*Video shot boundary detection and key-frame extraction using mathematical models*," 2017.

[38] S. Abdulhussain, et al., "*Methods and challenges in shot boundary detection: a review*," Entropy, vol. 20, p. 214, 2018.

[39] Rashmi, B.S., Nagendraswamy, H.S.," *Effective video shot boundary detection and keyframe selection using soft computing techniques*," Int. J. Comput. Vis. Image Process. 8, 27–48, 2018.

[40] Liang, J.-S.; Wen, H.-P,"*Key frame abstraction and retrieval of videos based on deep learning*," Control Eng. China 2019, 26, 965–970.

[41] Gharbi, H., Bahroun, S., Zagrouba, E , " *Key frame extraction for video summarization using local description and repeatability graph clustering* ," Signal, Image Video Process. 13, 507–515, 2019.

[42] Kanagaraj, K., Priya, G.G.L, " *Curvelet transform based feature extraction and selection for multimedia event classification*," J. King Saud Univ. - Comput Inf. Sci. 0–8, 2018.

[43] Yao, H.F.; Guo, B.L ,"*An ORB-based feature matching algorithm*," Electron. Des. Eng. 2019, 27, 175–179.

[44] Kan, S.; Cen, Y.; He, Z.; Zhang, Z.; Zhang, L.; Wang, Y ," *Supervised Deep Feature Embedding with Hand Crafted Feature* ," IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 2019, 28, 5809–5823.

[45] Li, S.-S.; Feng, H.-C.; Su, X.-H ," *Improved algorithm for shot mutation detection based on SIFT feature points*," J. Wuhan Polytech. Univ. 2019, 38, 67–72.

[46] Zhou, S.; Wu, X.; Qi, Y.; Luo, S.; Xie, X," *Video shot boundary detection based on multi-level features collaboration*," J. VLSI Signal Process. Syst. Signal Image Video Process. 2020, 15, 627–635.

[47] Wang, R.-J.; Niu, Z.-X.; Song, C.-H.; Niu, B.-N ,"*An improved shot boundary detection algorithm based on mutual information*," Sci. Technol. Eng. 2018, 18, 228–236.

[48] Zhou, S.; Wu, X.; Qi, Y.; Luo, S.; Xie, X ," *Video shot boundary detection based on multi-level features collaboration* ," J. VLSI Signal Process. Syst. Signal Image Video Process. 2020, 15, 627–635.

[49] I. H. Ali and T. T. Al-Fatlawi, " *Key Frame Extraction Methods*," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 10, pp. 485–490, 2018.

[50] Tang, H.; Xiao, W.; Liu, H.; Sebe, N ,"*Fast and Robust Dynamic Hand Gesture Recognition via Key Frames Extraction and Feature Fusion*," Neurocomputing 2018, 331, 424–433.

[51] Huang, C.; Wang, H," *Novel Key-frames Selection Framework for Comprehensive Video Summarization*," IEEE Trans. Circuits Syst. Video Technol. 2019, 30, 577–589.

[52] Yu, L.; Cao, J.; Chen, M.; Cui, X ," *Key frame extraction scheme based on sliding window and features* ," Peer-Netw. Appl. 2018, 11, 1141–1152.

[53] Rubleee, Rabaudv and Konoligek, " *ORB: an efficient alternative to SIFT or SURF* ", Proc of IEEE International Conference on Computer Vision, (2011), pp. 2564-2571.

[54] Xie, Y.; Wang, Q.; Chang, Y.; Zhang, X ,"*Fast Target Recognition Based on Improved ORB Feature*," Appl. Sci. 2022, 12, 786.

[55] Y. D. Feng and Y. sun ,"*Image. Matching Algorithm Based on SURF Feature Extraction and FLANN Search* ," Journal of Graphics, 2015 (4), pp. 650.

[56] Shigang Wang , Zhenjin Guo and Yang Liu ," An Image Matching Method Based on SIFT Feature Extraction and FLANN Search Algorithm Improvement ," Journal of Physics: Conference Series, ICAIIT 2021.

[57] *Jadon, S., & Jasim, M. (2019) ,"Video Summarization using Keyframe Extraction and Video Skimming ,"*arXiv preprint arXiv:1910.04792.

[58] Muhammad Ajmal, Muhammad Husnain Ashraf, Muhammad Shakir, Yasir Abbas, and Faiz Ali Shah ," *Video Summarization: Techniques and Classification* ," ICCVG 2012, LNCS 7594, pp. 1–13, 2012.

[59] K.-L. Du and M. N. Swamy, "*Neural networks and statistical learning: Springer Science & Business Media*, " 2013.

[60] Y. Guo, et al., "*Selecting video key frames based on relative entropy and the extreme studentized deviate test*," Entropy, vol. 18, p. 73, 2016.

[61] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie ,"*Feature pyramid networks for object detection*," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2117–2125, 2017.

[62] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar ,"*Focal loss for dense object detection*," arXiv preprint arXiv:1708.02002, 2017.

[63] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016).,"*You Only Look Once: Unified, RealTime Object Detection* ," arXiv. Seach date 12.11.2020. https://arxiv.org/pdf/1506.0264.

[64] Do Thuan ," *Evolution of YOLO Algorithm and YOLOv5: The State-of-the-art Object Detection Algorithm* ," Oulu University of Applied Sciences, Spring 2021.

[65] Y. Zhang, M. Zhang, Y. Zhang, G. Lai, Y. Liu, H. Zhang, and S. Ma ,"*Daily-aware personalized recommendation based on feature-level time series analysis* ,"In WWW, 2015, pages 1373–1383.

[66] H. Ma, " *On measuring social friend interest similarities in recommender systems* ," In SIGIR, 2014, pages 465–474.

[67] J. J. McAuley, R. Pandey, and J. Leskovec, " *Inferring networks of substitutable and complementary products* ," In KDD, 2015, pages 785–794.

[68] T. Zhu, P. Harrington, J. Li, and L. Tang, " *Bundle recommendation in ecommerce* ," In SIGIR, 2014, pages 657–666.

[69] Ali Elkahky, Yang Song, Xiaodong He, "*A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems*," the International World Wide Web Conference Committee (IW3C2). ACM, 2015.

[70] P. Jiang, Y. Zhu, Y. Zhang, and Q. Yuan," *Life-stage prediction for product recommendation in e-commerce* ," In KDD, pages 1879– 1888, 2015.

[71] O. Sar Shalom, N. Koenigstein, U. Paquet, and H. P. Vanchinathan, " *Beyond collaborative filtering: The list recommendation problem* ," in Proc. Int. World Wide Web Conf , pp. 63–72, 2016.

[72] Liu, Hongyan, et al, "*Combining user preferences and user opinions for accurate recommendation*," Electronic Commerce Research and Applications 12.1 , 14-23, 2013.

[73] E. Cano and M. Morisio," *Hybrid Recommender Systems: A Systematic Literature Review*," : Intelligent Data Analysis, vol. 21 n. 6, pp. 1487-1524. - ISSN 1088-467X, 2017.

[74] I. Ronen, I. Guy, E. Kravi, and M. Barnea ,"*Recommending social media content to community owners*,". In SIGIR, pages 243–252, 2014 .

[75] GK, Kiran Kumar Bathi, and N. Sheela ,"*Personalized Recommender System* ," PARIPEX-Indian Journal of Research 4.8 , 2016.

[76] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun, " *Xdeepfm: Combining explicit and implicit feature interactions for recommender systems*," in the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1754–1763.

[77] K. Kolomvatsos, M. Koziri and T. Loukopoulos," *A recommendation system for allocating video resources in multiple partitions, In Big Data Recommender Systems: Recent Trends and Advances* ," Enlighten: Institute of Engineering and Technology (IET), 2018.

[78] L. Viktoratos, A. Tsadiras and N. Bassiliades, " *Combining Community-Based Knowledge with Association Rule Mining to Alleviate the Cold Start Problem in Context-Aware Recommender Systems,* " Expert Systems with Applications, 101 (2018) 78-90.

[79] J. Chen, C.Z. Uliji and L. Chen," *Collaborative ltering recommendation algorithm based on user correlation and evolutionary clustering*," Complex & Intelligent Systems, (2019) 1-10.

[80] Liu, H.; Zhao, X.; Wang, C.; Liu, X.; and Tang, J ," *Automated Embedding Size Search in Deep Recommender Systems* ," In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2307–2316, 2020.

[81] L. C. Bezerra, M. Lopez-Ib ´ anez, and T. St ´ utzle, "*Automatic ¨ component-wise design of multiobjective evolutionary algorithms*," IEEE Transactions on Evolutionary Computation, vol. 20, no. 3, pp. 403–417, 2016.

[82] M. Zhao, H. Ge, Y. Lian, and C. L. P. Chen, " *Online decisioning meta-heuristic framework for large scale black-box optimization*," arXiv preprint arXiv:1812.06585v1, 2018.

[83] P. Kerschke, H. H. Hoos, F. Neumann, and H. Trautmann, "*Automated algorithm selection: Survey and perspectives*," Evolutionary Computation, vol. 27, no. 1, pp. 3–45, 2018.

[84] X. Chu, F. Cai, C. Cui, M. Hu, L. Li, and Q. Qin, "*Adaptive recommendation model using meta-learning for population-based algorithms*," Information Sciences, vol. 476, pp. 192–210, 2019.

[85] H. Zhu, X. Li, P. Zhang, G. Li, J. He, H. Li, and K. Gai ,"*Learning tree-based deep model for recommender systems* ," In KDD, pages 1079–1088, 2018.

[86] N. Lynn and P. N. Suganthan, ‘‘*Heterogeneous comprehensive learning particle swarm optimization with enhanced exploration and exploitation*,’’ Swarm Evol. Comput., vol. 24, pp. 11–24, Oct. 2015.

[87] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, ‘‘*Long short term memory hyperparameter optimization for a neural network based emotion recognition framework*,’’ IEEE Access, vol. 6, pp. 49325–49338, 2018.

[88] T. Kim and S. Cho, ‘‘*Particle swarm optimization-based CNN-LSTM networks for forecasting energy consumption*,’’ in Proc. Congr. Evol. Comput. (CEC), 2019, pp. 1510–1516.

[89] N. Xue, I. Triguero, G. P. Figueredo, and D. Landa-Silva, ‘‘*Evolving deep CNN-LSTMs for inventory time series prediction*,’’ in Proc. Congr. Evol. Comput. (CEC), 2019, pp. 1517–1524.

[90] Seyedali Mirjalili et al, *" Grey Wolf Optimizer* ," Advances in Engineering Software Volume 69, March 2014, Pages 46-61.

[91] Jie-ShengWang , Shu-Xia Li ," *An Improved Grey Wolf Optimizer Based on Diferential Evolution and Elimination Mechanism* ," Scientific Reports, 2019.

[92] P. Niu, S. Niu, N. Liu, and L. Chang, ‘‘*The defect of the Grey wolf optimization algorithm and its verification method*,’’ Knowl.-Based Syst., vol. 171, pp. 37–43, May 2019.

[93] Aidean Sharghi, Jacob S. Laurel, and Boqing Gong “*Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach*”, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[94] W. Bruce Croft, Donald Metzler , and Trevor Strohman ,"*Search Engines Information Retrieval in Practice,*" Pearson Education, Inc,2015.

[95] Afzal Godil et al ," *Performance Metrics for Evaluating Object and Human Detection and Tracking Systems* ," NISTIR 7972,2014.

[96] R. Vio et al ," *The correct estimate of the probability of false detection of the matched filter in weak-signal detection problems* ," A&A 604, A115 (2017).

[97] Itiel Dror ," *The Error in 'Error Rate': Why Error Rates Are So Needed, Yet So Elusive* ,"Journal of Forensic Sciences 65(5),2020.

[98] Gunnar Schr¨oder, Maik Thiele, and Wolfgang Lehner ,"*Setting goals and choosing metrics for recommender system evaluations*," In CEUR Workshop Proceedings, volume 811, pages 78–85, 2011.

[99] P. Pu, L. Chen, R. Hu, "*A user-centric evaluation framework for recommender systems*," in: Proc. Fifth ACM Conf. Recomm. Syst.—RecSys'11, 2011, p.157.

[100] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020 ,"*Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance*," Proc. ACM Hum.-Comput. Interact. 4, CSCW2, Article 153 (October 2020), 22 pages.

[101] Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A ,"*TVSum: Summarizing web videos using titles* ," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5179–5187.

[101] Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A ,"*TVSum: Summarizing web videos using titles* ," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5179–5187.

[102] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "*Creating summaries from user videos*," in ECCV, 2014.

[103] TV, 2004. TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

[104] http://research.google.com/youtube8m/download.html/.

[105] http://www.metacafe.com/.

[106] https://www.crcv.ucf.edu/research/data-sets/ucf101/

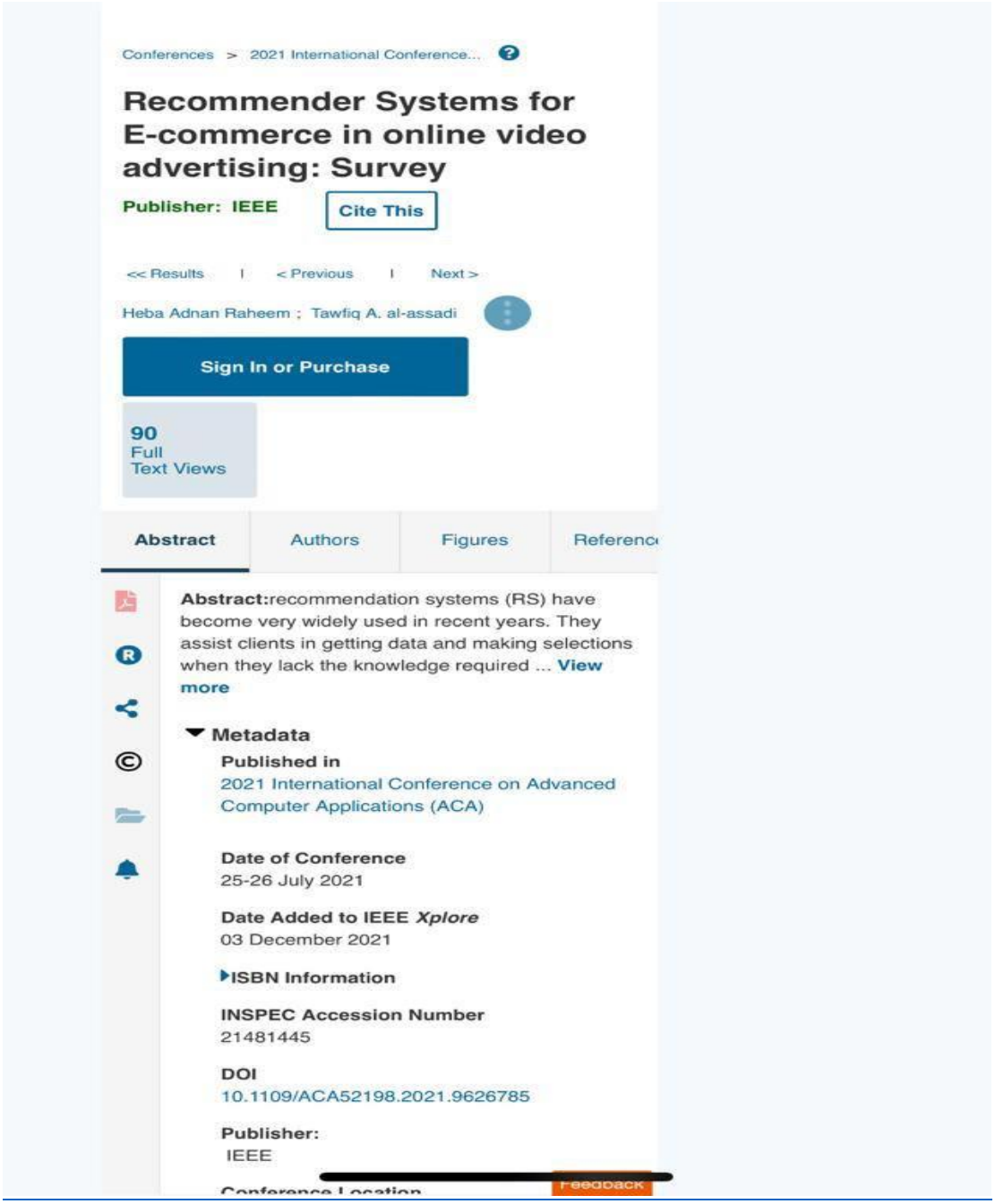
[107] "The Open Video Project ", 2018. [Online].Available: https://open-video.org/

[108] Google Video. http://video.google.com/

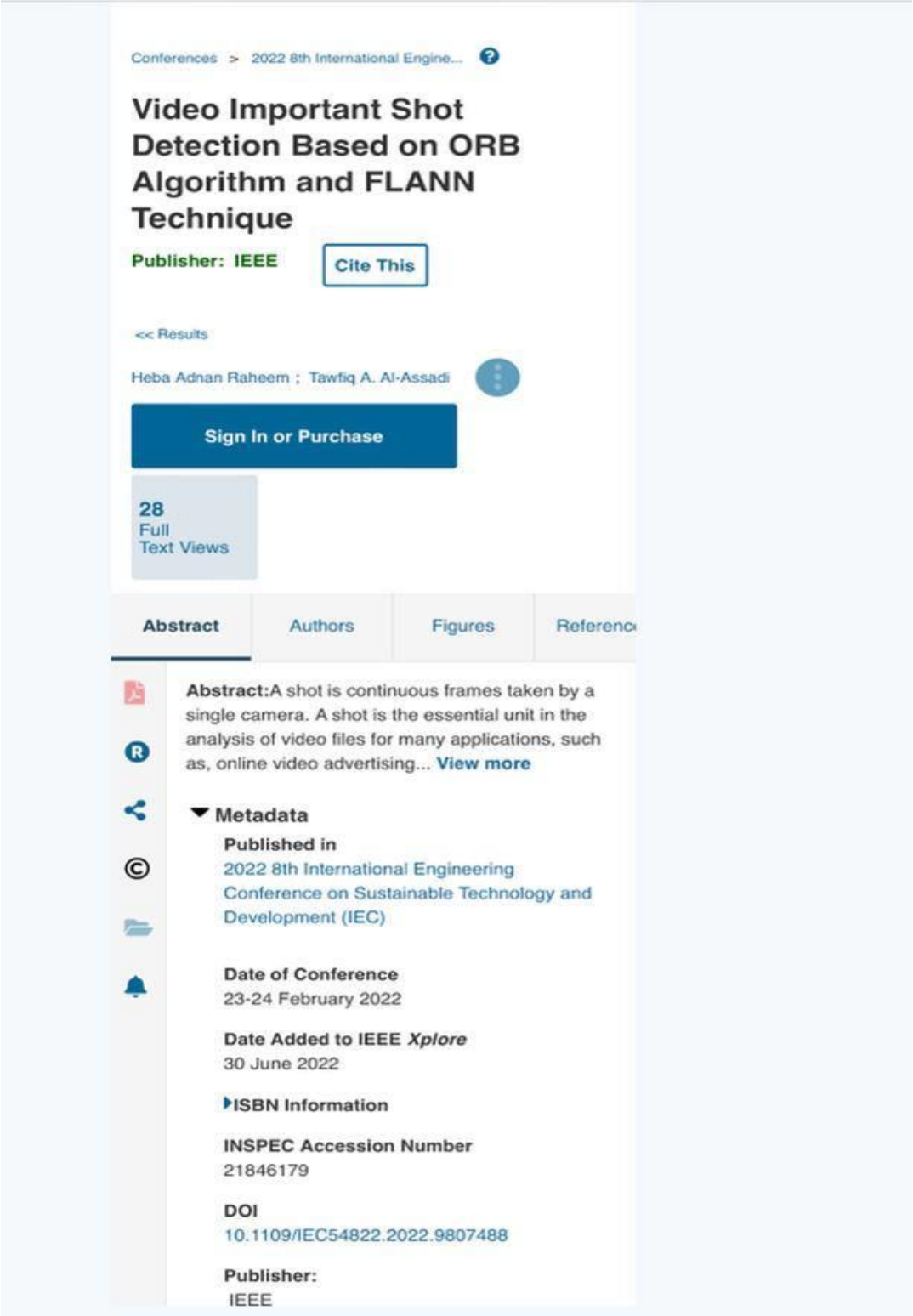
APPENDICES

APPENDIX_1

https://ieeexplore.ieee.org/document/9626785/



https://ieeexplore.ieee.org/document/9807488/




2022
AEST
2nd International Conference on Advances in Engineering Science and Technology
Certificate of Acceptance
This certificate is granted to
"Hba Adnan Raheem, Ali; Tawfiq Al-Assadi"
Certifies the acceptance of research paper entitled:
(Video Shot Importance Detection Based on Video Summarization with Feature Extrac)
for presentation in AEST-2022
Which will be held on 24-25 October, 2022 at University of Babylon, College of Engineering, Babylon, Iraq.
Sponsored by IEEE Iraq-section. The final decision of publication in IEEE Xplore is subject to terms and conditions of Conference Scientific Committee and IEEE.
Prof. Dr. Hatem Hadi Al-Taei
Dean of the College of Engineering
Prof. Dr. Sattar B. Sadkhan
IEEE Iraq Section
IEEE Iraq Section
IEEE Xplore
Scopus

الخلاصة

الإعلان هو إشعار أو عرض في وسيلة عامة للترويج لسلعة. تلعب الإعلانات دورًا مهمًا في إدخال منتجات جديدة إلى السوق وحث الناس على شرائها. كما أنها توفر لشركات التجارة الإلكترونية فرصة للتوصية بمنتجاتها في مقاطع الفيديو.

تهدف هذه الأطروحة إلى تنفيذ نظام إعلان فيدوي فعال لعرض فيديو مناسب لإعلانات المنتج لمستخدم معين في توقيتات مناسبة باستخدام طرق التوصية. هذا يأخذ في الحسبان محتوى الفيديو (الدلالات أو العلاقة بين الفيديو والمنتجات)، وكذلك سلوك المشاهدة الحالي (العلاقة بين المستخدم والفيديو).

يتألف النظام المقترح من المراحل الرئيسية التالية. في المرحلة الاولى يتم تحليل بنية الفيديو إلى مكوناتها الأساسية. في المرحلة الثانية يتم الكشف عن العلاقة بين المستخدمين ومقاطع الفيديو عن طريق استخلاص الخصائص باستخدام خوارزمية ORB ومطابقتها باستخدام تقنية FLANN ومن ثم تطبيق تقنية تلخيص الفيديو الأصلي لتحديد اللقطة المهمة فيه. يعتمد التلخيص على تقنية تجميع محسّنة تستخدم حساب مقياس entropy والذي يعطينا خلاصة فيديو تشتمل على استخلاص الإطار الرئيسي من اللقطة والحصول على اللقطة الأكثر أهمية مع إطار رئيسي تمثيلي لها. في المرحلة الثالثة يتم الكشف عن وجه التشابه بين الفيديو الأصلي والمنتج بالاعتماد على نموذج التعلم العميق (Yolo) مع محرك بحث يوتيوب. في المرحلة الرابعة تم اقتراح خوارزميات التوصية المستندة على نموذج مُحسّن لدمج ارتباط المنتج والإطار الرئيسي للفيديو مع أهم لقطة للفيديو الأصلي وذلك للتوصية بإعلان فيدوي مناسب للمستخدم. اقترحت خوارزمية (GWO) وخوارزمية البحث الشامل (comprehensive) للبحث والعثور على أفضل إعلان فيدوي مشابه للفيديو الأصلي والتوصية به للمستخدم.

تبين النتائج التجريبية أن النظام المقترح يعطي نتائج جيدة في كفاءة التشغيل وتحديد اللقطة المهمة واستخلاص الإطار الرئيسي وتصنيف فئة الفيديو أو التعرف على محتوى الفيديو، والتوصية في الإعلان الفيديوي. مقارنة بالدراسات الأخرى تظهر النتائج أن الخوارزمية القائمة على ORB لها ميزة السرعة لاستخلاص الصفات.

وصلت نتائج النظام إلى ٩٥٪ للدقة و٩٨٪ للاسترجاع. هذا يجعل طريقتنا تكتشف اللقطات بنجاح وتوفر معدل استدعاء مرتفع. أعطت نتائج تحديد فئة الفيديو باستخدام نموذج Yolo للتعلم العميق (Yolov5s) دقة ١٠٠٪. وبالتالي فإن الطريقة المقترحة أكثر دقة للحصول على محتوى الفيديو. تُظهر نتائج التوصية وخوارزمية الإعلان الفيديوي أن جميع مقاطع الفيديو الموصي بها مصنفة بدقة واقترحت على النوع الفعلي لمشاهدة مقاطع الفيديو في خوارزمية (GWO) وخوارزمية (comprehensive).

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية تكنولوجيا المعلومات
قسم البرامجيات

اقتراح نظام توصية لعرض إعلانات المنتج في طوابع زمنية مناسبة للفيديو

أطروحة

مقدمة إلى مجلس كلية تكنولوجيا المعلومات / جامعة بابل والتي هي جزء من متطلبات الحصول على درجة الدكتوراه في تكنولوجيا المعلومات – البرامجيات

من قبل
هبة عدنان رحيم علي

بأشراف
أ.د. توفيق عبد الخالق الأسدي

١٤٤٤ هـ ٢٠٢٣ م