TRANSMIXR

IGNITE THE IMMERSIVE MEDIA SECTOR BY ENABLING NEW NARRATIVE VISIONS

D2.1: Initial media ingestion, understanding and summarisation components

Lead author/organisation: Lyndon Nixon (MOD)

Co-authors/organisations: Ioannis Kontostathis, Maria Tzelepi, Vasileios Mezaris (CERTH), John Dingliana, Nivesh Gadipudi (TCD)

<table>
<tr><th colspan="2">Versioning (by)</th><th>Date</th><th>Version</th></tr>
<tr><td colspan="2"><b>Lyndon Nixon | Created initial template</b></td><td>17.11.23</td><td>0.1</td></tr>
<tr><td><b>Lyndon Nixon |</b></td><td><b>Introduction</b></td><td>29.11.23</td><td>0.2</td></tr>
<tr><td><b>John Dingliana</b></td><td>Added TCD inputs to 3.4 &amp; 4.3</td><td>01.12.23</td><td>0.3</td></tr>
</table>

Nivesh Gadipudi			
Vasileios Mezaris **Ioannis Kontostathis** **Maria Tzelepi**	Added CERTH inputs	06.12.23	0.4
Lyndon Nixon	Added MOD inputs	13.12.23	0.5
Lyndon Nixon	Included text summarization and conclusion, sent for review	15.12.23	0.7
Lyndon Nixon **John Dingliana** **Ioannis Kontostathis** **Maria Tzelepi**	Edits following reviewer comments	21.12.23	0.9
Lyndon Nixon	Submission version	22.12.23	1.0

Table of Contents

1. Introduction

This deliverable introduces the first prototypes of the media ingestion, understanding and summarisation components of TRANSMIXR. These components together form the CONTENT UNDERSTANDING part of the TRANSMIXR project (see Fig.1 below), enabling users to identify relevant content items, extract metadata about them, store that metadata for query and retrieval purposes, as well as retrieve suitably adapted versions of those content items such as summarised text or video which can be inserted into immersive scenes. These components make the aforementioned functionalities available to the TRANSMIXR use case pilots, through the CONTENT CREATION and CONTENT DELIVERY workflows (WPs 3 and 4), where professional users can effectively discover relevant content items and adapt them for use in the XR environment.

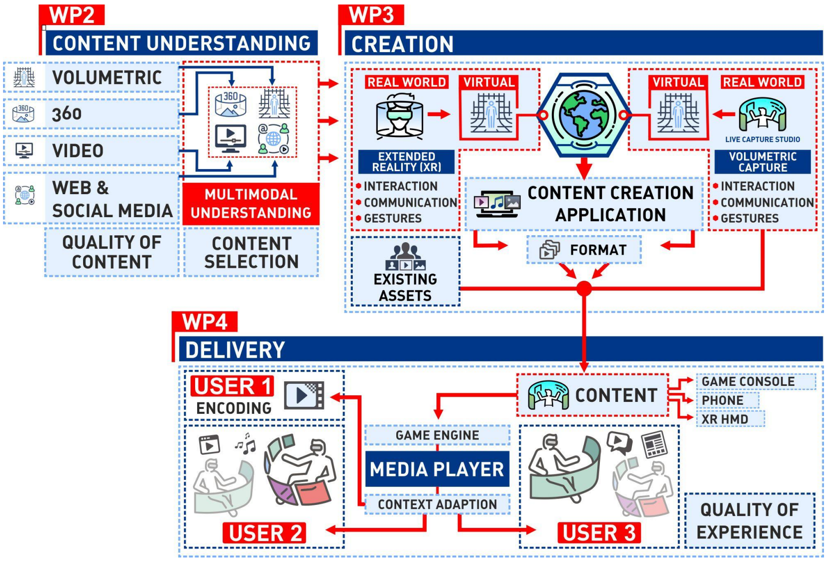


Figure 1: Place of the work of Work Package 2 in the TRANSMIXR project overview.

Component Overview

The developed components address content collection, metadata extraction and media retrieval for the following content types: Web & social media (text-based), classical video, 360 video and volumetric video.

Content collection tasks involve discovery of relevant content items on the public Web, social networks, media archives and company repositories.

Metadata extraction tasks involve the application of appropriate content analysis algorithms on the collected content items in order to extract technical and

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

descriptive metadata about the item and store it in a machine readable and shareable format.

Media retrieval tasks involve the access to the content metadata in a repository for query purposes as well as access to the content itself for adaptation purposes, i.e. requesting a video file but receiving it in a modified form, e.g. a summarisation, cropped or converted to image(s) for the purpose of re-use in an immersive environment.

Table 1 provides an overview of the components covered in this deliverable. The remaining Chapters 2-4 detail the current functional status of each component individually. Chapter 5 concludes with an outlook for further development of all components related to content understanding until the end of the TRANSMIXR project.

Table 1:Overview of all components.

Purpose	Content type	Component (lead development partner)
Content Collection	Web & social media	Content mirror for different online sources (MODUL)
	Classical video	As above (with relevant video assets stored temporarily at CERTH)
	360 video	As above (with relevant video assets stored temporarily at CERTH)
	Volumetric video	Manual directory (TCD)
Metadata Extraction	Web & social media	Data processing pipeline with NLP, NER and NEL (MODUL)
	Classical video	Video analysis service (CERTH)
	360 video	Video analysis service (CERTH)
	Volumetric vídeo	Semantic annotation service (TCD)
Media Retrieval	Web & social media	Document repository with Web dashboard and Search API (WLT) Entity search and description tool (MOD) Text summarization service (MOD/WLT)
	Classical video	Video summarization service (CERTH)
	360 video	Video summarization service (CERTH)
	Volumetric vídeo	Semantic summarization service (TCD)

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

Role in the TRANSMIXR Architecture

The above components may be used as part of a workflow (see Fig. 2) which focuses on the creation of interactive and immersive experiences for both professional users as well as the public, whenever external content (text, image, video…) needs to be identified and used. This may be for any one of the following three purposes:

(1) Use of metadata analytics to identify topics of interest. For example, keyword analysis across news articles enables a *story graph* which shows which news stories are in the focus of online reporting over time. This will be further explored in D2.2.

(2) Direct retrieval of content to illustrate a topic. Both a Web based dashboard and an API are available to query for content items via their metadata. Content items returned as a result of a search query are identified by a reference to their retrievable location (such as an URL if available on the Web). It is part of the content creation tools of WP3 to be able to handle any reference that is acquired to retrieve the content itself.

(3) Use of content items to generate new content for an immersive scene. Alternatively, the content which is found by the creator may be used as the input into a further content adaptation or generation step. For example, text or videos may be summarised to focus on their most salient parts for delivery in an immersive experience.

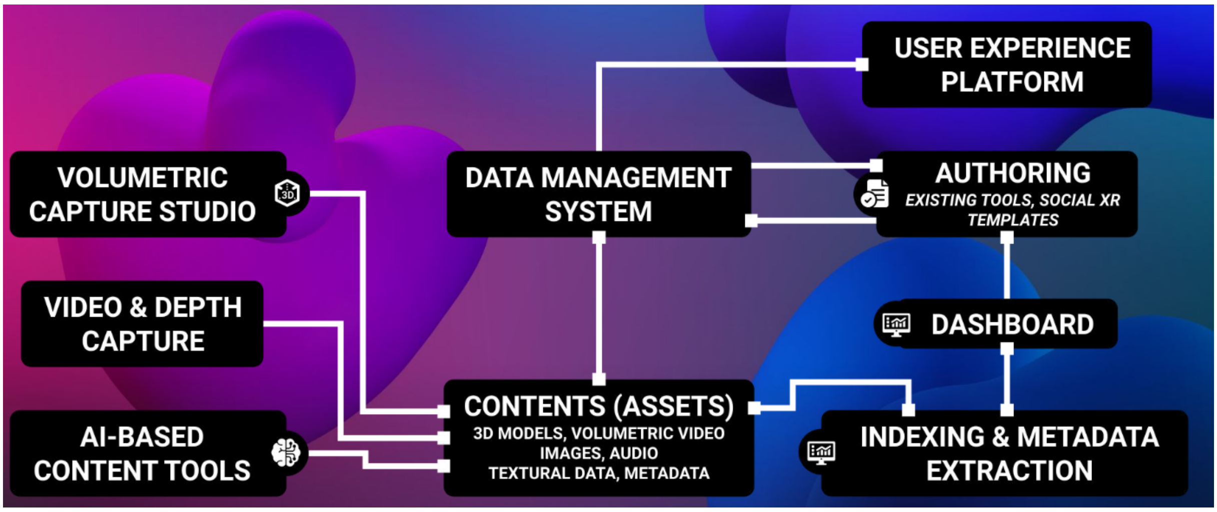


Figure 2. Place of this work in the TRANSMIXR architecture

Once content items or their metadata (CONTENTS (ASSETS)) have been found and selected (WP2: INDEXING & METADATA EXTRACTION), it is the task of the user to manage the assets in a CMS (DATA MANAGEMENT SYSTEM) and proceed to the content creation step (WP3: AUTHORING). Finally, they can deliver the immersive scene with the authored content to the end device (WP4: USER EXPERIENCE PLATFORM).

2. Content Collection Components

The purpose of the content collection components is to identify relevant and useful content items from larger and broader content collections, which may then be passed to the metadata extraction components. Our focus is on open and Web-based collections, but the same approach may be applied to specific (also closed and organisation-internal) collections on a case-by-case basis in the project.

2.1. Webpage and Social Media Ingestion

Web and social media (textual) content is extracted using components built by MOD known as "*content mirrors*". They are specific to a social network API or a Web crawl of a predefined set of Webpages. The purpose of collecting online textual sources, analysing and annotating them, is threefold:

1. Extracting keywords from news articles and a time-based clustering of the articles by keyword as the basis for our **story detection**, which in turn may be used by journalists to track which news stories emerge and persist over time within certain news communities. This is part of the content metrics which will be presented in D2.2.
2. Supporting text search as well as visual analytics over the collected documents for use by professional users for discovery of online content, e.g. journalists in the news pilot can explore the different reporting of the same news story by source; curators in the cultural heritage pilot can identify audience interest in certain topics from social media analysis and incorporate those topics in their immersive exhibition. The search and visual analytics over the analysed text documents is made possible through the **webLyzard dashboard** which will also be presented in D2.2.
3. Identifying relevant text that may be directly **inserted into an immersive scene**, or, if the text is too long to be convenient for direct insertion, summarised to a much shorter text (cf. Section 3.1), or alternatively could form the basis for a textual prompt into a Generative AI system that can generate as a result a visual, audiovisual or 3D representation of the text input. For example, a journalist may wish to illustrate a news story in XR through insertion of some of the more recent article titles, or a two sentence summary of the story (derived from a longer article), or a set of images and/or 3D models which

represent aspects of the story (derived through text prompts to a Generative AI system).

The requirements analysis task of the TRANSMIXR use cases identifies which Web and social media sources should be used to collect relevant documents. A document is an abstraction of one Webpage or social media post. We store as a document not only the extracted and cleaned text but also additional metadata about the page/post based on our analysis components.

As demonstrated in the case of Twitter/X, where access to tweets for research purposes was heavily restricted after the change of ownership, we are aware that we can and should not rely on social network data, which is already limited for research purposes and may become even more restrictive in the future. Therefore, while we do have content mirrors for the Twitter and Facebook content APIs, we focus our online data collection rather on Web pages. We use an open source Web crawler and follow the respective terms and conditions of crawled sites (e.g. 'robots.txt', a text file Web admins can add to their Website which specifies what can and can not be crawled by automated bots). Since Webpage structure (HTML+metadata) tends to differ across sites, we generally have to modify our content mirrors for each Website or group of sites in order to correctly extract the main text and associated metadata (e.g. author, publication date).

This data collection pipeline has been used to collect more than 1.75 billion online documents in the past 10 years and the repository grows by approximately 50 million documents per month, primarily sourced from global news media in English, French, German, Spanish and Dutch as well as (previously) Twitter news content. However, a part of the daily data collection is research project specific, e.g. predefined lists of Websites, social media channels or topical searches.

In TRANSMIXR, while providing the whole existing repository to partners, we will focus in the news pilot on the collection of **local French media sources** and in the cultural heritage pilot we can include all relevant stakeholder Websites (e.g. the news section of museums and galleries) as well as selected cultural heritage objects from Europeana (https://www.europeana.eu/. We can make queries to the API, for which a content mirror exists).

2.2. YouTube Video Ingestion

The collection of video material (or references to online content) is a special case as the video is not processable by our existing text document ingestion pipeline. In the case of YouTube, for example, while we create a "textual" document for each retrieved YouTube video based on the metadata retrieved via the API (e.g. a title and description is present), we extend our analysis workflow by calling CERTH's video analysis service (see Section 3.2). The video metadata response is then appended to the document and video search as well as visual analytics based on the metadata will also be available via the webLyzard dashboard (more details in D2.2).

From our side, YouTube video retrieval is based on a set of textual search queries executed daily on the YouTube API endpoint, collecting all matching videos (the results are ordered by relevance) until our daily limit is reached. To ensure only recently posted videos are collected (as our focus is on current news), we use the `publicationAfter` filter set to 24 hours before the API call.

Currently, we have primarily two sources for YouTube video collection set up:

1. **YouTube News** automatically collects daily videos posted to YouTube which match the current trending news stories. Depending on the search queries generated automatically by keyword analysis over the global news feed, we collect anything from 100 to 800 videos daily.
2. **YouTube Media** will be used to collect daily YouTube videos matching topics defined in the news pilot, beginning in agreement with AFP for their news pilot with the term “climate change”.

2.3. AFP Video Ingestion

We have also been given access to an API from project partner AFP in order to receive news articles from their official news feed with media (videos) that match our predefined topics for the news pilot (beginning with “climate change”). A search endpoint receives the query string and other parameters (e.g. only news articles from the last 24 hours), and returns documents which match the query string in either the title or description of the news article and the other parameters. We will collect both English and French language news articles. Referenced media files (videos) will be passed to CERTH’s video analysis service and the metadata integrated into the news article’s document just as described for YouTube in the previous section.

2.4. Future Work

Content collection can be extended to other sources as long as it is technically feasible and conformant to the respective data access rights (Website terms for crawling, API definitions).

Additional Websites, or sets of Websites, can be added at request of partners, either to an existing content feed (e.g. global news) or as a new content feed. Our content mirror for Webpage collection includes a generic data ingestion component which uses the typical HTML page structure and metadata to identify and extract the page content (title, main text, author, publication date). When a Website has a particularly distinct structure which does not work well with the generic component, we can also implement content mirrors for specific Web domains.

Since social network APIs tend to be restrictive in terms of quotas for research purposes, we tend to be very conservative about setting up social media collection feeds. We can collect public posts on Facebook and Instagram according to specific accounts (or FB groups) or keywords, as requested. YouTube also provides an API with quota, which can be reconfigured as needed to use different search queries.

We can also add support for other sources provided an API is available, as can be seen by the integration of AFP content. We consider adding the Europeana API for the cultural heritage pilot. In these cases, a wrapper needs to be implemented which can map the API response into our document format and make the content available to our data ingestion pipeline (to complete the analysis and annotation of the content item).

3. Metadata Extraction Components

The purpose of the metadata extraction components is to apply appropriate content analysis algorithms to the raw content, in order to produce structured metadata about that content.

3.1. Textual Understanding

All data collected by MOD is stored in the same document structure (JSON) and indexed in an Elasticsearch index (scalable and fast retrieval) by webLyzard. These documents already contain directly extractable metadata fields such as title, main text, author and publication date. MOD provides a “data ingestion pipeline” which analyses the documents produced by the content mirrors (Section 2.1) and using NLP and NER technologies (Natural Language Processing, Named Entity Recognition), we extract additional metadata from the text such as keywords and entities which act as additional annotations of each document. The NLP components are built using the SpaCY library (https://spacy.io/) and our NER component is self-implemented, known as Recognyze (Weichselbraun, 2019), and uses lexical rules and dictionaries to identify references in text to Persons, Organisations, Locations or Events. We will look in due course at potential improvements to this pipeline through the use of large language models (LLMs) as initial experiments have been promising (e.g. fine tuning ChatGPT to take text as input and respond with entities and relations in that text).

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

This pipeline functions multilingually for multiple primary European languages (English, French, German, Spanish, Italian, Dutch). The detected entities are linked to known entities in our local knowledge graph (which is a curated copy of Wikidata content, where additional entities have been introduced on a project-specific basis). In the case of documents with videos (YouTube, AFP news), the metadata produced for each document will be enriched by a call to CERTH's video analysis service - the metadata response is stored in our document model and available to support video search and retrieval (see the next section for more details).

3.2. Video Understanding

This task produces semantic descriptions of multimodal media items to support computational understanding.

3.2.1 Problem Statement

The scope of this task is the multimodal media understanding. There are several integral steps towards the video analysis and understanding goal, including temporal segmentation, keyframe extraction, etc. To enable early experimentation at the starting phase of the project, CERTH initially deployed an analysis REST service, leveraging relevant state-of-the-art algorithms that the CERTH team has developed in previous research projects. Our goal is to build upon established methodologies in order to provide improved in-depth video analysis. To this aim, CERTH also developed improved training pipelines focusing on the fundamental target of event recognition, delivering enhanced performance.

3.2.2 State of the Art Survey

The fundamental tasks involved in the context of multimodal media understanding are briefly discussed in this Section.

For video content understanding, the shot boundary detection (SBD) is one of the most essential components. In general, the performance of a shot detection algorithm is based on its ability to detect transitions (shot boundaries) in a video sequence. An approach by (Hassanien, 2017), proposes a spatial-temporal feature-based Convolutional Neural Network (CNN) that detects wipe transitions and other gradual transitions. Similarly, in (Mondal, 2018), an effective detection technique is proposed, which is capable of discriminating the changes caused by the transitions from one shot to the other in the presence of different types of disturbances, like large movements of objects or camera. Moreover, a recent

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

approach from (Zhu, 2023) proposes a framework for video shot boundary detection, spanning various sophisticated 3D CNNs and Transformer models.

For video scene segmentation, preliminary shot boundary detection is performed, involving the task of identifying transition points within videos by assessing frame similarities. (Rao, 2020) approaches the challenge by transforming it into a binary classification task on a shot-by-shot basis. It employs a boundary-centric model that aggregates features from neighboring shots within a preset sliding window to formulate predictions. (Wu, 2022) introduces a Self-Supervised-Learning (SSL) strategy to ensure scene consistency. Rather than focusing on learning the scene boundary features, their approach involves the introduction of a straightforward temporal model with less inductive bias to verify the quality of the shot features.

In the concept detection part, the limitation of Deep Neural Networks (DNNs) to treat all classes fairly during the training procedure has been mostly studied in the context of class imbalanced learning (Sarafianos, 2018). Moreover, the identification of classes receiving little attention during training, as described above, is a relatively unexplored topic. The work of (Kiziltepe, 2021) integrates CNNs and Recurrent Neural Networks (RNNs) in video classification problem, highlighting that key-frame extraction is a crucial pre-processing step. An approach of 3D CNN pipeline is utilized for concept segmentation in video classification by (Syed, 2022). This method leverages a model pre-trained on a substantial action recognition task as an encoder, enabling the performance of unsupervised video classification.

Video event recognition describes the task of recognizing high level events and actions in a video sequence (Jiang, 2013). Event recognition is associated with various applications, rendering it a task of critical importance (Vieira, 2022), (Oh, 2011), (Herath, 2017). The advent of deep learning in a wide spectrum of computer vision tasks, including event recognition, provided powerful models achieving improved performance (Yao, 2019), (Ma, 2017). For example, a model called ViGAT, consisting of a ViT backbone (a neural network architecture based on Vision Transformers) in order to obtain feature representations of frames and objects and an attention-based network head in order to identify the most interesting scene parts, is proposed in (Gkalelis, 2022). Next, an approach for the unsupervised pre-training of a graph attention network block is incorporated into the ViGAT model in (Daskalakis, 2023). Finally, following the outstanding performance of LLM-based methodologies in NLP and computer vision tasks, they have recently been applied for video recognition tasks. For example in (Wu, 2023a) the authors aim at improving the transferability of powerful vision-language pre-training models for downstream video classification tasks, employing their textual encoders, while in (Wu, 2023b) a method that uses a video attribute association mechanism and a temporal concept spotting mechanism, aiming to build a bridge between visual and textual domain of vision-language models is proposed.

Visual sentiment analysis is the problem of identifying the sentiment tone expressed within an image. It presents a challenge due to the increased level of human subjectivity in the classification process, compared to other image classification tasks. Similarly to such tasks, deep CNNs are widely used; many works, e.g., (Campos, 2016) (Islam, 2016) rely on transfer learning by performing fine-tuning on pre-trained networks. In (Huang, 2019), the authors exploit the discriminative features in texts and images using a mixed fusion framework for sentiment analysis that involves investigating the underlying correlation between visual and semantic content, while (Chandrasekaran , 2022) based on an unique approach, uses existing pre-trained transfer learning models for predicting sentiment.

Cross-modal information retrieval refers to the task where queries from one or more modalities (e.g., text, images etc.) are used to retrieve items from a different modality. To perform text-video retrieval, typically the videos or video parts, along with the textual queries, need to be embedded into a joint latent feature space. Early approaches to this task (Markatopoulou, 2017) (Habibian, 2017) try to annotate both modalities with a set of pre-defined visual concepts, and retrieval is performed by comparing these annotations. Although various DNN architectures have been proposed to this end, with their general strategy being the same: encode text and video into one or more joint latent feature spaces where text-video similarities can be calculated. In contrast, (Bain, 2021) relies on a transformer architecture and does not employ trained image DNNs as feature extractors. BridgeFormer (Ge, 2022) incorporates a specialized module, where an encoder is trained to have an enhanced awareness of regional objects and temporal dynamics.

3.2.3 Initial Video Analysis Service

This REST service supports video, images and text. The video analysis is performed in three levels of temporal segmentation. The videos are fragmented into scenes and shots. Shots are segments of the video that are captured uninterruptedly by a single camera. Scenes are semantically and temporally coherent segments corresponding to the video's story-telling parts. Scenes are larger temporal fragments of a video and are comprised of one or more shots.

- Video level:
    - Visual event detection
    - Sentiment analysis
    - Scene segmentation

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

- Scene level:
    - Keyframe extraction
    - Shot segmentation
- Shot level:
    - Keyframe extraction
    - Visual concept detection
    - Sentiment analysis
    - Cross-modal signature extraction

The methods supported for image analysis:

- Visual concept detection
- Sentiment analysis
- Cross-modal signature extraction

The text analysis supports:

- Cross-modal signature extraction

Cross-modal signature extraction involves retrieving or associating information across various modalities, such as text, images, and videos. Those signatures can be later used for video retrieval by non-CERTH services.

An overview of the analysis REST service is given in Fig. 3. The scene segmentation of the videos relies on a method introduced by (Gygli, 2017). The algorithm considers a highly efficient CNN architecture by making it fully-convolutional in time. It poses the shot boundary detection as a binary classification problem. The objective is to correctly predict if a frame is part of the same shot as the previous frame or not. From this output, it is trivial to obtain the final shot boundaries. The process of segmenting the video into shots is based on the method (Souček, 2019). The proposed TransNet architecture, takes as input a sequence of N consecutive video frames and by applying a series of 3D convolutions, it returns a prediction for every frame in the input. Each prediction expresses how likely a given frame is a shot boundary. The events detection task relies on the existing method from (Gkalelis, 2022). As previously mentioned, a ViT network derives feature representations of the objects and frames, obtaining rich bottom-up information about the video scenes. An attention-based network head (called ViGAT head) is factorised along the spatial and temporal dimensions in order to identify the most interesting scene parts and thus to understand and encode temporal correlation between frames in video. For the visual concept detection, a proposed method (Gkalelis, 2020) called Subclass Deep Neural Network (SDNN), enhances neglected classes by creating subclasses. By emphasising subclass separation and creating linear subspaces for neglected classes, it reduces the need for complex features, improving DNN training effectiveness and overall generalisation performance. For the sentiment method, (Pournaras, 2021) achieves top performance in image sentiment analysis by leveraging knowledge from five neural networks with diverse architectures. Feature vectors are extracted from each network, classified as either in-domain (trained for image classification) or out-of-domain (trained for other tasks), contributing to the analysis based on their respective expertise in different domains. For the signature extraction, we leveraged the effectiveness of (Galanopoulos, 2023). This study introduces a novel cross-modal network architecture, termed TxV, to effectively merge diverse textual and visual features for text-based video retrieval. It extends a text processing strategy to visual information and employs a multiple latent space learning approach.

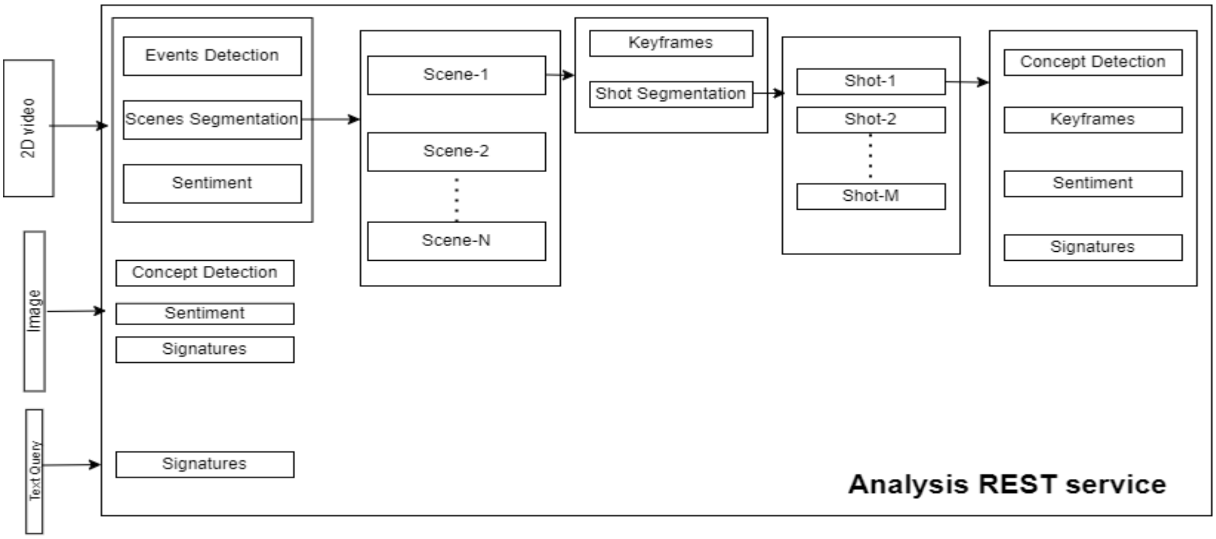


Figure 3: An overview approach of the REST analysis service.

3.2.4 Video Event Recognition

CERTH focused also on event recognition, proposing two novel training approaches, aiming to improve the baseline recognition performance in terms of accuracy. The proposed approaches are model-agnostic, i.e., can be applied to any baseline model for recognition tasks, regardless of their complexity. The proposed approaches are guided by the insights provided in the anchored-based object detection methodologies (Liu, 2016), (Ren, 2009). The fundamental concept underpinning the aforementioned methodologies, where the goal of the task is to predict the bounding boxes of objects of interest, is that it is easier for the network to learn offsets instead of absolute coordinates. That is, the network is provided with predefined boxes, also known as anchors, and the goal is formulated as prediction of offsets, instead of predicting the absolute coordinates of the bounding box.

Considering the event recognition task, we correspondingly propose to define an anchor, and instead of learning the event recognition label, to learn a percentage change of the one-hot ground truth event label with respect to the defined anchor. That is, the problem is transformed into predicting the offset instead of predicting the ground-truth label. More specifically, considering a neural network $f(x;\theta)$, with input x and parameters θ, and the one-hot event label y, our new targets for training the network are formulated as:

$$y_{ch} = y/\alpha - 1, \quad (1)$$

where α is the anchor. Thus, the network is trained to predict offsets. Then, during the test phase, the predictions are converted back to the original space of the ground-truth labels according to the equation, and the performance of the model in terms of test accuracy is evaluated:

$$\hat{y} = \alpha \cdot (\hat{y}_{ch} + 1). \quad (2)$$

We proposed two distinct approaches for defining the anchor: the Prototype Anchoring for image and video Recognition (PAR) and the Online Anchored-based Training for image and video recognition (OAT). The aforementioned approaches for defining the anchors are briefly presented in the subsequent subsection.

3.2.4.1 Prototype anchoring for image and video recognition

As previously mentioned, we aim at improving the event recognition performance in terms of accuracy. To this end, we propose to train a model in order to learn the percentage change of the one-hot ground truth event labels with respect to an anchor, instead of learning the ground truth labels. In order to define the anchors, in this approach, named Prototype Anchoring for image and video Recognition (PAR), we train a network with the ground truth event labels only for a few epochs (10 epochs of conventional supervised training are realised in the conducted experiments). Then, we utilise this model in order to extract the feature representations at the output layer, and use them in order to compute the anchors. The anchors are computed as the class centres at the output layer of the model. That is, the mean vectors of the extracted feature representation for each event class. Then, we transform the problem into learning the offsets, and train the network with the formulated anchored targets. It should be highlighted that the anchors are computed once, thus this process negligibly affects the computation cost.

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

During the test phase, the predictions are converted back to the original ground-truth label space in order to evaluate the performance of the method in terms of accuracy, according to eq. (2). Since each anchor is associated with the class of each sample, in order to avoid employing such information, we use the nearest anchor in the space generated by the output layer of the model, in terms of Euclidean space, to perform the aforementioned transformation. The above training and test phases of the proposed PAR approach are illustrated in Figure 4.

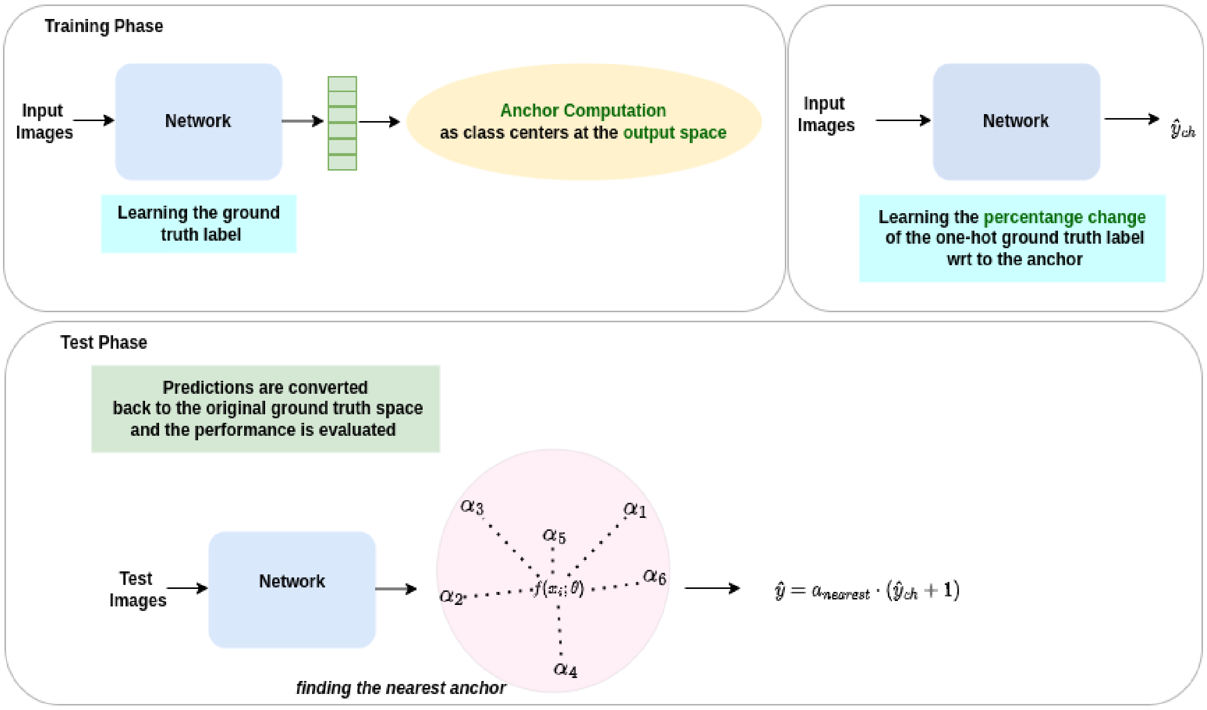

Training Phase
Input Images
Network
Anchor Computation as class centers at the output space
Learning the ground truth label
Input Images
Network
$\hat{y}_{ch}$
Learning the percentange change of the one-hot ground truth label wrt to the anchor
Test Phase
Predictions are converted back to the original ground truth space and the performance is evaluated
Test Images
Network
α_3
α_1
α_5
α_6
α_2
$f(x_i;\theta)$
α_4
$\hat{y} = a_{nearest} \cdot (\hat{y}_{ch} + 1)$
finding the nearest anchor

Figure 4: A schematic presentation of the proposed PAR approach.

3.2.4.2 Online anchored-based training for image and video recognition

In the second proposed approach, named Online Anchored-based Training (OAT), we propose to compute the anchors in an online fashion, dynamically, and also without using any class label information. To do so, the anchors are computed as the mean vectors of all the feature representations of the samples in a batch at the output space of the model. Thus, the network is trained to predict the percentage change of the one-hot ground truth labels with respect to the computed anchors. During the test phase the reverse transformation is performed, according to eq. (2). In this approach, we use as anchors the corresponding mean vector of the feature representations of the samples in the test batches, based on the fundamental assumption of supervised learning that training and test samples are drawn from the same distribution. The training and test phases of the proposed OAT approach are illustrated in Figure 5.

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

It should be emphasised that both the proposed approaches can be applied for generic recognition tasks, other than event recognition. To validate the aforementioned claim, we have included evaluation results on Cifar-10 datasets, in the Results Section. Both the described approaches are being prepared to be submitted to international conferences.

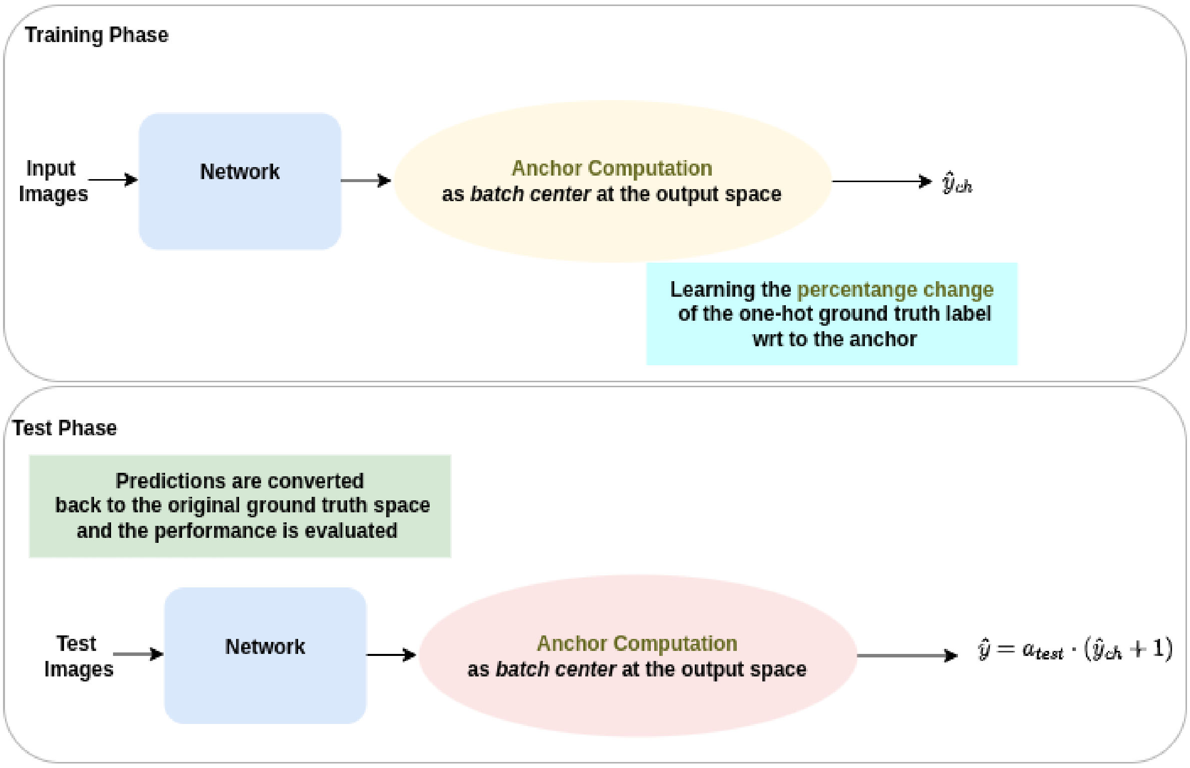

Training Phase
Input Images
Network
Anchor Computation as *batch center* at the output space
$\hat{y}_{ch}$
Learning the percentange change of the one-hot ground truth label wrt to the anchor
Test Phase
Predictions are converted back to the original ground truth space and the performance is evaluated
Test Images
Network
Anchor Computation as *batch center* at the output space
$\hat{y} = a_{test} \cdot (\hat{y}_{ch} + 1)$

Figure 5: A schematic presentation of the proposed OAT approach.

3.2.5 Implementation Details and Use

Regarding the REST service, state-of-the-art models are incorporated, trained on popular datasets. The concept detection model is pre-trained on the publicly-available YoutTube-8M dataset with 3862 classes (Sami Abu-El-Haija, 2016), whereas the event detection method is pre-trained on the ActivityNet dataset with 197 classes (Heilbron, 2015). The scene segmentation model is pre-trained on a dataset of 79 videos obtained (by the authors) from YouTube with automatically generated transitions such as cuts, dissolves and fades, whereas the shot segmentation model is pre-trained on the TRECVID IACC.3 dataset (Awad, 2017), including segments of 3000 IACC.3 randomly selected videos with only hard cuts and dissolves considered. The sentiment model is pre-trained by extracting vectors from

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

models trained on diverse datasets, including 1000-classes ImageNet, 11k-ImageNet (Russakovsky, 2015), YouTube-8M, MSR-VTT (Xu, 2016), among others. The cross-modal model is also trained on various datasets, such as ActivityNet (Heilbron, 2015), MSR-VTTT (Xu, 2016), TGIF (Li, 2016), and Vatex (Wang, 2019).

To request the processing of a video, submit a POST call to https://transmixr-idt.iti.gr/video-annotation. The URL of the video and the user key must be included as parameters in the body of the request in JSON format. The service supports videos from various online platforms and social media such as YouTube, Facebook, Twitter, Vimeo etc. as well as custom URLs. The reply to the request is a small JSON document that includes a short message (Table 2) and, if the request was valid, an item id that uniquely identifies the video and is later used to retrieve the status and the results.

Table 2: Possible JSON messages as reply to the video request.

message	*A short response message. It can be one of the following:*	
	status code	***Message***
	200	**The REST call has been received. Please check the status of the analysis via the appropriate REST call** *Explanation: The request was valid and video processing has started*
	403	**Not valid user key** *Explanation: The user_key used in the parameters is not valid*
	403	**Limit exceeded. Try again later** *Explanation: The maximum total video duration sent for processing the last 10 hours was exceeded*
	404	**The video URL is broken** *Explanation: The content of the video_url parameter is not valid*
	404	**XX : Error. No such mode** *Explanation: The service allows specific analysis mode, such as “video-annotation”. This error message informs that the mode requested does not exist.*
	404	**Bad request** *Explanation: The request was not correctly formed*
	409	**Video is currently being processed** *Explanation: The request sent is currently being processed*
	409	**Video already in queue** *Explanation: The video has already been sent and is waiting in the queue*

To request the processing of an image/image collection, submit a POST call to https://transmixr-idt.iti.gr/image-annotation. The image/image collection can be either a zip file containing one or more images or individual image URLs that can be downloaded. In the case of the zip file, the URL of the zip file should be provided in the “zip_url” parameter. In the case of individual image URLs, a list of the URLs should be provided in the “image_urls” parameter. In both cases the image collection can consist of between 1 and 1000 images. The JSON reply to the request includes the following possibilities (Table 3).

Table 3: Possible JSON messages as reply to the image request.

Message	*A short response message. It can be one of the following:*	
	status code	***Message***

	200	***The REST call has been received. Please check the status of the analysis via the appropriate REST call*** *Explanation: The request was valid and image processing has started*
	403	***Not valid user key*** *Explanation: The user_key used in the parameters is not valid*
	404	***Bad request*** *Explanation: The request was not correctly formed*
	404	***XX : Error. No such mode*** *Explanation: The service allows specific analysis mode, such as "image-annotation". This error message informs that the mode requested does not exist.*
	409	***Image collection is currently being processed*** *Explanation: The request sent is currently being processed*
	409	***Image collection already in queue*** *Explanation: The image collection has already been sent and is waiting in the queue*

Text processing consists of signature extraction only. In contrast to the image/video analysis, which is asynchronous, the text signature extraction is synchronous. This means that the processing result (the signature vector) is instantly returned to the processing request. To process, submit a POST call in https://transmixr-idt.iti.gr/text-sign-extr with body including <text> and the <user_key>.

The second step in the workflow is to check the status of the analysis. Typically, this should be done periodically until the status reports that the analysis has ended. Some of the status messages are temporary and some are final. Final status messages indicate that processing has finished (either successfully or unsuccessfully). The table with the possible status messages makes the distinction between temporary and final status messages clear. To issue a request for the status, submit a GET call to https://transmixr-idt.iti.gr/status/<item_id>, where <item_id > is the item_id previously retrieved. The status of the analysis can be one of the following (Table 4).

Table 4: Possible JSON replies to the request for status analysis

status		
	status code	***Message***
	200	**VIDEO_WAITING_IN_QUEUE** *Explanation: The video is waiting to be processed in the queue* *Type: Temporary*
	200	**ITEM_WAITING_IN_QUEUE** *Explanation: The image collection is waiting to be processed in the queue* *Type: Temporary*
	200	**VIDEO_DOWNLOAD_STARTED** *Explanation: Downloading of the video has been initiated* *Type: Temporary*
	200	**IMAGE_COLLECTION_DOWNLOAD_STARTED** *Explanation: Downloading of the image collection has been initiated* *Type: Temporary*
	200	**VIDEO_DOWNLOAD_FAILED** *Explanation: Video downloading has failed* *Type: Final*

	200	**VIDEO_DOWNLOAD_TIMEOUT** *Explanation: The video was taking too long to download so downloading was cancelled* *Type: Final*
	200	**MAX_VIDEO_DURATION_EXCEEDED** *Explanation: The is an 1 hour limit to the duration of the video that can be submitted* *Type: Final*
	200	**VIDEO_ANALYSIS_STARTED** *Explanation: The analysis of the video has started* *Type: Temporary*
	200	**IMAGE_COLLECTION_ANALYSIS_STARTED** *Explanation: The image collection analysis has started* *Type: Temporary*
	200	**VIDEO_ ANALYSIS_COMPLETED** *Explanation: The analysis of the video has completed successfully* *Type: Final*
	200	**IMAGE_COLLECTION_ANALYSIS_COMPLETED** *Explanation: The image collection analysis has completed successfully* *Type: Final*
	200	**VIDEO_ANALYSIS_FAILED** *Explanation: The analysis of the video has failed* *Type: Final*
	200	**IMAGE_COLLECTION_ANALYSIS_FAILED** *Explanation: The image collection analysis has failed* *Type: Final*
	404	**Wrong file name or status file does no longer exist** *Explanation: No status of the requested item_id exists*

The final step of the workflow is the retrieval of the results. Once the results have been created, they can be retrieved for 48 hours. After this point, they are no longer available on our server. To get the result, issue a GET call to https://transmixr-idt.iti.gr/result/<item_id>, where <item_id> is the identifier of your video/image collection.

Regarding the proposed anchored-based training approaches, they were implemented using the PyTorch framework. The mini-batch gradient descent is used for the networks' training. The learning rate is set to 0.001 and the momentum is 0.9. Mini-batch size is set to 32 samples. All the models are trained with a NVIDIA GeForce GTX 4080 with 24 GB of GPU memory for 100 epochs.

3.2.6 Results

Concerning the output (Listing 1) JSON file of the REST service for video analysis, the first level of the output contains the following fields: "expires_at", "framerate", "generated_at", "generated_by", "version" and the following arrays: "scenes", "shots", "events", "key_frames" and "sentiment". The "shots" are included in their "scene". For each scene, the JSON contains an id of the scene ("scene_id"), the "begintime" and "endtime" of the scene in seconds, the list of "keyframes" with the timestamp ("time") and "url" of the keyframe included, and the list of "shots". For each shot the JSON contains an id of the shot ("shot_id"), the "begintime" and "endtime" of the shot in seconds, the list of "keyframes" with the timestamp ("time") and "url" of the keyframe included, a list of the top 30 visual concepts (“concepts”) with their confidence score (the higher the more relevant), and the sentiment in both binary (positive, negative) and numerical form (0-1; higher values indicate more positive sentiment). Each shot, also, contains a “signature” which is a list of floating-point numbers that acts as a visual descriptor for the shot. Please note that the returned JSON includes only URL references for the produced scenes and shot keyframes and those graphic files can be downloaded independently, if needed.

Reply

```
Status code: 200
{   "expires_at": "2020-07-15 10:51:13.304437",
   "framerate": 25.000,
   "generated_at": "2020-07-01 10:51:13.304426",
   "generated_by": "https://transmixr-idt.iti.gr",
   "scenes": [
      {
         "scene_id": "Sc1",
         "begintime": 0.040,
         "endtime": 53.800,
         "keyframes": [
            {
               "time": 9.000,
               "url": "https://transmixr-idt.iti.gr/keyframe/becdf9cc16b2358b59b1cc12d938e580/shot4_1"
            },
            …
         ],
         "shots": [
            {
               "shot_id": "Sh1",
               "begintime": 0.040,
               "endtime": 1.400,
               "concepts": {
                  "Walking": 0.004,
                  "Smartphone": 0.005,
                  …
               },
               "keyframes": [
                  {
                     "time": 0.360,
                     "url":
"https://transmixr-idt.iti.gr/keyframe/becdf9cc16b2358b59b1cc12d938e580/shot1_1"
                  },
                  …
               ],
               "sentiment": {
                  "positive",  "0.866345"
               },
               "signature": [
                  0.05675,
                  0.00458,
                  …
               ]
            },
            …
         ],
      },
      …
   ],
    "sentiment": {
         "negative", "0.1745"
    }
   "events": {
      "Cleaning windows": -8.943937301635742,
      "Discus throw": -6.931890487670898,
      …
},

"summary": "https://transmixr-idt.iti.gr:443/summary/becdf9cc16b2358b59b1cc12d938e580",
"thumbnails": [
" https://transmixr-idt.iti.gr: 443/thumbnail/716797cdedd2d4f577f3503/1",
" https://transmixr-idt.iti.gr: 443/thumbnail/716797cdedd2d4f577f3503/2",
…
" https://transmixr-idt.iti.gr: 443/thumbnail/716797cdedd2d4f577f3503/5",
   ],
},    "version": "v1.1" }
```

Listing 1: The JSON output for a video analysis component

For the images analysis (Listing 2), the output JSON includes their names e.g. “img100.jpg”. For each image, it contains a “concepts” attribute with the top 30 visual concepts, the sentiment and the signature, all in the same form as the JSON for a video analysis. The dimensions of the image are also included as well as the boolean “analysis” attribute which states whether analysis has been performed on the image.

Reply

```
Status code: 200
{
   "img100.jpg": {
      "analysis": "True",
      "concepts": {
         "yt8m_top30": {
            "Animal": 0.010635197162628174,
            "Boat": 0.9999825954437256,
            …
         }
      },
      "sentiment": {
         "negative",
         "0,078937"
      },
      "dimensions": "150x200",
      "signature": [
         0.10345,
         0.02353,
         …
      ]
   },
   …
}
```

Listing 2: The JSON output for an image analysis

The text analysis consists of signature extraction only (Listing 3).

Reply	Status code: 200
	{
	"signature": [0.00345, 0.04586, 0.01345, …]
	}

Listing 3: The JSON output for a text analysis

Regarding the experimental evaluation of the anchored-based training approaches, four datasets are used to validate their effectiveness, i.e. Cifar-10 (Krizhevsky, 2009), UCF-101 (Soomro, 2012), BAR (Nam, 2020), and ERA (Mou, 2020). UCF-101 is an action recognition data set of 13,320 YouTube action videos, consisting of 101 action categories. The middle frame is derived from each video, forming a train set of 9,537 images and a test set of 3,783 images. BAR is a real-world image dataset with six action classes, consisting of 1,941 train images and 654 test images. ERA dataset is an event recognition in unconstrained aerial videos, consisting of 1,473 train images and 1,391 test images, divided into 25 event classes. Finally, Cifar-10 consists of 50,000 train images and 10,000 test images divided into 10 classes.

Network architectures of varying complexity are used in all the utilised datasets. More specifically, in all the event recognition datasets, ResNet-18 (He, 2016) and Wide-ResNet-50- 2 (Zagoruyko, 2016) (abbreviated as WRN-50-2), pretrained on ImageNet weights, are used. A linear layer is added to the output of the networks, with neurons equal to the number of the classes of each dataset. In the Cifar-10 dataset, two networks are also utilised, i.e., a simple lightweight model, consisting of two convolutional layers with six and sixteen kernels of size 5 × 5 respectively and three fully connected layers (128 × 64 × 10), and a heavyweight Wide-ResNet-28-10 model (abbreviated as WRN-28-10).

We apply the proposed PAR and OAT approaches on the aforementioned models and compare their performance against the baseline. As baseline is denoted a network of identical architecture trained with the ground-truth labels. Test accuracy is used as the evaluation metric. Best performance is printed in bold.

First, in Table 5, we present the evaluation results for the PAR approach on the generic recognition dataset, i.e. Cifar-10. As it is demonstrated the proposed training approach provides improved performance in terms of accuracy for both the utilized model architectures.

Table 5: Test accuracy for the proposed method against baseline using both the utilized models.

Method	Lightweight	WRN-28-10
Baseline	65.660	89.130

PAR	**67.130**	**91.440**

Next, in Tables 6 and 7 the corresponding results on the event recognition datasets are presented for the ResNet-18 and the WRN-50-2 models, respectively. The proposed method achieves considerably improved accuracy in all the considered cases.

Table 6: Test accuracy for the proposed method against baseline using the ResNet-18 model on the three event recognition datasets.

Dataset	Baseline	PAR
UCF-101	72.403	**74.174**
BAR	61.621	**63.150**
ERA	54.565	**55.356**

Table 7: Test accuracy for the proposed method against baseline using the WRN-50-2 modelon the three event recognition datasets.

Dataset	Baseline	PAR
UCF-101	78.853	**79.170**
BAR	70.183	**73.547**
ERA	56.506	**59.310**

As it was mentioned, the proposed PAR training approach requires the model to be trained with the ground truth labels, in order to acquire the feature representations for computing the anchors as the class centres at the output layer. Only a few epochs of conventional training are adequate. In all the experiments, 10 epochs of training are used, however an investigation of the optimal number of epochs is performed, validating also that the method provides enhanced test accuracy for various epochs of anchoring. The evaluation results on the UCF-101 dataset, using the ResNet-18 model, are provided in Table 8.

Table 8: UCF-101: Test accuracy of different epochs of anchoring using the ResNet-18 model.

Epoch	Test Accuracy
5	73.222
10	**74.174**
20	72.746

Subsequently, we provide the corresponding evaluation results considering the second proposed anchored-based training approach, i.e., the OAT method. First, the evaluation results on the Cifar-10 dataset for both the used model architectures are provided in Table 9, validating the effectiveness of the OAT approach in generic recognition tasks.

Table 9: Cifar-10: Test accuracy for the proposed method against baseline using both the utilized models.

Method	Lightweight	WRN-28-10
Baseline	65.660	89.130
OAT	**66.540**	**92.020**

Next, in Tables 10 and 11 the corresponding experimental results are provided for the event recognition datasets, using the ResNet-18 and WRN-50-2 models, respectively. As it can be observed, the proposed OAT method significantly improves the baseline accuracy for all the considered cases.

Table 10: Test accuracy for the proposed method against baseline using the ResNet-18 model on the three event recognition datasets.

Dataset	Baseline	OAT
UCF-101	72.403	**74.253**
BAR	61.621	**64.526**
ERA	54.565	**57.565**

Table 11: Test accuracy for the proposed method against baseline using the WRN-50-2 model on the three event recognition datasets.

Dataset	Baseline	OAT
UCF-101	78.853	**79.038**
BAR	70.183	**73.853**
ERA	56.506	**58.519**

3.3. 360 Video Understanding

As part of our research efforts in this period, we developed and implemented technologies for 360-degree video analysis that will be incorporated in the analysis REST service. These technologies rely on visual-content analysis specifically tailored

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

for 360 videos. In the following, we outline the initial method used to generate a 2D traditional video, focusing on the most important parts of the 360 video.

3.3.1 Problem Statement

The scope of this task involves understanding of the 360-degree videos by focusing on detecting the most salient events within a 360-degree video, considering both stationary and moving-camera videos. Initially, the focus is on extracting saliency maps for every frame of the 360-degree video. The goal is to convert these saliency maps and frames from the 360-degree video to a 2D traditional video, including the most salient parts in chronological order, so that can be analyzed by the REST API service described in Section 3.2.

3.3.2 State of the Art Survey

To simplify the storage and processing of 360-degree video content from a spherical domain, it is commonly projected onto a 2D plane. The most common projections are equirectangular and cube-map (called ERP and CMP in the following). The ERP projection directly employs the latitude and longitude on the sphere as the horizontal and vertical coordinates, respectively, on the original frame. However, this approach leads to increased distortion in the polar regions. The CMP projection involves mapping a spherical video onto an external six faces cube. In this projection, the upper and lower faces of the cube correspond to the polar regions, while the four faces in the middle correspond to the equatorial region. (Nguyen, 2018) used CNN-based estimation networks, predicting saliency map for each independent ERP frame, without taking account the temporal dynamic. Previous works (Zhang, 2018), (Vo, 2020) and (Qiao, 2021) introduced a spherical U-Net (Ronneberger, 2015), to reduce the distortion of the ERP frames.

3.3.3 Approach

An overview of the proposed approach is given in Figure 6. The input 360-degree video is subjected to ERP to form a set of omnidirectional planar frames. This set of frames is then analyzed by a Camera Motion Detection Algorithm (CMDA) that makes a decision on whether the 360-degree video has been captured by a static or a moving camera. The use of such an algorithm in combination with two different methods (including in the saliency detection part) was based on our study of the relevant literature that most methods deal with 360-degree videos that have been captured by either static or moving view conditions. So, based on the output of the camera motion algorithm, the ERP frames are forwarded to one of the saliency detection methods, which produce a set of frame level saliency maps. The ERP

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

frames and the extracted saliency maps are then given as input to the 2D Video Production, containing the detected salient events in the 360 video.

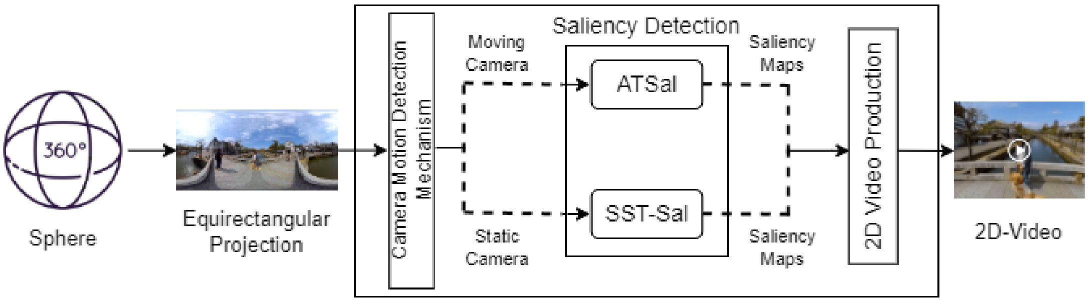


Figure 6: An overview of the proposed approach for 360 video analysis. Dashed lines indicate alternative paths of the processing pipeline.

The primary action zones within ERP frames are typically situated near the equator (Figure 7). Considering this, determining whether the 360-degree video was captured by a static or moving camera, is based on analyzing the northern and southern regions of the ERP frames. Such regions should exhibit limited variation across a sequence of frames when the 360-degree video is captured by a static camera. So, given the ERP frames of the input video, the CMDA focuses on specified frame regions and computes the phase correlation between consecutive frames. If the computed scores frequently exceed a threshold t_0, the algorithm declares the use of a moving camera; otherwise, it indicates the use of a static one.

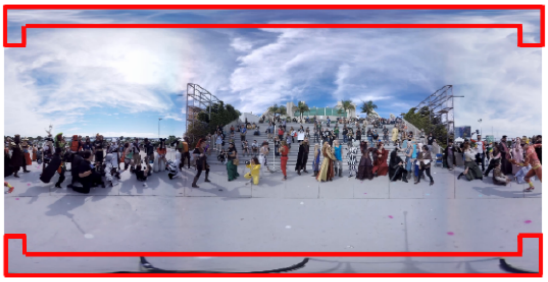


Figure 7: An ERP frame and its north and south regions (red-coloured bounding boxes) that are used by the camera motion mechanism.

The saliency detection part includes two models. For the moving view cases, we assessed the performance in ATSal (Dahou, 2020). The authors proposed a sophisticated architecture (Figure 8), in which they integrate saliency maps obtained from two distinct models: i) an attention model that is specialized in extracting global statistics through the application of the attention mechanism between the encoder-decoder on the ERP frame, and ii) SalEMA (Linardos, 2019) expert model, designed to learn more efficient temporal characteristics, via CMPs frames. The SalEMA expert model is composed of the SalEMA_Poles model, responsible for the North and South views and a SalEMA_Equator model, which focuses on the front, left,

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

right and back views. Given as input the ERP frame to the attention model and the six CMP faces to the SalEMA (SalEMA_Poles, SalEMA_Equator), the final saliency map is obtained after pixel wise multiplication between the outputs. For the stationary cases, we based on the approach of SST-Sal (Bernal-Berdun, 2022). The authors introduce an architecture (Figure 9) with encoder-decoder, based on Conv-Lstms that incorporates temporal information. Employing operations that take into account the specific characteristics of ERP projection enhances the accurate interpretation of 360 content. Their proposal is to extract the spatial features with temporal awareness, based on stationary cameras, and convert them to saliency maps.

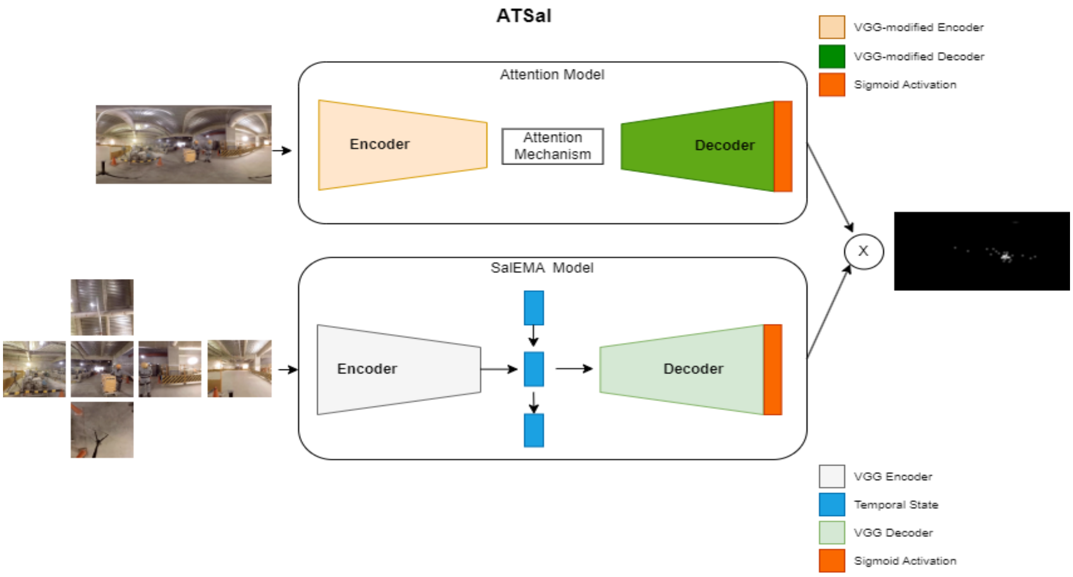


Figure 8: The architecture of the ATSal method. As input are the ERP and the CMP frames.

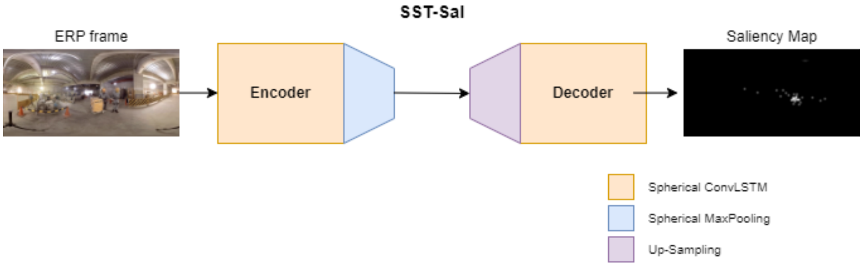
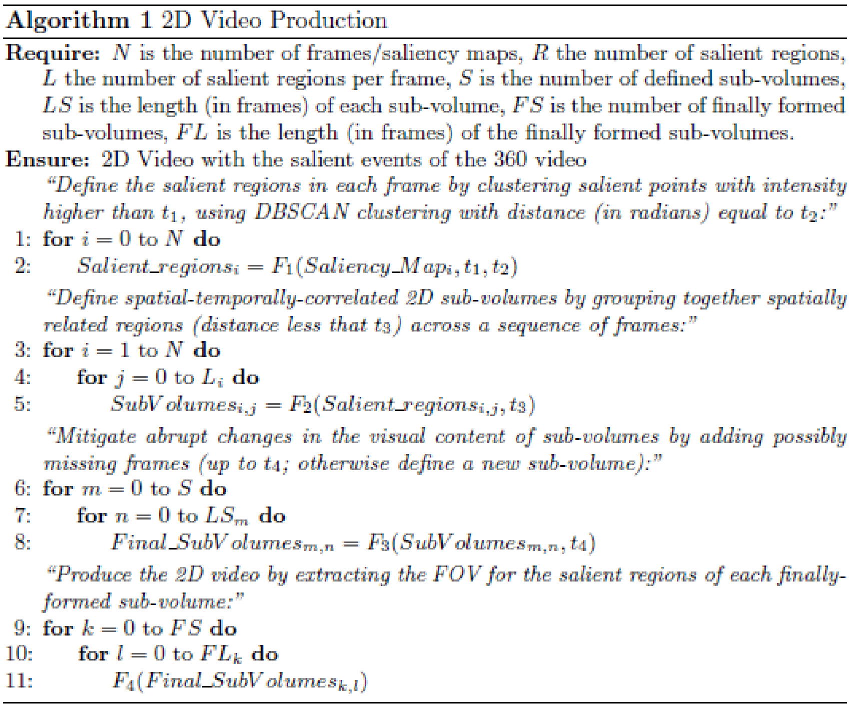


Figure 9: The architecture of the SST-Sal method.

The 2D Video Production algorithm takes as input the ERP frames and their extracted saliency maps (by the saliency detection methods), and produces a 2D video that contains the detected salient events in the 360 video. The procedure starts by identifying the salient regions of each ERP frame. As presented in Algorithm 1, this procedure starts by identifying the salient regions of each ERP frame. To form such a region in a given ERP frame, our method focuses on points of the associated saliency map that exceed an intensity value t_1, converts their coordinates to radians and clusters them using the DBSCAN algorithm (Ester, 1996) and a predefined distance t_2. Following, the salient regions that are spatially related across a sequence of frames are grouped together, thus establishing a spatial-temporally-correlated sub-volume of the 360 video. By taking into account the entire frame sequence, our method examines whether the salient regions of the f_i frame are close enough to the salient regions of one of the previous frames (f_{i-1}...f_1). If this distance is less than t_3, then the spatially-correlated regions over the examined sequence of frames (f_1...f_i) are grouped and form a sub-volume; otherwise, it creates a new sub-volume for each of the salient regions in the f_i frame. Given the fact that a spatially-correlated salient region can be found in non-consecutive frames (e.g., appearing in frames f_t and f_{t+2}), the applied grouping might result in sub-volumes that are missing one or more frames. To mitigate abrupt changes in the visual content of the formulated sub-volumes, our method adds the missing frames within each sub-volume. Moreover, in the case that a sub-volume contains a large sequence of missing frames (higher than t_4) our method splits this sub-volume and considers that the aforementioned sequence of frames does not contain a salient event. For each sub-volume, the developed method extracts the FOV of the salient regions, thus creating a short spatio-temporally-coherent 2D video fragment. Finally, the 2D video is formed by concatenating the created 2D video fragments for the different sub-volumes, in chronological order.



Algorithm 1 2D Video Production

```
Require: $N$ is the number of frames/saliency maps, $R$ the number of salient regions,
    $L$ the number of salient regions per frame, $S$ is the number of defined sub-volumes,
    $LS$ is the length (in frames) of each sub-volume, $FS$ is the number of finally formed
    sub-volumes, $FL$ is the length (in frames) of the finally formed sub-volumes.
Ensure: 2D Video with the salient events of the 360 video
    "Define the salient regions in each frame by clustering salient points with intensity
    higher than $t_1$, using DBSCAN clustering with distance (in radians) equal to $t_2$:"
for $i = 0$ to $N$ do
    $Salient\_regions_i = F_1(Saliency\_Map_i, t_1, t_2)$
    "Define spatial-temporally-correlated 2D sub-volumes by grouping together spatially
    related regions (distance less that $t_3$) across a sequence of frames:"
for $i = 1$ to $N$ do
    for $j = 0$ to $L_i$ do
        $SubVolumes_{i,j} = F_2(Salient\_regions_{i,j}, t_3)$
    "Mitigate abrupt changes in the visual content of sub-volumes by adding possibly
    missing frames (up to $t_4$; otherwise define a new sub-volume):"
for $m = 0$ to $S$ do
    for $n = 0$ to $LS_m$ do
        $Final\_SubVolumes_{m,n} = F_3(SubVolumes_{m,n}, t_4)$
    "Produce the 2D video by extracting the FOV for the salient regions of each finally-
    formed sub-volume:"
for $k = 0$ to $FS$ do
    for $l = 0$ to $FL_k$ do
        $F_4(Final\_SubVolumes_{k,l})$
```

Algorithm 1: Algorithm for the 2D traditional video production.

3.3.4 Implementation Details and Use

We used 208 videos of the publicly-available VR-EyeTracking dataset (Dahou, 2020) which is based on the work of (Xu, 2018). We excluded two videos due to the

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

ambiguity in their saliency maps. The 206 remaining high definition videos include diverse content, such as indoor scenes, outdoor activities, and music shows. The duration of the videos are between 20s and 60s with at least 25 fps. Further, 147 videos are recorded from a stationary camera perspective, while the other 59 are shot with a moving camera. We also used the high definition 85 ERP images of Salient360! (Gutiérrez, 2018) and 22 ERP images (Sitzmann, 2018) dataset. For extra evaluation, we used the 104 videos of Sports-360 (Zhang, 2018). The dynamic video contents involve five sports (i.e. basketball, parkour, BMX, skateboarding, and dance), and the duration of each video is between 20s and 60s. The dataset includes 84 stationary and 18 moving view videos.

The parameter t_0 of the CMDA, was set equal to 0.5. For the training process of the attention model of ATSal, we first trained the model for 90 epochs from scratch on the Salient360! and Sitzman dataset. After applying augmentation techniques, the dataset contains 2140 ERP images, where 1840 used for training and 300 for validation.

For the second stage of the training process, we used the VR-EyeTracking dataset. The training set includes 140 videos, while the validation set had 66 videos. We trained it with batches of 10 frames. For the fine-tuning of the SalEMA-Poles and SalEMA-Equators models, we employed different batch sizes: 10 frames for the Equator model and 80 frames for the Poles model. This decision was grounded in the observation that the Equator regions generally exhibit more motion compared to the Poles. The number of epochs for both models was 20. For the training process of the SST-Sal model, we changed the number of hidden-layers of the model to 9 (since we do not use the optical-flow frames) and the number of input channels equals 3. We trained the model on the static view data of the VR-EyeTracking, containing training set equal to 92 videos and validation set equal to 55 videos. We kept the number of sequences to 20 frames as input for 100 epochs. Concerning the 2D video production step:

- t_1 can range in [0, 255] and pixels with saliency score lower than 150 were not considered as important;
- t_2 represents the Haversine distance that affects the number of possibly identified salient regions within a frame by the applied the DBSCAN algorithm, and was set equal to 1.2;
- t_3 is associated to the Euclidean distance between salient regions in consecutive ERP frames, and given the input frames' resolution it was set equal to 100; and
- t_4 affects the duration of the created sub-volumes, and was set equal to 100.

3.3.5 Results

For the CMDA evaluation, we used a total of 329 equirectangular video tests, including Vr-EyeTracking, Sports-360 and Salient360! videos. The results (Table 12) indicate an overall accuracy rate of 88.75%. Significantly, we achieved a high accuracy rate of 94.85% specifically on predicting moving videos. These results confirm the effectiveness of our CMDA in handling a diverse set of video scenarios.
Since our model was applied in 360-degree video cases, we employed the metrics including Pearson Correlation Coefficient (CC) and similarity (SIM) established by (Bylinskii, 2019) to assess the performance of the saliency prediction models. CC measures the linear correlation between the distributions of the saliency maps, while the SIM focuses on the similarity between them. The results of the training process for the two models are presented in Table 13. The ATSal method performs better in the VR-EyeTracking dataset, both for static and moving videos. In contrast, SST-Sal achieves comparable results in the Static VR-EyeTracking dataset with the ATSal method, with minimal variance in similarity scores. However, SST-Sal outperforms ATSal on the Static Dataset of the Sports-360 dataset. Finally, to evaluate the impact of the utilized CMDA, we formed a large set of test videos (merged 104 test videos of the Sports-360 dataset with 66 test videos from the VR-EyeTracking dataset) and considered three different processing options:

a) the use of ATSal only,
b) the use of SST-Sal only, and
c) the use of both methods in combination with the developed CMDA.

The results in Table 14 document the positive impact of the CMDA, as its use in combination with the integrated saliency detection methods results in higher performance. Based on these findings, we have decided to include both models in our task. SST-Sal for stationary views and ATSal for moving scenarios. Figure 10 presents two sequences of ERP frames, the produced saliency maps by ATSal and SST-Sal, and the ground-truth saliency maps. Starting from the top, the first frame sequence was extracted from a 360 video captured using a static camera. From the associated saliency maps we observe that SST-Sal performs clearly better compared to ATSal and creates saliency maps that are very close to the ground-truth; instead, ATSal fails to detect several salient points. For the second frame sequence, which was obtained from a 360 video recorded using a moving camera, we see the exact opposite behavior. ATSal defines saliency maps that are very similar with the ground-truth, while the saliency maps of the SST-Sal method contain too much noise.

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

Table 12: Performance (Accuracy in percentage) of the CMDA

	Static Camera	Moving Camera	Total
Number of Videos	232	97	329
Correctly Classified	200	92	292
Accuracy	86.21%	94.85%	88.75%

Table 13: Ablation study about the use of the decision mechanism.

	CC	SIM
ATSal Only	0.290	0.241
SST-Sal Only	0.377	0.279
CMDA & ATSal or SST-Sal (Proposed)	0.379	0.280

Table 14: Performance of the trained ATSal and SST-Sal models on videos of the VR-EyeTracking (upper part) and Sports-360 (lower part) datasets

	VR-EyeTracking					
	Static View Videos (55)		Moving View Videos (11)		Total Videos (66)	
	CC↑	SIM↑	CC↑	SIM↑	CC↑	SIM↑
ATSal	**0.336**	**0.240**	**0.230**	**0.172**	**0.322**	**0.229**
SST-Sal	0.309	0.167	0.168	0.106	0.285	0.157
	Sports-360					
	Static View Videos (86)		Moving View Videos (18)		Total Videos (104)	
	CC↑	SIM↑	CC↑	SIM↑	CC↑	SIM↑
ATSal	0.270	0.251	0.270	0.243	0.270	0.249
SST-Sal	**0.464**	**0.372**	**0.273**	**0.283**	**0.436**	**0.358**

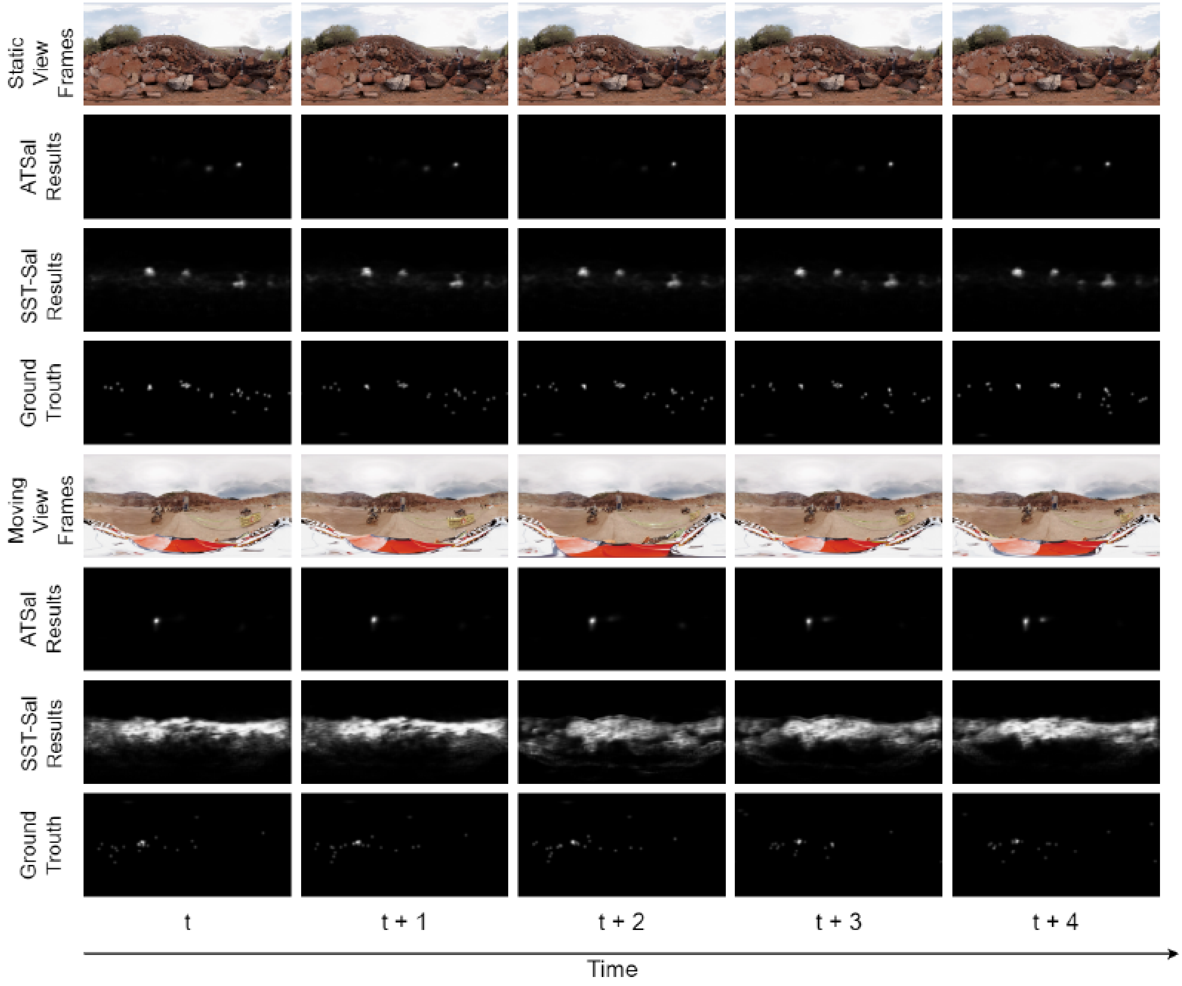


Figure 10: Qualitative comparisons between the output of the ATSal and SST-Sal methods on frame sequences of videos from the VR-EyeTracking dataset.

3.4. Volumetric Video Understanding

3.4.1 Problem Statement

This sub-task focuses on semantic understanding of volumetric video, consisting of 3D geometry and texture information of objects captured from the real world. Volumetric video is a relatively new form of media asset, with few publicly available datasets and very limited existing off-the-shelf solutions for media search-and-retrieval, processing and integration into XR environments. In particular, TCD's research expertise lies in volumetric videos captured of human performers usable, for instance, in mixed reality holograms. Current work within TRANSMIXR involves the analysis and deployment of state-of-the-art techniques for recognition, classification, segmentation, and automatic rigging of human volumetric video. The purpose of this is, on one hand, to integrate volumetric video search into media asset search tools (on par with other assets such as video, static 3D models, and 360 images/video) by developing techniques for semantic summarization and semantic annotation. Furthermore semantic analysis, as discussed in the grant proposal, will be used to improve usage and ease of integration of volumetric video with more traditional assets in XR experiences such as the use case pilots.

3.4.2 State of the Art

Rapid progress in 3D acquisition technologies, CAD models, and research on trainable 3D methods has led to a surge in the availability of 3D models, and consequently a demand for development of 3D shape retrieval solutions (Funkhouser, 2003). In more recent years, such research has extended also into the 3D animation domain, including estimation of parametric human models from motion capture data (Mahmood, 2019), but there is still a paucity of work related to semantic understanding of volumetric video.

On the other hand, video summarization and understanding of human actions in a 2D context is a well-established field, utilising cues such as features, objects, actions and trajectories in video sequences. Action and trajectory-based clustering approaches, for instance, entail the extraction of skeletal points from humans in 2D videos (Cao, 2019; Zhu, 2023), conducting clustering of actions in lower dimensional space, and analysing changes in joint keypoints. Standardised benchmarks already exist for human activity understanding in 2D (Caba, 2015). Some of these previous works in 2D have notably influenced recent research on volumetric video summarization including automated rigging, pose generation from textual descriptors, and clothed human reconstruction from single images. Keyframe extraction of volumetric videos can be executed by leveraging 3D shape descriptors (Kazhdan, 2003; Novotni, 2003) and assessing the similarity score between frames in a volumetric video sequence (Moynihan, 2021). The ideal properties of a descriptor for shape retrieval include invariance to scale, orientation, position, and non-rigid transformations. However, when considering human volumetric videos, a descriptor that is sensitive to non-rigid transformations becomes crucial for effectively discriminating various actions.

3.4.3 Approach

TCD compiled a database of human volumetric videos in order to serve as reference and training data for continued development. The data was gathered from scientific research datasets and a small number of publicly available commercial data samples. Further datasets were captured in-house by the team and their collaborators. Based on an analysis of the scope and varying formats in this compiled database, TCD designed a pipeline for semantic understanding of volumetric video, as shown in Figure 11. The first stages of this pipeline have already been implemented and work is ongoing on refining, extending and testing this with more varied datasets.

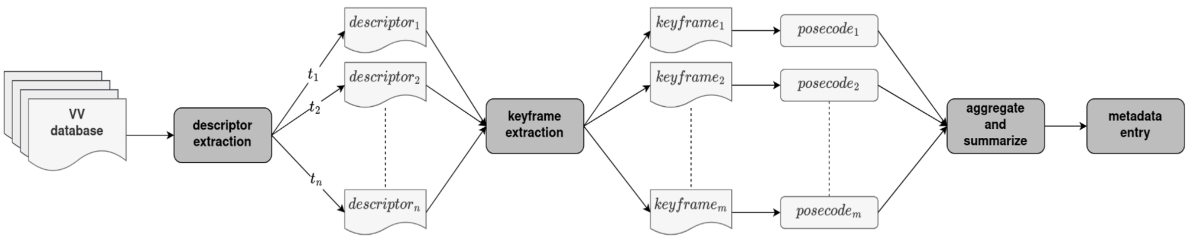


Figure 11. Methodology for keyframe extraction and posecode annotation.

A description of the components of the pipeline follows:

- <u>Volumetric video summarization based on shape descriptors</u>: Descriptors to effectively distinguish vastly different actions within human 3D volumetric data require specific invariances, including scale, pose, orientation, and position. However, unlike tasks associated with standard 3D shape retrieval, it is necessary to account for non-rigid transformations. To date, TCD have investigated the use of spherical harmonic descriptors (Kazhdan, 2003) for the purpose of segmenting volumetric videos depicting human actions. This process facilitates the extraction of keyframes from the video sequence.
- <u>Automatic Rigging:</u> Extracting human pose from a volumetric video frame requires aligning the captured model with a standard rig (Bhatnagar, 2020). Networks trained for this task should not only learn how to fit a random pose to the canonical pose but also encode non-rigid deformations, which can be used for summarising human volumetric video by capturing overall body movements and shape changes.
- <u>Pose annotations:</u> *Posecodes* (Delmas, 2022) provide a means to represent combined joint pose information in a 3D human model, capturing angles, distances, and relative positions between different body parts. The data obtained from *posecodes* will undergo thresholding and statistical analysis to generate higher-level textual descriptors.
- <u>Action semantics:</u> Exploration into the concept of a joint embedding space connecting text information and body part poses should allow processing of posecodes to obtain action semantics (Liu, 2023).

The next stages of work will enable several outputs of relevance. Firstly, the keyframe extraction component, which was completed in December 2023, identifies characteristic keyframes in the volumetric video sequence and enables visual summarization of volumetric video into descriptive animated thumbnails that efficiently depict the key elements in the sequence. This alone could allow creators to efficiently browse large databases of volumetric video assets, which, due to their size and complexity, are currently difficult to work with without specialised tools and technical skills. The current approach, which extracts keyframes as illustrated in Figure 12, will be further improved for robustness and tested with a wider range of datasets by January 2024.

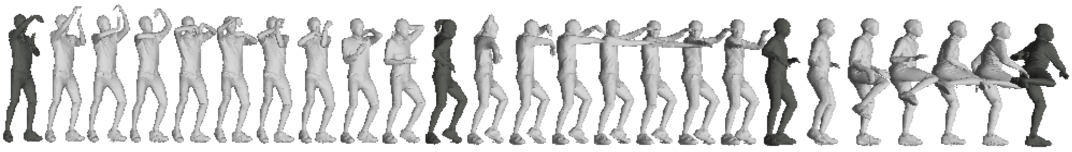


Figure 12. Extracted keyframes (highlighted in grey) from a volumetric video sequence

Later stages of the pipeline will allow semantic summarization to provide textual descriptors for the volumetric video that can be used by a traditional search engine to retrieve relevant volumetric videos for a particular XR experience, as required by creators. Furthermore, semantic annotation of keyframe segments will allow more fine-grained search and also the ability to trim and extract subsets of larger volumetric datasets for specific user needs. Finally, high-level meta-data (such as the spatial or temporal bounds of the volumetric video, anchor points and interaction points) extracted using the semantic information, will allow easier integration of volumetric video to more traditional 3D assets used in XR and animation.

4. Media Retrieval Components

The purpose of the media retrieval components is to provide a search or query mechanism over the collected content items on the basis of the exposure of their metadata descriptions such that specific content items can be found for a given purpose. The component may return: (1) metadata about the content item (e.g. identifier for a document stored in the webLyzard repository), (2) the content item itself (the file, if online it may be referenced by an URL) or also (3) a modified content item based on its planned usage (e.g. a summarisation of a longer text or video).

The first case is covered by a Web based dashboard and Search API provided by the partner webLyzard. All content items collected and analysed by MOD's content mirrors and data processing pipeline (respectively), which includes CERTH's video analysis service, are stored as metadata documents in the webLyzard document repository. These documents can then be searched and browsed using various data visualisations, allowing users to select the desired content items.

Regarding the second case, a reference to the location of the content item, e.g. an URL, may be derived from the metadata documents or another source (e.g. TCD's directory of volumetric videos). The retrieval of the original content item is usually managed by the standard tools available to resolve the given content location, e.g. content at an URL may be accessed via a Web browser or through different software libraries in various programming languages.

The webLyzard dashboard and its visualisations of the content metadata, as well as the search API, cover these two cases and will be described in D2.2. Therefore the focus on this section is on the third case, which is the need to access a modified version of a content item for the purposes of immersive content creation.

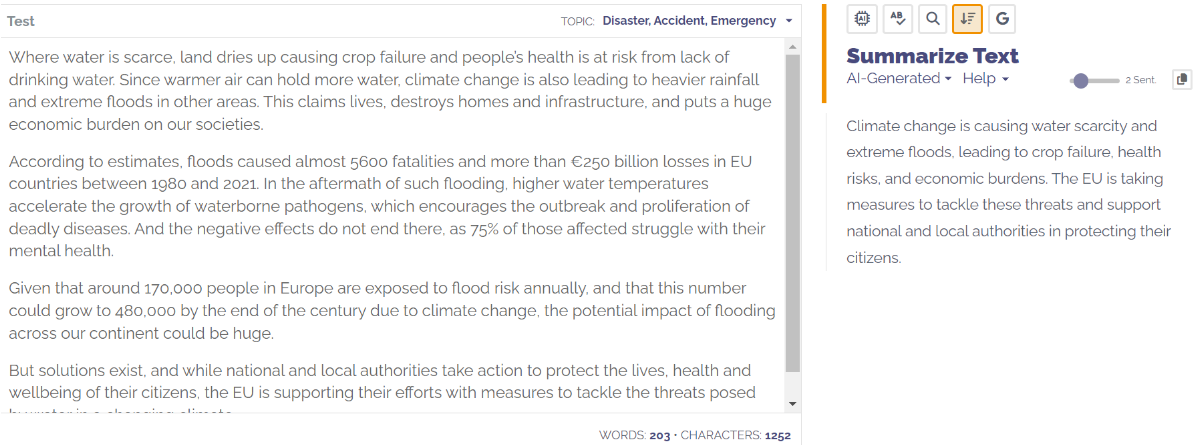
4.1. Textual Summarisation

MOD and WLT have worked on a text summarization component which can capture the most important parts of a longer text and reduce it to a shorter, more compact description which does not lose the key meaning. Prompt engineering was used to find the optimal means of soliciting a shortened text summary from Large Language Models (LLMs). Apart from reducing the total length (number of sentences) of the input text to a specified, smaller, number of sentences (generally between 1 and 10), our criteria for success was that this summary captures the essential information contained within the larger text.

While NLP algorithms could be used for *extractive summarization*, which means extracting the most important sentences and groups of words from the text, results typically lose legibility as sentences (or groups of words) are put together which were originally not together, and the loss of other words in the original text which provided necessary context might reduce their understandability. With deep learning techniques such as sequence-to-sequence models (where the model trained on sets of words also retains information about the order in which they occur), a new possibility has emerged which is *abstractive summarization*, the generation of new textual content to express the same thing as the original text (Clark, 2019). Given the ability of the model to rephrase the text in its output, it is possible to avoid the use of words which have lost important context through the summarization, and produce more natural and understandable summaries. The current generation of pretrained language models on which we will build our text summarization component are based on the Transformers architecture (Vaswani, 2019) which has been demonstrated to be effective on picking up on both syntactic cues (prepositions, co-references) and semantic cues (entities, relations).

Our implementation of abstractive summarization takes as input the full text of an online news article, and the intended output is a short summary of a few sentences - even just one - which captures the essential information about that news item. The current crop of Large Language Models (LLMs) are all based on the GPT architecture (Generative Pre-trained Transformer), which is effective for abstractive summarization tasks. We initiated our tests using the OpenAI API for ChatGPT 3.5 Turbo, testing various prompts that indicate we wanted a summary in one or a few sentences of the longer text in the input that would still capture the most important aspects of the news story. We now continue these tests using ChatGPT 4.0.

Since the use of ChatGPT would mean dependence on an API hosted by an external organisation (OpenAI) where access and pricing could change at any time, we have begun to load and test smaller LLMs within our own infrastructure. Currently best results (through human expert evaluation of the output summaries) could be seen with Mistral7B and Zephyr7B (as the names suggest, both have around 7 billion parameters). A practical use case of the text summarization is as part of the Storypact text editor tool (see Figure 13, screenshot below). Once we shift to a locally deployed LLM fine-tuned to do our abstract summarization, we will also make an API available.



TOPIC: Disaster, Accident, Emergency

Where water is scarce, land dries up causing crop failure and people's health is at risk from lack of drinking water. Since warmer air can hold more water, climate change is also leading to heavier rainfall and extreme floods in other areas. This claims lives, destroys homes and infrastructure, and puts a huge economic burden on our societies.

According to estimates, floods caused almost 5600 fatalities and more than €250 billion losses in EU countries between 1980 and 2021. In the aftermath of such flooding, higher water temperatures accelerate the growth of waterborne pathogens, which encourages the outbreak and proliferation of deadly diseases. And the negative effects do not end there, as 75% of those affected struggle with their mental health.

Given that around 170,000 people in Europe are exposed to flood risk annually, and that this number could grow to 480,000 by the end of the century due to climate change, the potential impact of flooding across our continent could be huge.

But solutions exist, and while national and local authorities take action to protect the lives, health and wellbeing of their citizens, the EU is supporting their efforts with measures to tackle the threats posed by water in a changing climate.

WORDS: 203 • CHARACTERS: 1252

2 Sent.

Climate change is causing water scarcity and extreme floods, leading to crop failure, health risks, and economic burdens. The EU is taking measures to tackle these threats and support national and local authorities in protecting their citizens.

Figure 13. Text summarization in the Storypact editor. A 203 word news article is summarized in 2 sentences.

4.2. Video Summarisation

4.2.1 Problem Statement

The scope of this task is to support the generation of visual summaries of 360 and traditional 2D videos. The developed method segments the video and selects the subshots which are most significant to the video content. Two methods are integrated in the analysis REST service: one dedicated to video summarization and the other for extracting thumbnails from it. The 360 summarization method has yet to be incorporated into the analysis REST service.

4.2.2 State of the Art Survey

Various approaches have been introduced to automate traditional 2D-video summarization, and the current state of the art is represented by methods utilising deep network architectures. One of the first approaches for video summarization was to model the variable-range temporal dependence among frames and learn how to estimate their importance according to ground-truth annotations. For this, some methods utilise structures of RNNs (Recurrent Neural Networks). Other few video summarization methods learn the frames or fragments importance by modelling the spatiotemporal structure of the video. For this, a couple of works use convolutional LSTMs in combination with typical CNN-based deep representations.

360-degree video summarization is a less-explored field**.** One of the early attempts to offer a more natural-looking NFOV (Natural Field of View) video that focuses on the interesting events of a panoramic 360 video, was made by (Su, 2016). Their method, known as AutoCam, learns a discriminative model of human-captured NFOV Web videos and employs this model to identify candidate view-points. Then, it uses dynamic programming to stitch them together through optimal human-like camera motions and create a new NFOV presentation of the 360 video. (Yu, 2018), addressed the problem of 360 video highlight detection, after defining NFOV segments. Their method uses a trainable deep ranking model which produces a spherical score map of composition per video segment and determines which view can be considered as a highlight via a sliding window kernel. Based on the composition score map, their method performs spatial summarization by finding out the best NFOV subshot per 360 video segment, and temporal summarization by selecting the N top-ranked NFOV subshots as a highlight for the entire 360 video. A similar approach was proposed by (Lee, 2018), for story-based summarization of 360 videos. It uses a trainable deep ranking network to score NFOV region proposals cropped from the input 360 video. Then, it performs temporal summarization using a memory network that models the correlation between past and future information (video subshots), based on the assumption that the parts of a story-based summary share a common story-line.

4.2.3 Approach

We propose an approach that considers both the spatial and the temporal dimension of the 360 video and takes into account several different events that might take place in parallel in order to form a summary of the video content. An overview of the proposed approach is given in Figure 14. The produced 2D video from Section 3.3 is processed by a video summarization method which makes estimates about the importance of each frame of the video and forms the video summary. The summarization task is performed using a variant of the CA-SUM (Apostolidis, 2022). This method includes an attention mechanism capable of focusing on specific regions of the attention matrix, improving estimations of their importance by considering the uniqueness and diversity of relevant video frames. The utilised variant of CA-SUM has been trained by taking into account also the estimated saliency score of each frame of the video. The saliency score is determined by computing the mean average of the salient object within the frame. Those scores, that are extracted by the 2D production algorithm (Section 3.2), are used to weight the extracted representations of the visual content of these frames. Therefore, the utilized video summarization model is trained based on a set of representations that incorporate information about both the visual content and the saliency of each video frame. At the output, this model produces a set of frame-level importance scores. By treating various sections of the 2D video as separate video fragments, importance scores for each fragment are calculated by averaging the importance scores of the frames within that fragment. These fragment-level scores are then used to select the key-fragments given a target summary length L, by solving the Knapsack problem.

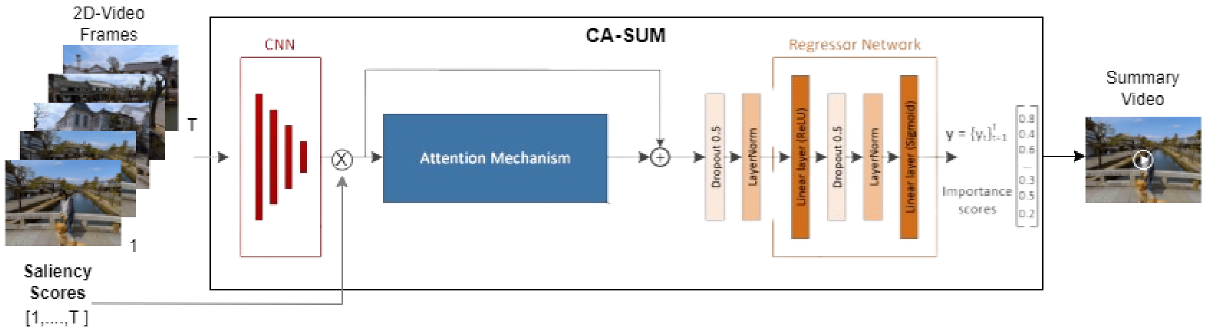


Figure 14: An overview of the approach method for the 360 summarization. The 2D-Video Frames and the saliency scores produced by our approach in section 3.3.

For the analysis REST service summarization (Figure 15), the method of (Apostolidis (a), 2021) is used, the PGL-SUM model, which introduces absolute positional encoding into multi-head self-attention mechanisms for video summarization. This integration allows our network architecture to comprehend the video frames' dependencies at various levels, capturing both the entire sequence and finer segments. By learning the importance of different parts of the video and incorporating temporal order modeling, PGL-SUM generates concise and coherent video summaries. For the thumbnail selection, a novel method introduced by (Apostolidis, 2021b) employs a blend of adversarial and reinforcement learning. It utilizes an adversarially-trained discriminator to estimate keyframe representativeness, combined with aesthetic quality measurements to form a reinforcement learning-based training pipeline.

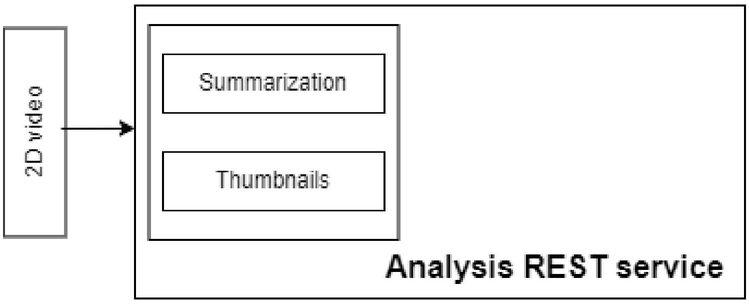


Figure 15: The overview of the analysis service for summarization and thumbnails

4.2.4 Implementation Details and Use

To train the developed variant of the CA-SUM video summarization model, deep representations were obtained for sampled frames of the videos using GoogleNet (Szegedy, 2015). The block size of the concentrated attention mechanism was set equal to 20. Finally, for training the employed CA-SUM model, we used 100 2D videos that were produced and scored in terms of frame-level saliency, based on the 2D Video Production algorithm described in Section 3.3.3. These videos consist of 57 videos of the VR-EyeTracking, 37 videos of the Sports-360, and 6 videos of the Salient360! dataset. Training was performed for 400 epochs in a full-batch mode for 80 videos of the formed dataset, and 20 videos were used for testing.

To improve the summarization method on 360 videos, our focus is on creating a specialised dataset. Using the 360 video frames and the ground truth saliency maps of the VR-EyeTracking dataset, a set of 2D videos is generated by the 2D Video Production algorithm described in Section 3.3.3, including the salient events of the 360 videos. To annotate these 2D videos for the needs of video summarization, an interactive annotation tool has been developed to allow users to indicate their preferences regarding the importance of each video fragment. This interactive tool will be the base of a web-based platform, enabling users to interact with the final summary video, considering user's needs and feedback.

The summarization and thumbnails methods are part of the analysis service described in 3.2. The summarization method is tested on two publicly available datasets, namely TVSum (Song, 2015) and SumMe (Gugli, 2014) including 50 and 25 videos, respectively. The thumbnails method is pretrained on OVP and Youtube datasets (Avila, 2011), including 50 videos from each dataset.

4.2.5 Results

Figure 16 shows a frame-based overview of an automatically-produced 2D video from a 360-degree video, generated after selecting one frame per shot (shots are directly related to the defined sub-volumes of the 360-degree video, as described in Section 3.3.3), and presents the produced summaries for two video summarization methods. The summary at the top was created by the trained variant of the CA-SUM method using videos of the VR-EyeTracking, Sports-360, Salient360! datasets and after taking into account the computed saliency scores for the frames of these videos. The summary at the bottom was created by a pre-trained model of CA-SUM using videos from the TVSum dataset for video summarization. As can be seen, the trained variant produces a more complete and representative video summary after including parts of the video showing the gathered people in a square with a Christmas tree, the people right in front of the avenue taking some photos, and the illuminated shopping mall behind the avenue. On the contrary, the pre-trained CA-SUM model focuses more on fragments of the video showing the avenue and ignores salient video parts presenting the shopping mall. Hence, taking into account the saliency of the visual content is important when summarising 360-degree videos, as it allows the production of more representative and thus useful video summaries.

For the summarization and thumbnail methods that are included in the analysis REST service, an output JSON is shown in Listing 4. The JSON file contains the outputs as URL links and must be downloaded independently, if needed.

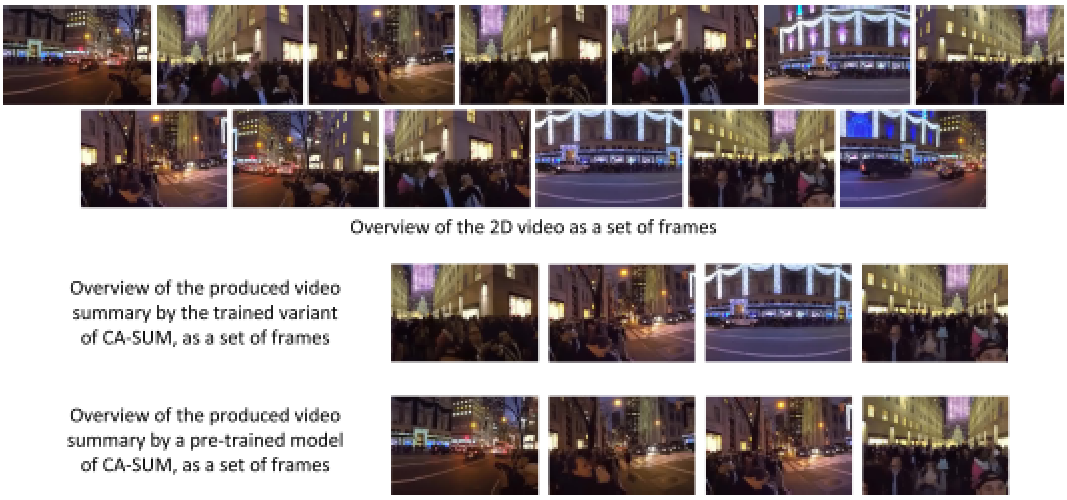

Overview of the 2D video as a set of frames
Overview of the produced video summary by the trained variant of CA-SUM, as a set of frames
Overview of the produced video summary by a pre-trained model of CA-SUM, as a set of frames

Figure 16: A frame-based overview (using one representative frame per shot), and example summaries by the summarization method on the produced 2D video for a 360 video of the VR-EyeTracking dataset

```
Reply   Status code: 200
        {   "expires_at": "2020-07-15 10:51:13.304437",
            "framerate": 25.000,
            "generated_at": "2020-07-01 10:51:13.304426",
            "generated_by": "https://transmixr-idt.iti.gr",
        "summary": "https://transmixr-idt.iti.gr:443/summary/becdf9cc16b2358b59b1cc12d938e580",
        "thumbnails": [
                " https://transmixr-idt.iti.gr: 443/thumbnail/716797cdedd2d4f577f3503/1",
                " https://transmixr-idt.iti.gr: 443/thumbnail/716797cdedd2d4f577f3503/2",
                 …
                " https://transmixr-idt.iti.gr: 443/thumbnail/716797cdedd2d4f577f3503/5",
            ],
        }
```

Listing 4: A JSON file output of the analysis REST service described in Section 3.2.

4.3. Volumetric Video Discovery

4.3.1 Overview and State of the Art

Volumetric video datasets, which represent the 3D shape and appearance (texture or colours) of captured subjects over an interval of time, are relatively new with currently little standardisation of formats. Although they are often conflated with similar formats such as 3D point-clouds (or RGB-D data), 3D models or traditional 3D animations, the unique characteristic value of volumetric video datasets is that they represent the authentic recognizable appearance and motion of specific subjects (typically human performers) captured from the real world using 3D imaging or multi-camera systems.

Recent advancements in multi-view capture systems have led to a substantial expansion of human volumetric video data (Işık, 2023; Pagés, 2021). However, because this specific data format has only been widely discussed within the last 5 years or so, there is a dearth of publicly available databases compared to other media formats. Beyond a handful of sources of data captured for scientific research, most volumetric video datasets are a result of bespoke captures for a specific production or use-case and often not shared widely. A small number of commercial volumetric video databases exist (e.g. renderpeople.com), but in such cases the data is of relatively generic humans for general animation purposes, and therefore such databases under-utilise the unique potential of volumetric video for capturing recognizable real world individuals.

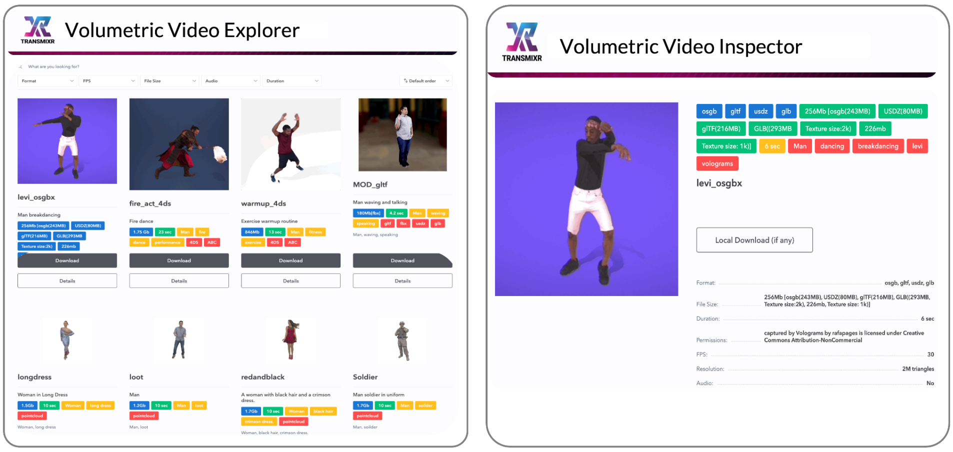
As a result of this, there is limited previous work in volumetric video discovery and retrieval beyond manually annotating models in a small number of isolated databases. Recently, there has been some work dedicated to automatically annotating human action sequences using pre-trained models (Delmas, 2022; Lin, 2023). These efforts predominantly focus on Skinned Multi-Person Linear (SMPL) sequences (Loper, 2015), rather than raw human meshes. The annotation pipeline involves several challenges including automatic rigging and bridging the relationship between pose parameters and textual descriptions. Additionally, some works concentrate on per-frame textual pose descriptions rather than deriving the action semantics of an entire sequence (Delmas, 2022).

Furthermore, unlike more established media formats such as 2D and 360 video, volumetric videos, due to their non-standard nature and complexity cannot be natively viewed in web browsers and commonly available digital media tools without specialised plug-ins/code, nor can they yet be automatically parsed by existing search tools. Amongst our goals within TRANSMIXR is to broaden the accessibility of volumetric video directly to XR creators and also to standard automated indexing tools.

4.3.2 Approach and Initial Results

Leveraging the work on Volumetric Video understanding, discussed in Section 3.4 of this report, TCD has developed an approach for extracting visual summaries of volumetric video in more traditional and more easily interoperable graphical formats, such as animated gif thumbnails or a compact videos, which can be more easily integrated into existing search engines. This will allow users to browse quickly through potential datasets of interest without the need for a specific volumetric video viewer. In addition, the next stages of TCD's work in the next 3 months will deal with automatic generation of text-based semantic annotations of volumetric video that will allow meaningful indexing of such datasets by standard web-search tools. When a search result of potential interest is selected, users would be directed to a specific landing page containing (a) metadata and natural language descriptions (b) download links in multiple formats and (c) an inline interactive 3D preview using a proprietary or custom viewer for volumetric video, based, for instance, on WebGL.

An online proof-of-concept of such a system that was developed by TCD is shown in Figure 17. In this prototype, the metadata and visual summaries are manually populated in the envisaged formats that will be required by the final system, but, in the next stages of the project, these will be replaced by data that is automatically extracted based on the work discussed in Section 3.4 above.



levi_osgbx

Man breakdancing

256Mb [osgb(243MB) USDZ(80MB) glTF(216MB) GLB((293MB Texture size:2k) 226mb

fire_act_4ds

Fire dance

1.75 Gb 23 sec Man fire dance performance 4DS ABC

warmup_4ds

Exercise warmup routine

846Mb 13 sec Man fitness exercise 4DS ABC

MOD_gltf

Man waving and talking

180Mb[fbx] 4.2 sec Man waving speaking gltf fbx usdz glb

Man, waving, speaking

longdress

Woman in Long Dress

1.5Gb 10 sec Woman long dress pointcloud

Woman, long dress

loot

Man

1.2Gb 10 sec Man loot pointcloud

Man, loot

redandblack

A woman with black hair and a crimson dress.

1.7Gb 10 sec Woman black hair crimson dress. pointcloud

Woman, black hair, crimson dress.

Soldier

Man soldier in uniform

1.7Gb 10 sec Man soilder pointcloud

Man, soilder

osgb gltf usdz glb 256Mb [osgb(243MB) USDZ(80MB) glTF(216MB) GLB((293MB Texture size:2k) 226mb Texture size: 1k)] 6 sec Man dancing breakdancing levi volograms

levi_osgbx

Format: osgb, gltf, usdz, glb
File Size: 256Mb [osgb(243MB), USDZ(80MB), glTF(216MB), GLB((293MB, Texture size:2k), 226mb, Texture size: 1k)]
Duration: 6 sec
Permissions: captured by Volograms by rafapages is licensed under Creative Commons Attribution-NonCommercial
FPS: 30
Resolution: 2M triangles
Audio: No

Figure 17. Left: prototype web-interface for browsing a volumetric videos database compiled by TCD comprising visual summaries using animated gif thumbnails. Right: individual landing page to inspect detailed information and in future will allow users to interactively inspect the data in 3D, such as by rotating the view.

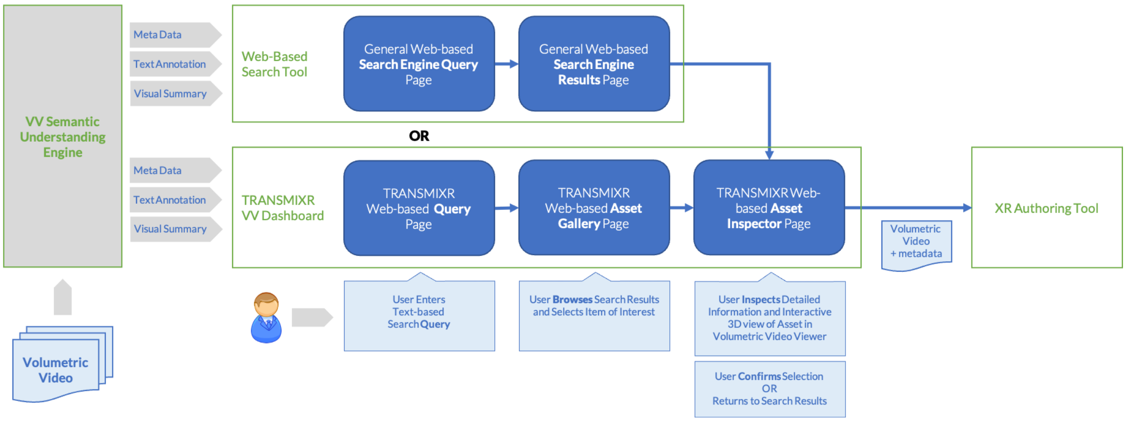

VV Semantic Understanding Engine
Meta Data
Text Annotation
Visual Summary
Web-Based Search Tool
General Web-based Search Engine Query Page
General Web-based Search Engine Results Page
OR
Meta Data
Text Annotation
Visual Summary
TRANSMIXR VV Dashboard
TRANSMIXR Web-based Query Page
TRANSMIXR Web-based Asset Gallery Page
TRANSMIXR Web-based Asset Inspector Page
Volumetric Video + metadata
XR Authoring Tool
Volumetric Video
User Enters Text-based Search Query
User Browses Search Results and Selects Item of Interest
User Inspects Detailed Information and Interactive 3D view of Asset in Volumetric Video Viewer
User Confirms Selection OR Returns to Search Results

Figure 18. Envisaged flowchart of the volumetric video asset search and retrieval system

Furthermore, the same techniques for summarization and annotation should also benefit search and query tasks in more specialised media asset databases of the future with dedicated support for volumetric videos, which are expected to become more widespread as this form of XR data matures. Figure 18 shows a flowchart of the envisaged volumetric video asset discovery system.

5. Conclusions

We have presented the initial set of components in the TRANSMIXR project for media ingestion, understanding and summarisation. Together, this forms the “content understanding” part of the TRANSMIXR architecture, where multimodal content items from different sources can be browsed, understood and selected based on their relevance to a topic, keyword or search term. All of the content assets are then available to the “content creation” stage where immersive and interactive social XR experiences can be created. Besides supporting the “core” data format of text – any other non-textual content may be converted to text through the act of annotation – TRANSMIXR also focuses on advancing research work in the retrieval, analysis and description of video assets, whether linear 2D video, 360 degree video or volumetric video. Through the three pilots of the TRANSMIXR project, emerging and changing requirements for “content understanding” may occur and our ongoing development of all components - testing state of the art and developing beyond state of the art solutions - will also take this into account in the subsequent, iterative, updates that will be made. A later deliverable will describe the final versions of each component, refined or extended throughout their use in the pilots.

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

References

Weichselbraun, A., Kuntschik, P., & Brasoveanu, A. M. (2019). Name variants for improving entity discovery and linking. In 2nd Conference on Language, Data and Knowledge (LDK 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. arXiv preprint arXiv:1906.04341.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. NeurIPS 2017: 5998-6008. arXiv:1706.03762.

Gkalelis, Nikolaos, Dimitrios Daskalakis, and Vasileios Mezaris. "ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network." IEEE Access 10 (2022): 108797-108816.

Jiang, Yu-Gang, et al. "High-level event recognition in unconstrained videos." International journal of multimedia information retrieval 2 (2013): 73-101.

Vieira, Joelton Cezar, et al. "Low-cost CNN for automatic violence recognition on embedded system." IEEE Access 10 (2022): 25190-25202.

Oh, Sangmin, et al. "A large-scale benchmark dataset for event recognition in surveillance video." CVPR 2011. IEEE, 2011.

Herath, Samitha, Mehrtash Harandi, and Fatih Porikli. "Going deeper into action recognition: A survey." Image and vision computing 60 (2017): 4-21.

Yao, Guangle, Tao Lei, and Jiandan Zhong. "A review of convolutional-neural-network-based action recognition." Pattern Recognition Letters 118 (2019): 14-22.

Ma, Shugao, et al. "Do less and achieve more: Training cnns for action recognition utilizing action images from the web." Pattern Recognition 68 (2017): 334-345.

Daskalakis, Dimitrios, Nikolaos Gkalelis, and Vasileios Mezaris. "Masked Feature Modelling: Feature Masking for the Unsupervised Pre-training of a Graph Attention Network Block for Bottom-up Video Event Recognition." arXiv preprint arXiv:2308.12673 (2023).

Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018, June). Gaze Prediction in Dynamic 360° Immersive Videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Wenhao, et al. "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023a.

Wu, Wenhao, Zhun Sun, and Wanli Ouyang. "Revisiting classifier: Transferring vision-language models for video recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 37. No. 3. 2023b.

Liu, Wei, et al. "Ssd: Single shot multibox detector." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).
Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

Mou, Lichao, et al. "Era: A data set and deep learning benchmark for event recognition in aerial videos [software and data sets]." IEEE Geoscience and Remote Sensing Magazine 8.4 (2020): 125-133.

Nam, Junhyun, et al. "Learning from failure: De-biasing classifier from biased classifier." Advances in Neural Information Processing Systems 33 (2020): 20673-20684.

Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 (2012).

He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." arXiv preprint arXiv:1605.07146 (2016).

Gutíerrez, J., David, E.J., Coutrot, A., Da Silva, M.P., Callet, P.L.: Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360° contents. In: 2018 10th Int. Conf. on Quality of Multimedia Experience (QoMEX). pp. 1–3 (2018). https://doi.org/10.1109/QoMEX.2018.8463369

Bernal-Berdun, E., Martin, D., Gutierrez, D., Masia, B.: SST-Sal: A spherical spatio-temporal approach for saliency prediction in 360 videos. Computers &Graphics 106, 200–209 (2022). https://doi.org/10.1016/j.cag.2022.06.002

C. H. Vo, J. -C. Chiang, D. H. Le, T. T. A. Nguyen and T. V. Pham, "Saliency Prediction for 360-degree Video," *2020 5th International Conference on Green Technology and Sustainable Development (GTSD)*, Ho Chi Minh City, Vietnam, 2020, pp. 442-448, doi: 10.1109/GTSD50082.2020.9303135

Campos, V., Jou, B., & Giró-i-Nieto, X. (04 2016). From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction. *Image and Vision Computing*, *65*. doi:10.1016/j.imavis.2017.01.011

Fei-Fei, O. R. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*, 211-252. doi:10.1007/s11263-015-0816-y

D. Galanopoulos, V. Mezaris, "Are All Combinations Equal? Combining Textual and Visual Features with Multiple Space Learning for Text-Based Video Retrieval", Proc. ECCV 2022 Workshop on AI for Creative Video Editing and Understanding (CVEU), Springer LNCS vol. 13804, pp. 627–643, Oct. 2022

Gkalelis, N., Mezaris, V. (2020). Subclass Deep Neural Networks: Re-enabling Neglected Classes in Deep Network Training for Multimedia Classification. In: Ro, Y., *et al.* MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science(), vol 11961. Springer, Cham. https://doi.org/10.1007/978-3-030-37731-1_19

N. Gkalelis, D. Daskalakis and V. Mezaris, "ViGAT: Bottom-Up Event Recognition and Explanation in Video Using Factorized Graph Attention Network," in IEEE Access, vol. 10, pp. 108797-108816, 2022, doi: 10.1109/ACCESS.2022.3213652.

Gygli, Michael. "Ridiculously fast shot boundary detection with fully convolutional neural networks." 2018 International Conference on Content-Based Multimedia Indexing (CBMI). IEEE, 2018.

Hassanien, Ahmed & Mohamed, Elgharib & Selim, Ahmed & Hefeeda, Mohamed & Matusik, Wojciech. (2017). Large-scale, Fast and Accurate Shot Boundary Detection through Spatio-temporal Convolutional Neural Networks.

J. Islam and Y. Zhang, "Visual Sentiment Analysis for Social Images Using Transfer Learning Approach," *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, Atlanta, GA, USA, 2016, pp. 124-130, doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.29.

Lee, S., Sung, J., Yu, Y., & Kim, G. (2018). A Memory Network Approach for Story-Based Temporal Summarization of 360° Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1410–1419.

Li, Z. G. (2021). *Temporal-attentive Covariance Pooling Networks for Video Recognition*. Retrieved from arXiv:2110.14381

Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "Simple vs complex temporal recurrences for video saliency prediction." BMVC 2019.

Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2017. Query and Keyframe Representations for Ad-hoc Video Search. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR '17). Association for Computing Machinery, New York, NY, USA, 407–411. https://doi.org/10.1145/3078971.3079041

Mondal, J., Kundu, M.K., Das, S. *et al.* Video shot boundary detection using multiscale geometric analysis of nsct and least squares support vector machine. *Multimed Tools Appl* 77, 8139–8161 (2018). https://doi.org/10.1007/s11042-017-4707-9.

Nguyen, Anh & Yan, Zhisheng & Nahrstedt, Klara. (2018). Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction. 1190-1198. 10.1145/3240508.3240669.

Pournaras, A., Gkalelis, N., Galanopoulos, D., & Mezaris, V. (2021, December 13). Combining Multiple Deep-learning-based Image Features for Visual Sentiment Analysis. MediaEval 2021 Workshop. https://doi.org/10.5281/zenodo.6655366

M. Qiao, M. Xu, Z. Wang and A. Borji, "Viewport-Dependent Saliency Prediction in 360° Video," in IEEE Transactions on Multimedia, vol. 23, pp. 748-760, 2021, doi: 10.1109/TMM.2020.2987682.

Ronneberger, O. a. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. a. Navab (Ed.), *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015"* (pp. 234--241). Springer International Publishing.

Sarafianos, N., Xu, X., & Kakadiaris, I. A. (2018, September). Deep Imbalanced Attribute Classification using Visual Attention Aggregation. *Proceedings of the European Conference on Computer Vision (ECCV)*

A. Habibian, T. Mensink and C. G. M. Snoek, "Video2vec Embeddings Recognize Events When Examples Are Scarce," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 2089-2103, 1 Oct. 2017, doi: 10.1109/TPAMI.2016.2627563.

Souček, Tomáš, Jaroslav Moravec, and Jakub Lokoč. "Transnet: A deep network for fast detection of common shot transitions." *arXiv preprint arXiv:1906.03363* (2019).

Y. Xu, Z. Zhang and S. Gao, "Spherical DNNs and Their Applications in 360◦ Images and Videos," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 7235-7252, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3100259..

This project has been funded by the European Union as part of the Horizon Europe Framework Program (HORIZON), under the grant agreement 101070109.

Y. Zhang, F. D. (2020). “Saliency prediction network for 360◦ videos,” . *IEEE Journal of Selected Topics in Signal Processing,*, pp. vol. 14, no. 1, pp. 27–37.

Yasser Dahou, M. T. (2020, Nov 20). ATSal: An Attention Based Architecture for Saliency Prediction in 360 Videos. *Computer Vision and Pattern Recognition*.

Yue, K. a. (2018). Compact Generalized Non-Local Network. *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 6511–6520). Montr\'{e}al, Canada: Curran Associates Inc.

Zhang, Z., Xu, Y., Yu, J., & Gao, S. (2018, September). Saliency Detection in 360° Videos. *The European Conference on Computer Vision (ECCV)*.

E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, "Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames", Proc. of the 2022 Int. Conf. on Multimedia Retrieval (ICMR ’22), June 2022, Newark, NJ, USA.

Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems 167 (2019), 26–37. DOI:https://doi.org/10.1016/j.knosys. 2019.01.019

Chandrasekaran G, Antoanela N, Andrei G, Monica C, Hemanth J. Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. *Applied Sciences*. 2022; 12(3):1030. https://doi.org/10.3390/app12031030

Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., Luo, P.: Bridging video-text retrieval with multiple choice questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16167–16176 (2022)(15)

Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)(2)

Syed, Arslan & Aldhahri, Eman & Iqbal, Muhammad & Ali, Abid & Muthanna, Ammar & Jamil, Harun & Jamil, Faisal. (2022). Intelligent 3D Network Protocol for Multimedia Data Classification using Deep Learning. 10.20944/preprints202207.0056.v1.

Savran Kızıltepe, Rukiye and Gan, John Q and Escobar, Juan José (2023) *A novel keyframe extraction method for video classification using deep neural networks.* Neural Computing and Applications, 35 (34). pp. 24513-24524. DOI https://doi.org/10.1007/s00521-021-06322-x

Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, Ji Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 2238-2247

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10146–10155, 2020.

Haoqian Wu, Keyu Chen, Yanan Luo, Ruizhi Qiao, Bo Ren, Haozhe Liu, Weicheng Xie, Linlin Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14021-14030

Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of the Second Int.Conf. on Knowledge Discovery and Data Mining. p. 226–231. AAAI Press (1996)

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. 2019. What Do Different Evaluation Metrics Tell Us About Saliency Models? IEEE Trans. Pattern Anal. Mach. Intell. 41, 3 (March 2019), 740–757.

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G.: Saliency in vr: How do people explore virtual environments? IEEE Transactions on Visualization and Computer Graphics 24(4), 1633–1642 (2018).

Su, Y.C., Jayaraman, D., Grauman, K.: Pano2vid: Automatic cinematography for watching 360 videos. In: Proc. of the Asian Conf. on Computer Vision (ACCV) (2016)

Yu, Y., Lee, S., Na, J., Kang, J., Kim, G.: A Deep Ranking Model for Spatio-Temporal Highlight Detection From a 360 Video. In: Proc. of the 2018 AAAI Conf. on Artificial Intelligence (2018)

Apostolidis, E., Balaouras, G., Mezaris, V., & Patras, I. (2021, December). Combining Global and Local Attention with Positional Encoding for Video Summarization. *2021 IEEE International Symposium on Multimedia (ISM)*, 226–234. (a)

E. Apostolidis, E. Adamantidou, V. Mezaris, I. Patras, "Combining Adversarial and Reinforcement Learning for Video Thumbnail Selection", ACM Int. Conf. on Multimedia Retrieval (ICMR), Taipei, Taiwan, Nov. 2021. DOI:10.1145/3460426.3463630 (b)

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE/CVF

Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594

Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: TVSum: Summarizing web videos using titles. In: 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 5179–5187 (2015). https://doi.org/10.1109/CVPR.2015.7299154

M. Gygli et al., "Creating Summaries from User Videos," in Europ. Conf. on Comp. Vision 2014. Cham: Springer Int. Publishing, 2014, pp. 505–520.

S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de A. Araújo. 2011. VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. Pattern Recognition Letters 32, 1 (Jan. 2011), 56–68.

J. Xu, T. Mei, T. Yao and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 5288-5296, doi: 10.1109/CVPR.2016.571.

Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., & Luo, J. (2016). TGIF: A New Dataset and Benchmark on Animated GIF Description. *arXiv [Cs.CV]*. Retrieved from http://arxiv.org/abs/1604.02748

Wang, X., Wu, J., et al.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proc. of the IEEE Int. Conf. on Computer Vision. pp. 4581–4591 (2019)

Bhatnagar, B. L., Cristian S., Christian T., and Gerard P-M. "Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration." Advances in Neural Information Processing Systems, vol. 33, pp. 12909-12922, 2022.

Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. "Activitynet: A large-scale video benchmark for human activity understanding." Proceedings of the ieee conference on computer vision and pattern recognition, pp. 961-970, 2015.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291-7299, 2019.

Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F. and Rogez, G. "PoseScript: 3D human poses from natural language." European Conference on Computer Vision, vol. 2022, pp. 346-362, 2022.

Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., & Jacobs, D. "A search engine for 3D models." ACM Transactions on Graphics (TOG), vol. 22(1), pp. 83-105, 2003.

Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., & Nießner, M. “Humanrf: High-fidelity neural radiance fields for humans in motion.” arXiv preprint, vol. 2305.06356, 2023.

Kazhdan, M., Funkhouser, T. and Rusinkiewicz, S. “Rotation invariant spherical harmonic representation of 3 d shape descriptors.” Symposium on geometry processing, vol. 6, 2003.

Liu, X., Li, Y-L., Zeng, A., Zhou, Z., You, Y. and Lu, C. “Bridging the Gap between Human Motion and Action Semantics via Kinematic Phrases.” arXiv preprint, vol. 2310.04189, 2023.

Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., & Zhang, L. “Motion-x: A large-scale 3d expressive whole-body human motion dataset.” *arXiv preprint, vol. 2307.00818*, 2023.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. “SMPL: A skinned multi-person linear model.” Seminal Graphics Papers: Pushing the Boundaries, vol. 2, pp. 851-866, 2015.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., & Black, M. J. “AMASS: Archive of motion capture as surface shapes.” Proceedings of the IEEE/CVF international conference on computer vision, pp. 5442-5451, 2019.

Moynihan, M., Ruano, S., & Smolic, A. “Autonomous tracking for volumetric video sequences.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1660-1669, 2021.

Novotni, M., & Klein, R. “3D Zernike descriptors for content based shape retrieval.” Proceedings of the eighth ACM symposium on Solid modeling and applications, pp. 216-225, 2003.

Pagés, R., Amplianitis, K., Ondrej, J., Zerman, E., & Smolic, A. “Volograms & V-SENSE Volumetric Video Dataset.” *ISO/IEC JTC1/SC29/WG07 MPEG2021/m56767*, 2021.

Xie, X., Bhatnagar, B. L., & Pons-Moll, G. “Chore: Contact, human and object reconstruction from a single rgb image.” European Conference on Computer Vision, pp. 125-145, 2022.

Xie, X., Bhatnagar, B. L., & Pons-Moll, G. “Visibility aware human-object interaction tracking from single rgb camera.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4757-4768, 2023.

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., & Wang, Y. “Motionbert: A unified perspective on learning human motion representations.” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15085-15099, 2023.