Video Analysis via Keyframe Extraction in a Classroom Recording Environment

María Fernanda Hernández Montes

November 2022

Abstract

Text-Video Retrieval is a field of great relevance for projects that aim to be resolved through the implementation of Machine Learning principles for the analysis of video scenes captured in any environment. For the research program *Intelligent Video Detection of Suspicious Behavior and Interpretation Using Deep Learning* the main challenge is to capture videos during virtual learning sessions and interpret them in a way that allows the algorithm to process their frames and, given a text query, retrieve videos that correspond to the given description via Natural Language Processing. The expected result is a ranked list of candidate videos scored via document retrieval metrics.

In this context, video abstraction has been developed as a process of presenting an abstract view and comprehensible analysis of a full-length video within the shortest period of time, thus by extracting a representative frame to produce a condensed version of the original video. The main idea revolves around managing a large amount of video data by selecting a set of representative frames as to simplify the process of video analysis and processing [24]. This leads us to defining our desired outcome, which is to have an ideal video analysis system where one could visually identify classroom states.

One of video retrieval major downsides is having to work with massive amounts of data, which will vary according to the nature of the videos being processed and the context that determines the information of interest. This is why the solution model is based on a first phase of keyframe extraction, which will permit the simplification of the dataset to be analyzed by creating a summarized set of relevant frames that will describe pivot moments from the original video, hence allowing to reduce processing costs without losing relevant information.

Through this work, we aim to propose a machine learning model that allows users to retrieve information of interest based upon a classroom recording video, thus through the implementation of a Keyframe Extraction approach that permits improvement in the accuracy of the results obtained while reducing and simplifying the set of data to be worked with.

1 Introduction

Recently, as a society, we have been facing different challenges that came as a consequence of the global pandemic caused by the spreading of the SARS-CoV-2 -COVID 19- virus, and among them we find the adaptation of education in an online environment, which has caused education providers to become increasingly aware of the diversity of their learners, since, as stated by Orlando and Attard, "teaching with technology is not a one size fits all approach, as it depends on the types of technology in use at the time and also the curriculum content being taught" [14].

As a result, we have seen the way different barriers to participation were being experienced by students, and how they became particularly evident in collaborative learning tasks. On account of this, many schools

have tried to come across with different solutions that could help improve the engagement students show during their online lessons, most of which have stayed even after the sanitary crisis mentioned above.

One of these efforts to improve and adapt online and face-to-face lessons according to the students' behavior is the actual research project, which aims to develop a video retrieval model that allows to receive an input, process and retrieve information from videos recorded during in-person learning sessions. The aimed result is a model that allows the extraction of candidate frames, along with their respective timestamps, obtained from a full-length video, according to queries made by the user through concise concepts of interest.

As we already know, digital video is an actual popular storage and exchange medium, since it allows users to own both visual and sound information as a representation of the human environment. Now, when it comes to creating a database with video archives, it becomes necessary to take many factors into account, such as the optimization of big data pools, the kind of query that will be worked with, the way the information has to be presented and which metrics will be used to measure success [6].

When comparison between images retrieved from videos is based upon analytical features, results tend to be quite ambiguous; while, on the other hand, features based upon content bring up more precise coincidences upon the query. Now, when it comes to working with entire videos, processing them becomes harder and costs tend to rise, this being the main reason why it becomes important to store the data effectively and allow convenient browsing [28]

Video retrieval consists of recovering videos from a space based upon a text-query made by a user, which turns it into a challenging part of a project, since this involves analyzing the actual context of the video and finding correlations between the semantic assessment and measurable concepts extracted by automatic techniques, as discussed by Hauptmann, Christel and Yan [15]. This plays a highly relevant role in multi-modal understanding, which makes reference to work with different modalities, such as text, speech and images [13].

When it comes to Video Retrieval, as a subset of Information Retrieval, there are methods that purely rely on text and make use of metadata to describe video content. Nevertheless, there are works that rely on different visual features that can be measured through each frame, such as color, texture and shapes, implementing different layers of abstraction. Given the fact that this particular problem works with both visual and textual information, the main goal must be approached through Multimodal Machine Learning (MML), which implies an embedded space where video and text are comparable.

As mentioned before, projects of this kind have already been developed, such as the thesis work presented by Portillo Quintero in Two-fold Approach for Video Retrieval: Semantic Vectors to Guide Neural Network Training and Video Representation Approximation Via Language-Image Models, which focuses on a Dual Encoder architecture to funnel video and text information through independent Neural Networks [22].

The model analyzed through this work consists of two points. One which will be the aggregated value to works previously done on Video Retrieval systems, and the second point establishes the way in which it impacts and brings value to the scope of the problem. This project works with the idea of the classroom students as a group instead of individual beings, which allows us to establish different scenarios that will be set for retrieval, which are: *There is a high level of movement in the classroom, the classroom presents an low level of movement.*

Some works that have laid foundation on video retrieval are exposed through models such as the Contrastive Language-Image Pre-Training model [23], which pretends to generate a tool that is able to describe images through text at a zero-shot glance, and its versions focused on migrating that to the description of videos, as will be explained later on this document. These works have been an inspiration for the present project, since they are distinguished because of the way they process information in order to avoid losing important

relations not only between sentences and videos, but between frames and words [4].

As it has been mentioned above, our system aims to implement a Keyframe Extraction approach in order to manage large amounts of video data. This is done by selecting unique sets of representative frames while preserving the essential activities of the original video, which will help us increase the simplicity in video processing afterwards [24]. Through this implementation we pretend to generate a pre-processing stage that simplifies the information to be analyzed by the image classification phase of the model.

As an outcome, we pretend to generate a model that allows the client to introduce certain queries according to the state of the classroom they are interested in and therefore obtain videos with timestamps that indicate the time interval during which the activity of interest is visually shown.

Throughout this document, the proposal and proper justification for it will be stated. First, the problem and necessary context to understand the present work will be defined in Section 2, covering elements relevant to our research. Afterwards, in Section 3, we'll focus on the explanation of the elements that take place during the development of the model; this will carry the document through the description, in Section 4, of the experiments that took place in order to finally conclude the research in the final part 5.

2 Background and Related Work

In order to establish a path to approach the proposed model, many other works have been revised, which have laid foundations for the different phases that will be boarded during this work. These works and the respective background for our project will be discussed during this section, such as Machine Learning fundamentals for Information Retrieval, Computer Vision, Keyframe Extraction, among other topics.

2.1 Machine Learning

Over the past few years, the interest on Machine Learning has increased, becoming one of the mainstays of technology, given the fact that the amounts of data we process as a society in different life areas is becoming larger and more vital for diverse processes.

Machine Learning is a branch of artificial intelligence, which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy . This area of AI has different classifications, where different statistical methods are used to train the algorithms in charge of classifying and predicting information with a large variety of purposes. UC Berkeley breaks down the learning process of a Machine Learning algorithm into three main parts [11]:

- **Decision process,** based on the input data that will be labeled, which leads to a pattern generation that will eventually become criteria for said process.
- **Error function,** which allows the evaluation of the prediction accuracy of the implemented model.
- **Model optimization process,** that is made possible given the fact that the weights that are implemented for the decision process can be adjusted in order to increase the accuracy of the model until a threshold has been met.

Now, as it has been mentioned before, Machine Learning covers a wide range of Artificial Intelligence approaches, where we can find supervised, weakly supervised, unsupervised and reinforcement learning.

2.1.1 Image Classification

Methods such as classification and clustering have become increasingly important, since they allow us to categorize instances according to their content. More specifically, classification is a data mining technicality that specifies classes according to a set of data, in order to help predict and analyze information.

There are many different methods implemented through reinforcement learning, unsupervised and supervised learning, among others, to learn a model that allows the machine to predict a result according to certain input given by the user. If we aim to talk about classification algorithms, there are many approaches that can be implemented, and one of the simplest is binary classification.

As we know, the process that surrounds classification is based on the description of features, such as colors and textures, of diverse items, in order to spot the similarities between them and create classes that are characterized by said features. On these terms, image classification consists of a database that contains predefined patterns that are compared to a specific image that will, therefore, be classified in a suitable category [3].

There are different approaches to this task, through supervised and unsupervised classification. In Supervised Classification, some previously known pixels or images are grouped and labelled in order to power the classifier with data to classify other images. On the other hand, Unsupervised Classification uses clustering to group pixels according to their properties, with the main difference that it does not count with available previously trained pixels [3].

For this task, Babu et. Al. identify three particular steps: Image Acquisition, Image Pre-Processing and Feature Extraction, as mentioned below [3]:

1. *Image acquisition.* Recollection of images for future processing.
2. *Image pre-processing.* Implementation of techniques such as image transformation, noise removal and morphological corrections.
3. *Feature extraction.* Extortion of important characteristics of the images.

Now, in order to approach this task, convolutional layers are commonly used for recognition, in which the features of an image are selected. In order to determine the image values, a filter must pass through the image -the usual window size for these filters is 3x3 or 5x5 pixels. The filter starts this recognition process from the upper left corner of the image, moving one pixel to the side, through the entire picture. This process implemented through Convolutional Neural Networks can be observed in the example depicted on figure 1 [27].

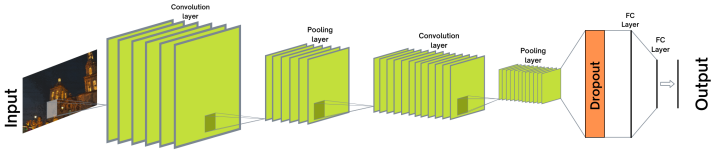

Figure 1: Convolutional Neural Network architecture.

2.2 Information Retrieval

According to Manning, Raghavan and Schütze, information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). Now, when it comes to the different kinds of data to be retrieved, we could be talking about many types of formats, but for this work we will focus on Video Retrieval (VR) [18].

Content video retrieval is an approach that aims to facilitate the process of searching and browsing through large collections of frames, from which the full-length video is composed [21]. For this task, many different approaches have been explored, mainly based on low level visual properties extracted from each frame.

2.3 Computer Vision

As stated by the British Machine Vision Association and Society for Pattern Recognition, computer vision is concerned with the automatic extraction, analysis and understanding of useful information from a single image or a sequence of images. It involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding [5].

This involves the development of mathematical techniques that permit the recovering of features that describe objects in the three-dimensional space, giving users the possibility of processing this information in order to obtain a vast variety of results. One of the areas that computer vision allows us to approach is image processing, based on the goal of pre-processing images and converting them into a suitable form of further analysis.

2.3.1 Keyframe extraction

Now, when it comes to video summarization, it is clear to say it generates an easier context for video browsing, indexing and retrieval, since it permits the simplification of the database to be worked with. This need comes from the difficulty of analyzing and processing large amounts of data, preventing an effective video management scheme [24].

To understand this concept, we must first lay out the foundations on how a video is structured, going from the video itself to scenes that capture a specific sequence of events, to shots that are the smallest unit of temporal visual information that contains a sequence of interrelated frames -which is itself the tiniest partition to be worked with [24].

Video summarization, as stated by Sadiq et al, "helps us save and improve the storage capacity of the video contents efficiently, decreasing the amount of data required when streaming or downloading video contents from the web". To achieve this, there exist a wide variety of techniques, such as *Edge-Detection Technique* and *Clustering-Based Technique*.

For **Edge-Detection Technique**, as it has been stated, the purpose of edge detection is to significantly reduce the amount of data in an image, while preserving the structural properties to be used for further image processing . On this context, *Canny Edge Detection* is implemented as an algorithm that optimally detects edges through the frames of the video, based upon a specific list of criteria [1]:

- Edges occurring in images should not be missed, and where there are no edges, there shouldn't be a response either,
- The distance between the edge pixels found by the algorithm and the actual edge should be as small as possible.

This leads us to a series of steps that must be followed in order to eliminate noise and therefore obtain a better and mor precise result:

1. The noise has to be filtered out by first determining a simple mask and then implementing the Gaussian filter using standard convolution methods.
2. Once the image has been smoothed and the noise has been eliminated, we have to find the edge strength. This is done by taking the gradient (directional change in the intensity of a frame) of the image, and then implementing the Sobel operator*, which uses a pair of 3x3 convolution masks (one estimates the gradient in the x axis and the other estimates the gradient in the y axis).
3. Since we used the Sobel operator [1], the edge direction is much easier to find. We should add that whenever the gradient in the x-axis equals zero, the edge direction has to be equal to 90 or 0 degrees, depending on the y direction and its value of the gradient; when the gradient in the y direction equals zero, the edge direction becomes 0 degrees. Edge direction can be found using the mathematical expression in Eq. 1

$$\theta = \arctan\left(Gy/Gx\right) \tag{1}$$

When the edge direction is known, it has to be related to a direction that can be traced in an image.

Now, when a **Clustering-Based Technique** is implemented, the algorithm in use consists of a unsupervised learning approach. It is based upon the idea of partitioning frames within a video file into different clusters when they have similar visual contents. When it comes to each cluster, the frame nearest to the center of the candidate cluster is the one extracted as a keyframe [24]. This reduces complexity by discarding similar frames with few to no difference.

[1] The Sobel operator computes the gradient of the intensity in an image for each pixel, which leads to us obtaining how softly or abruptly an image changes for each pixel analyzed and how probable it is for it to represent an edge

2.4 Video Retrieval

On big video databases, usefulness is determined by the efficiency to locate certain videos according to the information that the user is interested in. When it comes to video retrieval, it is of main interest that video indexing is analogous to text document indexing, where we perform certain analysis in order to decompose it into paragraphs, lines and words. Therefore, to facilitate speed and accuracy in content access to video data we should segment these archives into clips and shots to create a table of contents based on the most important information to be retrieved from said video database.

2.4.1 CLIP

As it has been mentioned before, there are diverse approaches and models that have been developed in order to work with a solution on text-based queries for video retrieval, one of which is the Contrastive Language-Image Pre-Training, better known as CLIP. [4]

CLIP is a multimodal model that can be instructed in natural language to predict most relevant text snippets given an image, having similar capabilities to zero-shot prediction. It consists of two sub-models called encoders:

- A text encoder that will embed text into mathematical space.
- An image encoder that will embed images into mathematical space.

When a set of images goes through an encoder, the output will be a series of numbers, same thing happens to text. What we pretend is for the series of numbers for the text to be very close to the series of numbers for the image. In order to obtain this "closeness", a mathematical approach called *Cosine Similarity* is used.

Cosine Similarity is a metric used to measure the cosine of the angle between two vectors projected in a multi-dimensional space. The smaller the angle, the higher the cosine similarity. This helps us determine how similar two documents are, irrespective of their size.

Contrastive Language-Image Pre-training model has achieved impressive performance learning representations from large databases and this makes it a very interesting alternative, though it was mainly designed for image retrieval, which takes us to the next developed approach.

2.4.2 CLIP2Video

Now, when it comes to understanding and describing a whole video instead of just an image with a zero-shot approach, it becomes quite relevant to use a more macroscopic point of view. This means that, instead of just relating videos and sentences, CLIP2Video pretends to break this problem down into two independent sub-problems: spatial representation of multi-modal image-text training, and temporal relationships of video frames and video-language [12].

This tool was designed in order to transfer the image-language pre-training model mentioned above to video-text retrieval, which adjusts better to the area of interest for our research. On this case, the pre-trained image-language model is simplified as a two-stage framework with co-learning of image-text, and after that the temporal relations between video frames and video-text are enhanced.

On different words, what CLIP2Video pretends is to search for certain videos by mapping both video and text into joint embedding space, just as we mentioned before, when we referred to CLIP model, but this time adjusted to a video dataset.

2.4.3 CLIP4CLIP

Finally, as we know, video clip retrieval will be the corner stone of this research, as it plays an essential role in multimodal analysis. The Contrastive Language-Image Pre-training model mentioned above lays an important ground on the way that visual concepts are able to learn from web collected image-text datasets, although it is still necessary to adapt it to a multimodal dataset where our main archives are composed by both image and sound through videos.

CLIP4CLIP model aimed to exploit the pre-trained CLIP model to design a similarity calculation module that can be used for video databases. The approach consulted as inspiration for the guidance of this work investigates three similarity calculation approaches: parameter-free type, sequential type, and tight type [17].

The parameter-free approach fuses the video representation without new parameters, and as well as the sequential type similarity calculation approach adopts two separate branches for video and text representation independently to calculate cosine similarity. While, on the other hand, the tight type similarity calculation module implements the transformer model for multimodal interaction and further calculates the similarity via a linear projection [17].

The experiments made are developed reusing similar parameters from the CLIP model -the position embedding in sequential type and tight type, which are initialized by repeating the position embedding from CLIP's text encoder. Meanwhile, the transformer encoder is initialized by the corresponding layers' weight of the pretrained CLIP's image encoder, and for the decoder a 3-layer transformer model is adopted for generation.

What differences this approach from the one proposed by the Contrastive Language-Image Pre-training model is the testing of the three mechanisms of similarity calculation mentioned above, the post-training of the CLIP model on a noisy large-scale video-text dataset to learn a better joint semantic space, and the adaptation of the State of the Art for both video retrieval and captioning tasks on different vastly used datasets.

2.5 Related Works

Prior to starting with the retrieval video, it is key to consider the factors that are involved at the time of creation or selection of the database that will be used, since it defines in large part the result of the implemented model. Among these factors are, the type of query, the results and the measure for success.

In the event that the comparison of images retrieved from videos is based on analytical features, the results tend to be ambiguous; on the other hand, content-based features provide more precise matches with the query, but the results are based on image processing that can only recognize a small amount of objects [26].

Queries in text form allow us to remove ambiguity in the query and work only with content-based functions, for which there is limited uncertainty. While the presentation of the result of a query is usually visual, textual presentations still provide more specific information and are useful when presenting very large collections of

data.

There are two main categories of work described, the first one consists of first extracting key frames for video data and then using some image retrieval technique to obtain video data indirectly, however it can lead to a time consuming data acquisition process [19]. On the other hand, the second category focuses on the emergence of objects in the recovery process, although it is an expensive technique, especially if the trajectories of the objects are tracked [2].

The principal work that has been revised as inspiration for the video retrieval approach being analyzed is the one presented by Jesús Andrés Portillo, *Two-fold Approach for Video Retrieval: Semantic Vectors to Guide Neural Network Training and Video Representation Approximation Via Language-Image Models*, in which he presents a method to bridge the semantic gap between video and text by employing a Multimodal Machine Learning model that is capable of mapping multiple types of information among themselves [22].

This work aims to measure similarity between a query and items from a database, transforming both modalities through Neural Networks as to learn a joint transformation between them both. It proposes the implementation of sentence embedders that generate vectors that contain contextualized encodings of query phrases, allowing the measurement of semantic similarity between vectors. On the other hand, as for an approximation of a video-text space for Video Retrieval, Portillo proposes an already trained image-text space by extending it, using CLIP model since it is trained with 400 million datapoints [22].

In order to process the textual information, Portillo implements BERT as a text backbone model, which stands for Bidirectional Encoder Representations from Transformers and is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [9]. It implements two steps in its framework: *pre-training* and *fine-tuning*. On the first phase, the model is trained on unlabeled data over different pre-training tasks; while, on the other hand, during fine-tuning, the model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks [9].

3 Solution Approach

As it has been stated above, the main goal of this research project is to generate an algorithm that allows the retrieval of video files according to certain queries made by the users, requesting information about the status of the classroom that has been recorded.

Although there are different approaches that have been explored when it comes to Text-Based Video Retrieval and data-set pre-training methods, for the purposes of the present project two of these have been taken into account as a ground for the desired aggregated value to the state of the art model in which our research is based upon: *Video summarization* through Keyframe extraction and *Image Classification* to classify said keyframes into status of interest.

This way, the first phase implemented in order to generate a simpler context to be analyzed and processed by our model is video summarization. This approach allows us to lay a visual foundation easier to be evaluated, since it is based on the idea that a video is composed by a finite number of frames. Some of these can be selected and stored apart in a non-aleatory form to represent key moments of the shots and scenes the video is made from, as seen in figure 2.

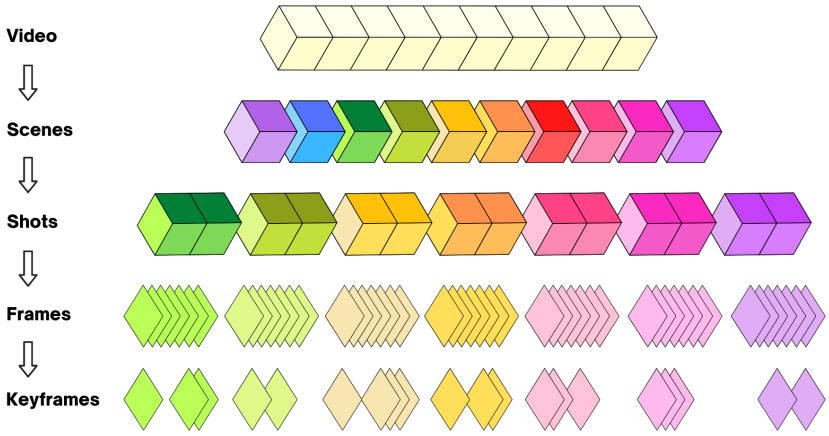

Figure 2: Video Summarization through keyframe extraction.

The approach being used is based on the processing of each frame of the video and its conversion into grayscale, in order to filter the image and make it easier to work with. In addition, a gaussian filter will allow the model to spatially reduce visual noise produced by small variations in the frame, such that it permits the detection of differences between all the frames through time, identifying the movements or elements that appear as new [8].

The Gaussian filter mentioned above is used to blur images and remove noise and detail implementing the function specified in equation x, where d is the standard deviation of the distribution -with the distribution being assumed to have a mean of 0. This leaves us with an image that has been filtered and ready to be processed [**?**].

$$G(y,x) = \frac{1}{\sqrt{2\pi\sigma}} \frac{(-(x^2+y^2))}{2\sigma^2} \tag{2}$$

After this, the keyframes extracted will be selected according to the iterative performing of a polynomial fitting in the data to detect its baseline. This implies that at every iteration, the fitting weights on the regions with peaks are reduced so that the baseline can be identified. This leaves us with the possibility to subtract said baseline from the image data and spot the indexes of the principal peaks where the images present relevant differences, as it can be seen in figure 3 [20].

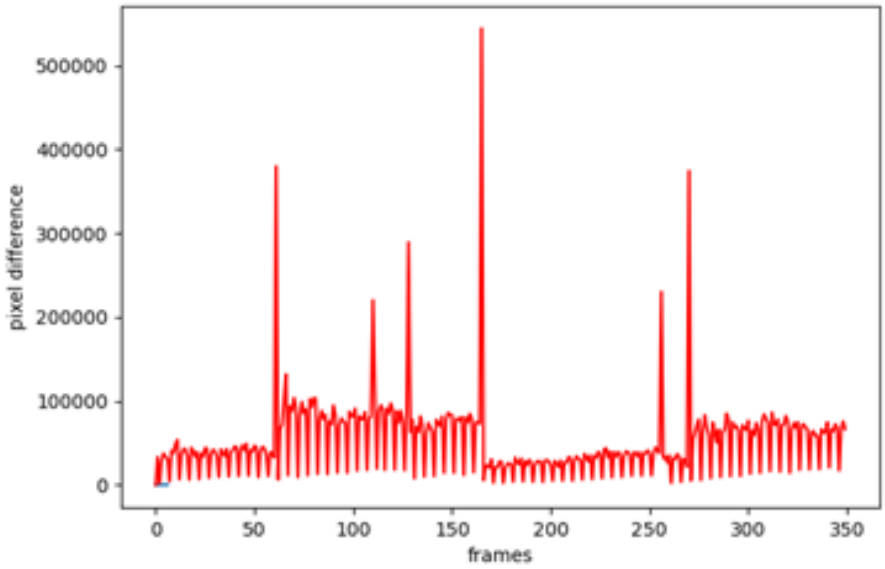

Figure 3: Pixel value differences shown between frames processed from a test video during experimentation.

It is there where the higher pixel differences are spotted that the frames are extracted as keyframes that will altogether represent the video when we go through the next phase of the model, which will classify these images according to the status of the classroom that the user has selected.

For the second phase, we implemented an Image Classificator to make use of it as an analyzer for the activity shown on the video, classifying the information according to whether a high or a low level of activity is visually perceptible. As we can see on figure 4, the Convolutional Neural Network is composed by four main 2D convolutional layers, which will allow the performing of an elementwise multiplication when the kernel slides over the input data, summing up the results into a single output pixel [25]. We include max pooling layers as well, in order to obtain a feature map that contains the most prominent features of the previous layer result. And finally, flatten and dense layers are used to collapse the spatial dimensions of the input into the channel dimension and classify the keyframe being processed [10].

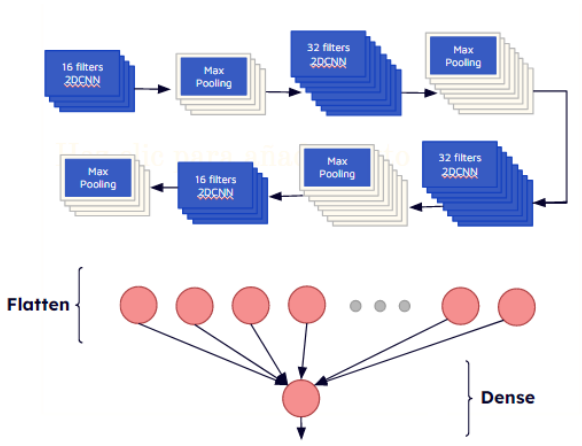

Figure 4: Convolutional Neural Network implemented for classification.

Therefore, once we have extracted the keyframes from the video given, our model will classify them according to the different states observed in the classroom. This, as it has been mentioned above, will permit the future visual analysis of the moments through the classroom recording where the two main types of activity (high level movement and low level movement) are being shown.

3.1 Datasets

For this model we firstly tried to implement VGG-16 Convolutional Neural Network, which works with ImageNet database, over about 14 million images belonging to 1000 classes. ImageNet is a large-scale ontology of images built upon the backbone of the WordNet structure; it aims to populate the majority of the 80 thousand synsets of WordNet with an average of 500-1000 clean and full resolution images [7].

ImageNet organizes the different classes of images in a densely populated semantic hierarchy, and one of the most impressive things about this dataset is its density, since it offers a vast number of categories for different things (for example, dog categories). Now, when it comes to training a Convolutional Neural Network model to perform classification on ImageNet and then adapt those features for a new task, there have been impressive results. These results are not only shown on image classification datasets, but also in action recognition, image segmentation, optical flow, among many others [16].

Nevertheless, the results obtained were not as expected, which led us to the creation of a local dataset conformed by classroom recording videos provided publicly by different schools and universities. This approach involved reducing the scope of the project, which is why we aim to improve the tools and databases implemented for future advances.

4 Experiments and Results

As we mentioned throughout the first sections of this work, there are many projects and approaches that have been implemented as a solution for Video Retrieval problems, many of which work with backbones that aim to process both text and images separately. Nevertheless, there are others that pretend to innovate the perspective from which this challenge is tackled, implementing different Machine Learning algorithms. On our counterpart, we propose a model that aims to reduce the costs of video processing and to facilitate the process of training a Convolutional Neural Network that is in charge of classifying the images that will be presented as results for the queries made by the user.

4.1 Keyframe Extraction

Just as it has been detailed along this paper, the present work is divided into two main phases, where the first one is Keyframe Extraction. This approach will allow us to reduce costs in video processing, since the solution aims to summarize the videos our model must work with, leaving us with representative frames only. As we detailed before, the Keyframe Extraction model implemented is based on a system that identifies the main peaks where the pixels throughout the images show important differences. On these terms, something provided by our model is the possibility to specify the level of summarization needed by the user, thus by giving the option to indicate the threshold used to discriminate those frames that will be extracted from the video given.

In table 1 we display the comparison in the level of abstraction shown when modifying the threshold value in one of the videos used for testing. When the threshold is higher, the number of keyframes extracted is significantly reduced. On figure 5, this is depicted by the blue horizontal graphic, that represents the number of frames surpassing the threshold, which in this case has a value of 0.15.

Threshold	0.05	0.12	0.14	0.2	0.25	0.5
Total frames per video	350					
Number of keyframes extracted	115	59	42	8	7	5
Percentage of keyframes extracted	32.86	16.86	12.0	2.29	2.0	1.43

Table 1: Level of video summarization according to different thresholds implemented during experiments

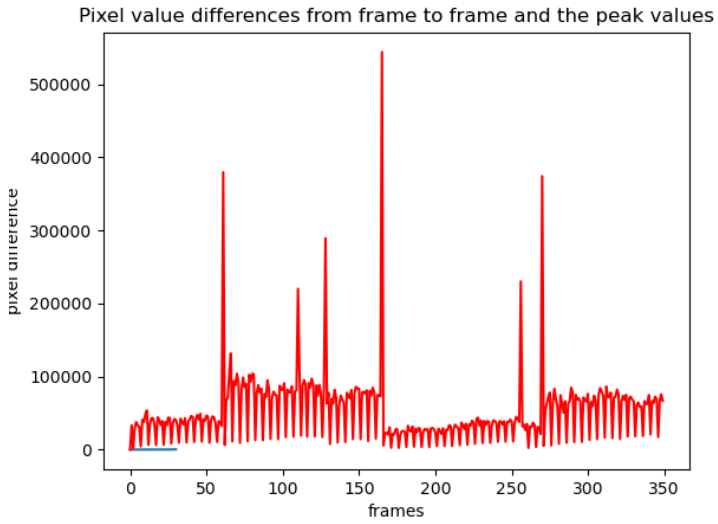

Figure 5: Pixel differences identified during the summarization of a test video, with a result of 32 keyframes extracted, 9.14 percent of total frames (depicted by blue graphic)

4.2 Image Classification

On the other hand, in order to be able to give the user certain moments where the classroom status of interest is being shown, we needed to generate a model that could classify the frames according to the activity of interest -in this case: low level movement and high level movement. This was obtained with respect to the time duration of the video, as we can observe in figure 6, showing a variation according to a classroom recording that shows us a scene of students entering the classroom -with a small time span where particularly high movement is not observed.

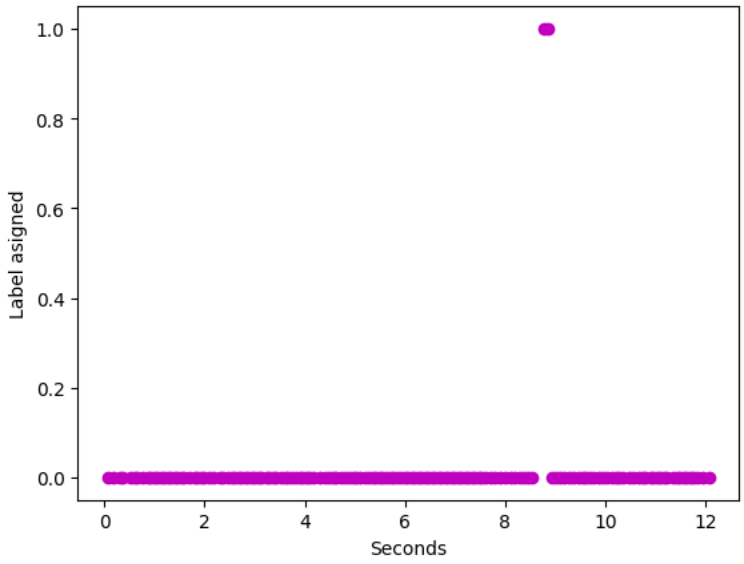

Figure 6: Results obtained from classification, where a high level movement is labeled as zero and low level movement is described by one.

4.3 Discussion

As we have observed through the past paragraphs, the model worked with two phases, which allowed us to obtain results that can be discussed both individually and through unification as final findings. For the first phase, we can observe in figure 7 that videos were successfully summarized in great proportions without losing relevant information, which means that time and processing costs were significantly reduced. On the other hand, for the second phase we are able to identify that, although an accurate analysis is visually shown according to the main activity states of interest, the final output can be considerably improved.

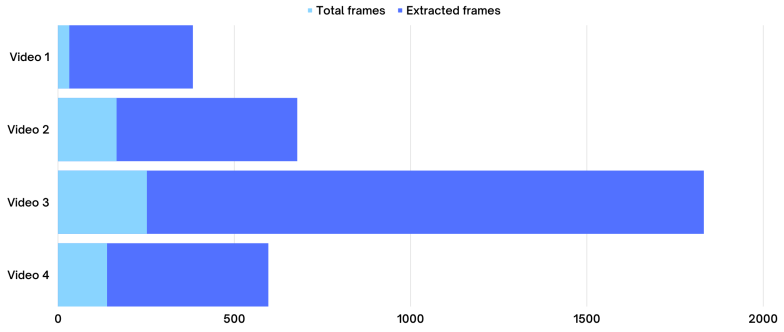

Figure 7: As we can observe, the first phase of our model allows us to summarize the video from 9 to 31 percent of the total length, thus reducing both processing times and costs

Now, according to the results that have been shown by the complete model, we consider that, although the present project facilitates the process of analyzing a classroom recording video in search of particular states

of activity, the results could be improved. This improvement could be achieved through the implementation of a Convolutional Neural Network backbone for transfer learning, as well as other technological tools for Multiclass Classification. Therefore, the current work will serve as a foundation for a project that allows video retrieval with a wider scope.

5 Conclusion and Future Work

As a brief overview of the model implemented during this work, we aimed to approach a video retrieval problem in physical classroom recordings through video summarization via keyframe extraction, and multiclass classification using transfer learning. The approach developed pretends to present a simple but adaptable solution to information retrieval not only in a classroom environment, but in any other environment that allows activity status classification.

During this work we found that, even though there are many different approaches that have been implemented in the past, through Image Classification we can propose a simpler method. Said method decomposes audiovisual information using keyframe extraction and generates sets of information that has been classified according to areas of interest. This contribution allows to reduce human time and effort in areas that require video analysis in order to identify certain situations, all of this while reducing the costs that implementing different approaches would generate.

Nevertheless, during this process we also found limitations that reduce the reaching of our model. Among them we find the implementation of reduced data sets for training, which directly impacts the performance and behavior our model shows during prediction and classification. To combat this particular situation, future steps implicate working with a bigger team in charge of labelling the visual data, so as to improve the model's results. We also consider that improvements to the present work can be obtained through the use of Convolutional Neural Network backbones for transfer learning, as well as Multiclass Classification models that will permit the coverage of a wider scope of interest states.

References

[1] 09gr820. Canny edge detection. 2009.

[2] Y Alp Aslandogan and Clement T. Yu. Techniques and systems for image and video retrieval. *IEEE transactions on Knowledge and Data Engineering*, 11(1):56–63, 1999.

[3] K. Ganapathi Babu, Dakannagari Harith Reddy, P. Divya Teja, and C. Yosepu. An overview on image classification methods in image processing. *International Journal of Current Engineering and Scientific Research*, 5:27–29, 2018.

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022.

[5] Emma Beauxis-Aussalet. Introduction to computer vision. *Digital Society School*, 2019.

[6] Saddam Bekhet and Amr Ahmed. Evaluation of similarity measures for video retrieval. *Multimedia Tools and Applications*, 79(9):6265–6278, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: a large-scale hierarchical image database. pages 248–255, 06 2009.

[8] Telecomunicación y Automática Departamento de Ingeniería Electrónica. *Reducción del Ruido en una Imagen Digital*. Universidad de Jaén, Jaén, Spain, 2005.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[10] Govinda Dumane. Introduction to convolutional neural network (cnn) using tensorflow, 2020.

[11] IBM Cloud Education. Machine learning. *IBM Cloud Learn Hub*, 2020.

[12] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

[13] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020.

[14] Jenna Gillett-Swan. The challenges of online learning: Supporting and engaging the isolated learner. *Journal of Learning Design*, 10(1):20–30, 2017.

[15] Alexander G Hauptmann, Michael G Christel, and Rong Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, 2008.

[16] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning?, 2016.

[17] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Lei Wen, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *Microsoft Research Asia*, 2021.

[18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[19] GS Naveen Kumar and VSK Reddy. Key frame extraction using rough set theory for video retrieval. In *Soft Computing and Signal Processing*, pages 751–757. Springer, 2019.

[20] Lucas Hermann Negri. *PeakUtils 1.3.3 documentation*. PeakUtils, 2020.

[21] B. V. Patel. and B. B. Meshram. Content based video retrieval. *The International Journal of Multimedia ist Applications*, 4(5), 2012.

[22] Jesús Andrés Portillo Quintero. Two-fold approach for video retrieval: Semantic vectors to guide neural network training and video representation approximation via language-image models. Master's thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey, 2021.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[24] Bashir Olaniyi Sadiq, Bilyamin Muhammad, Muhammad Nasir Abdullahi, Gabriel Onuh, Ali Abdulhakeem Muhammed, and Adeogun Emmanuel Babatunde. Keyframe extraction techniques: A review. *Journal of Electrical Engineering*, 19(3):54–60, 2020.

[25] Irhum Shafkat. Intuitively understanding convolutions for deep learning, 2018.

[26] Newton Spolaor, Huei Diana Lee, Weber Shoity Resende Takaki, Leandro Augusto Ensina, Claudio Saddy Rodrigues Coy, and Feng Chung Wu. A systematic review on content-based video retrieval. *Engineering Applications of Artificial Intelligence*, 90:103557, 2020.

[27] Andrii O. Tarasenko, Yuriy V. Yakimov, and Vladimir N. Soloviev. Convolutional neural networks for image classification. pages 101–114, 12 2019.

[28] Qiang Zhang, Shao-Pei Yu, Dong-Sheng Zhou, and Xiao-Peng Wei. An efficient method of key-frame extraction based on a cluster algorithm. *Journal of Human Kinetics*, 39:5–13, 2013.