An Overall Video Summarizer

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction

The ever-increasing rate of videos surfacing on the internet has been astonishing. YouTube, With the ease it offers, has been the best and a popular platform for people all across the globe to upload their videos online on diverse and versatile topics. More than 500 hours of video were uploaded to YouTube per minute as of June 2022. This comes up to almost 30,000 hours of fresh information uploaded per hour.

A vast array of videos is available on the internet on a wide spectrum of topics from education and entertainment to self-help guides and social activism. Videos are one of the most effective means to convey and comprehend information. It can be considered as one of the most effective tools for learning as they can visually and audibly convey concepts and ideas.

The covid-19 pandemic has boosted the growth of such online contents to the next level. Videos had become the major source of information across domains during the pandemic period allowing people to know things happening world-wide staying at home. Videos played major role especially in the field of education not hindering the learning of students.

While Videos play extremely important roles in our life these days, it is even important that the user gets access to the best video in search for without wasting much of their time. Often the videos available can waste our precious time as they may not necessarily contain the information we are looking for.

Students may waste enormous time in searching for the best lecturing videos. The video thumbnails may sometimes be misleading and may waste our time making us watch it without extracting the required information which makes it necessary to adopt a methodology which can help people knowing the context of the videos before watching the videos so that the viewer has an idea and understanding of the video before devoting his precious time into it. Beyond the realm of education, video summarizing has a number of advantages and is frequently used in industries like media, law, and business, where it may be used to glean important information from voluminous video material.

A concise video summary aids users in aligning their needs with the content, benefiting individuals with diverse preferences. An audio version of the summary is valuable for those who prefer not to read, ensuring accessibility for individuals with visual or auditory impairments.

So, a method is imposed that provides a concise summary of the input video. Video’s Audio transcript is summarized also the content extracted from the video frames is summarized and an overall summary is obtained. Our proposed approach deals mainly with three parts i.e., obtaining the audio transcript of the video, extracting the text present in the images of video’s (video frames) at regular intervals, obtaining the overall summary of both transcripts both as text and as audio.

Although there may be many existing models for the problem there is less focus on the overall summary of the video i.e., an approach that summarizes not just the audio transcript but also the video frames i.e., images of the video. It is very important to even extract the useful, significant information from images of videos along with the audio part of the video.

Our Proposed approach concentrates on this part. The video id/URL is fed as input and inbuilt python modules such as YouTube-Transcript API for videos with subtitles otherwise the videos’ audio is downloaded and is converted to wav format (using ffmpeg or any other pythons’ built-in modules) and Automatic Speech Recognition (ASR) is used for audio to text conversion. The video is split into multiple frames and the text is extracted from the frame using some Optical Character Recognition (OCR) Technique. The transcripts obtained from both i.e., from audio and images is summarized independently and compared to get an overall summary.

For Summarization, which is basically of two types extractive and abstractive. The texts extracted from images and audio is summarized by both the methods. While our major focus remains on improvising the already existing methods of abstractive summarization. Also, a hybrid summarization method can be adopted by combining both extractive and abstractive summarization. For conversion of summarized text to audio format a python module ‘pyttsx3’ can be used.

The Summarized part is evaluated Using ROUGE ((Recall-oriented understudy for Gisting evaluation) which helps in measuring the overlap or similarity between the reference text and the summarized text.

# Literature Survey

The field of video summarization has been extensively studied. However, as it remains challenging to construct an accurate summary, it continues to be a study area. The videos can be summarized most effectively using machine learning and natural language processing techniques, based on earlier approaches.

*A. Audio Extraction from the video and its transcript*

Researches in the field of audio to text have focused on developing models that can accurately transcribe the audio using various methods such as Speech recognition and Natural Language Processing (NLP). In few of the Earlier Approaches the Audio Transcript was obtained by YouTube Transcript API which is a python API used to obtain the transcript for the given YouTube Video. While, in some other approaches the transcript was obtained by downloading the audio of the Video and converting it into wav (mp3 or any other format to wav) format and then speech to text conversion could be done with the help of a Transformers’ such as Hugging face Transformers’ Automatic speech Recognition (ASR). Automatic Speech Recognition (ASR) converts speech signals to text by mapping a sequence of Audio inputs to text outputs.

Also, Pythons’ several built in modules have been used in earlier researches such as MoviePy which allows to work with videos such as video editing, processing etc., pydub which can be used for converting the audio from any format to wav format (in context of audio transcription) as the wav format is useful for transcription.

Apart from these there has been significant researches in the field of Audio transcription such as using convolutional transformers with ASR’s (Automatic speech recognition) and also using Mel-Frequency Cepstral Coefficient (MFCC) feature extraction technique and Minimum Distance Classifier, Support Vector Machine (SVM) for speech classification is used in one of the earlier research projects.

Long short-term memory (LSTM) can also be used for tasks involving text, speech, textual sequences etc.

*B. Text extraction from video frames(image)*

The previous researches in the field of image to text have tried developing a model to accurately transcribe the text present in the audio using techniques such as Optical Character Recognition (OCR). Optical Character Recognition is a technology used to convert scanned images of text or handwritten documents into editable and searchable text. It automates the process of extracting characters and words from images.

Also, pythons’ built in modules such as tesseract, which is an open-source Optical Character Recognition (OCR) developed by Google. It can be used to extract text from images, scanned documents or printed materials and converts it into text.

Some other researches include, using Maximally Stable Extremal Regions (MSER) which is an image processing technique used for detecting regions in images (in our case text) and Fuzzy means Clustering technique (FCM).

Multi-frame text tracking is a technique used in computer vision to track and maintain the continuity of text content across multiple consecutive frames in a video sequence. By analysing text present in previous frames, deep neural networks can enhance the accuracy and robustness of text recognition in real-time or dynamic scenarios.

Convolutional Neural Networks (CNN) can be used for image and video processing tasks. It can be used for extracting visual features from images containing text. It is useful in finding regions of interest in images (video frames in our case) where text is present which makes the text recognition more accurate.

OpenCV which is an open-source computer vision and machine learning library that provides the functionalities to perform Optical Character Recognition (OCR) , enabling extraction of text from images.

*C. Text (Transcript) Summarization*

From the earlier researches made it is known that there are two types of summarization methods i.e., extractive and abstractive. While, Extractive summarization deals with extracting the important parts of the given text to be summarized to create a condensed version of the original text. It tokenizes the input text into sentences or phrases, calculates features for each which may include metrics like Term Frequency- Inverse Document Frequency (TF-IDF), word frequency, position in text etc. It then assigns the scores to each sentence based on the computed features using approaches such as TF-IDF scores, cosine similarity, graph-based algorithms or many other machine learning models. It the selects the top-ranking sentences and arrange them in a coherent and meaningful way to form an extractive summary. Whereas, Abstractive summary is a method of summarization which summarizes a given piece of text in a way that is not constrained to selecting existing sentences or phrases from the original text. Instead, it may also contain new phrases and sentences to convey the main ideas present in the text. Similar to extractive summarization input text is tokenised converting the words, phrases or sentences into a numerical representation suitable for the model which may include word embeddings such as Word2Vec, GloVe which are components of Natural language Processing (NLP). This can further be trained using machine learning or deep learning model such as sequence-to-sequence models, transformer-based models (e.g., BERT) or RNN’s (Recurrent Neural Network) to generate the summary.

Some of the previous works show an yet another type of summarization method called as hybrid model which forms a blend of both extractive and abstractive summarization methods. One such example is the split and merge method which uses both the basic types of summarizations to obtain the best effects by utilising both. It is particularly useful when the length of the summarization text exceeds a certain level when dealing with Pythons’ inbuilt modules, As Pythons’ inbuilt modules do not support their functionality after a specific level.

Python has several built-in modules for both extractive and abstractive summarization techniques.

SpaCy for extractive summarization which is an open-source Natural Language Processing library which can be used for performing various NLP tasks including tokenization, sentence segmentation and more.

Google T5 and Pegasus on the other hand are used for abstractive summarization which can be used to obtain a short synopsis of the entire Video. Google T5 is a transformer-based model designed for NLP tasks. Pegasus is another transformer-based model developed by Google Research, primarily designed for abstractive text summarization. However, both these models have input text size limitation i.e., 512 for Google T5 and 1024 for Pegasus thereby introducing another type of summarization in the previous researches. These limitations have invented some hybrid methods in the earlier publications which first finds the extractive summary of a given text and then reduces it further by finding it abstractive summary to get an overall concise summary.

A transformer-based summarization pipeline is structured methodology integrating various components which can be used for text summarization. It does not autonomously generate summaries but relies on a transformer model, a sophisticated neural network architecture. The pipeline processes tokenized input data and applies learned patterns to generate condensed representations. It is a tool that aids in summarizing and extracting essential information from input text.

An yet another approached way is using BERT, a general-purpose language model, although not specifically designed for developing summaries it can be used in summarization pipeline to generate summaries. It has more complex pipeline and some addition layer for summarization.

Artificial Neural Network (ANN), a computational model inspired by human’s brain structure and Multi-Layer Perceptron (MLP) is a type of artificial neural network. ANN and MLP’s can be utilised to build models for text summarization. ANN’s support both extractive and abstractive summarization.

From all the earlier observations and researches made it can be outlined that tremendous work has been done in the context of audio transcription, extraction of text from images, video processing and text summarization. There has not been significant work in Overall video summarization. Hence, our approach will work towards an overall video summarizer i.e., an overall and concise text summary of the video’s audio and the video frames using appropriate machine learning and deep learning algorithms.

* econdary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not

consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**