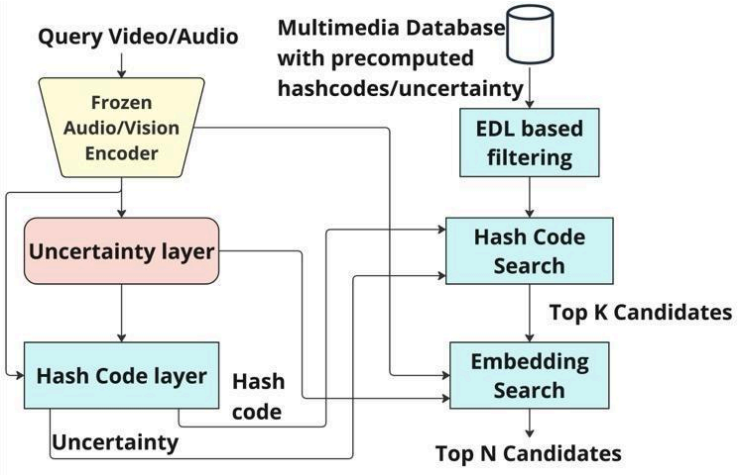
Uncertainty-driven Multimedia Source File Matching

Premium content, such as Hollywood movies, TV episodes, news, and professional sports, involves complex production processes with multiple stages, including pre-production, production, and post-production. During these stages, various assets such as scripts, storyboards, audio, video, images, animations, and visual effects (VFX) are created and combined to produce the final content. These assets, also called Premium data in this document, additionally, there is a need for a reliable way to extract metadata with high confidence levels, enabling users to rely on the extracted information. Raw file directories can often exceed 100TB in size, making manual retrieval through directories highly challenging, even when using Edit Decision Lists (EDLs) generated by editing software. EDLs provide metadata indicating the timeline and sequence of source files used during the editing process, but they can be insufficient for accurate file retrieval due to changes in directories or files during post-production edits.

Recent advancements in the representation learning and large-scale vision/audio language models, such as CLIP, OpenL3, offer promising solutions to effectively retrieve/match multimedia files of interest. A common practice is to extract the embedding of a given video/audio clip and compare it with the embeddings of the other videos stored in the database using distance-based measures such as Euclidean distance, cosine similarity, etc. Some of the drawbacks are that these techniques do not have a holistic approach to account for data uncertainty (affecting the quality of matching), high computation and storage costs.



In this paper, we present a method for efficiently matching source files, such as raw footage, to the final cut of a movie. Our method takes as input a query video, associated EDL files, and a media archive with hash codes and clip metadata, and outputs the matched source files corresponding to the video. Using EDL timecodes, the query video is segmented into clips, and candidate source files are first shortlisted based on duration. These candidates are compared with the segmented clips through a lightweight hash-code similarity function, after which the top-k matches are encoded using pretrained models (e.g., CLIP, OpenL3) to generate embeddings. An uncertainty estimation module then evaluates the reliability of each embedding match, and final mappings are retained only when similarity scores exceed a threshold and uncertainty values remain below a predefined cutoff.

We conducted preliminary experiments using the UCF101 dataset, where we tried to match similar video files. We observed a 10X improvement in the latency with a comparable Mean Average Precision Score of 0.87 with only embedding-based retrieval.