Known-Item Search

Aakash Rathee

Goal

To develop a content-based video retrieval system that can be used for finding small video segments of interest based on a query in a large dataset of videos. This type of task is referred to as known item search.

Problems

- How to find different segments in a video ?
- What type of queries to allow ?

Dataset - V3C100

- It is subset of a much larger dataset(V3C-1), consisting of 100 videos.
- Each item in the dataset consists of a MP4 file and meta-data.

Shot Detection

- FFMPEG’s scdet was used to detect scene change in the video, with a threshold of 2 and sc_pass of 0

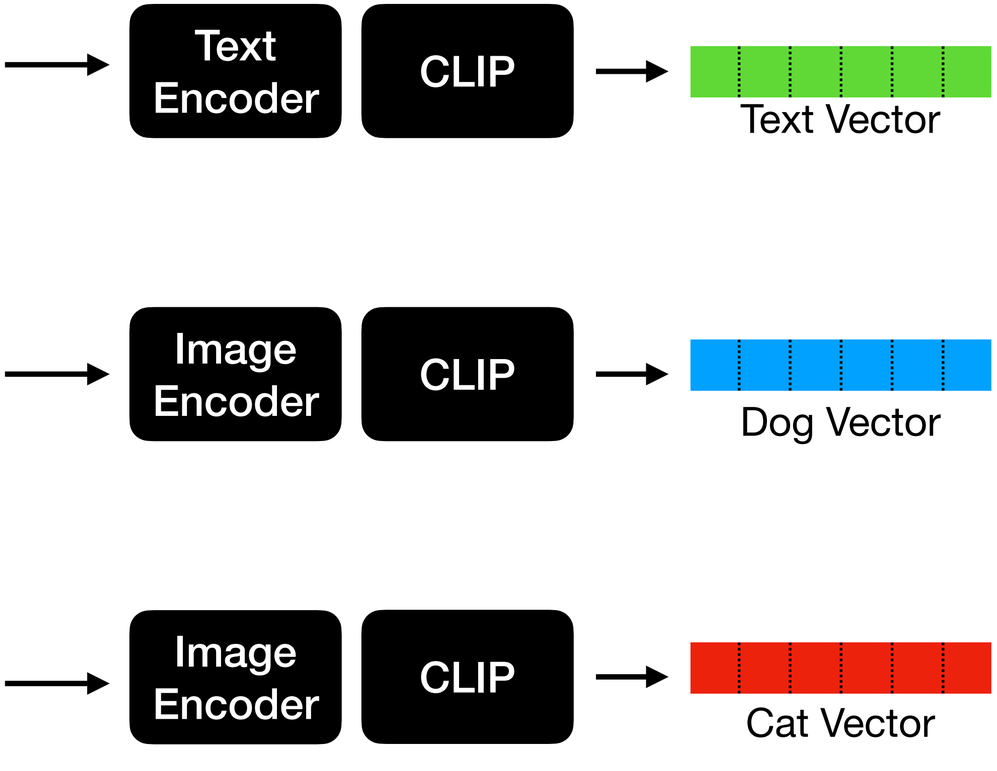
```
ffmpeg -i <src_path> -vf "scdet=s=0:t=2" <target_path>
```

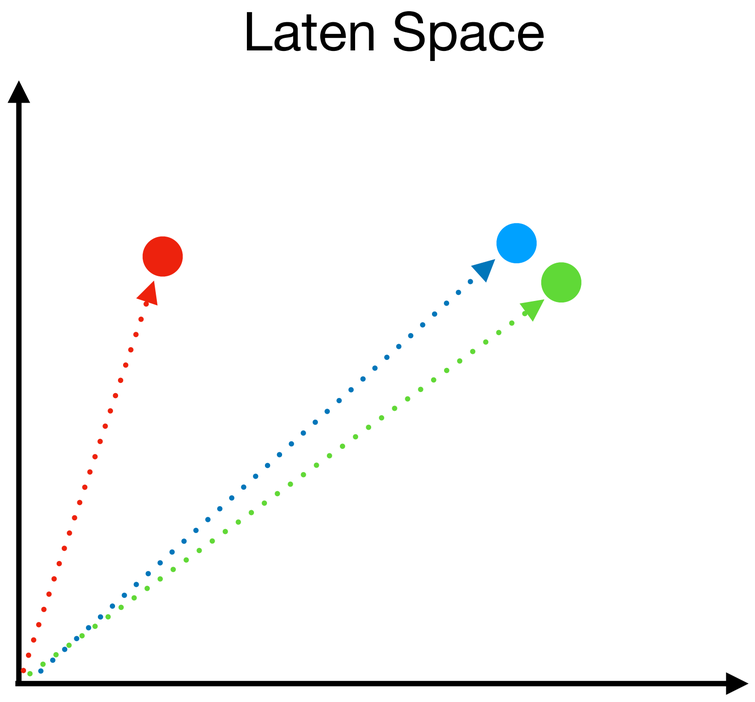
Queries

Query	Model / Library	Description
Text	CLIP ViT-L/14	Find relevant segments using a text query
Image	CLIP ViT-L/14	Find relevant segments which a similar to the target image
Object	YOLOv8	Filter segments based on objects present in it
Word	EasyOCR	Filter segments based on words present in it
Color	Pillow	Filter segments based on the dominant colour in it

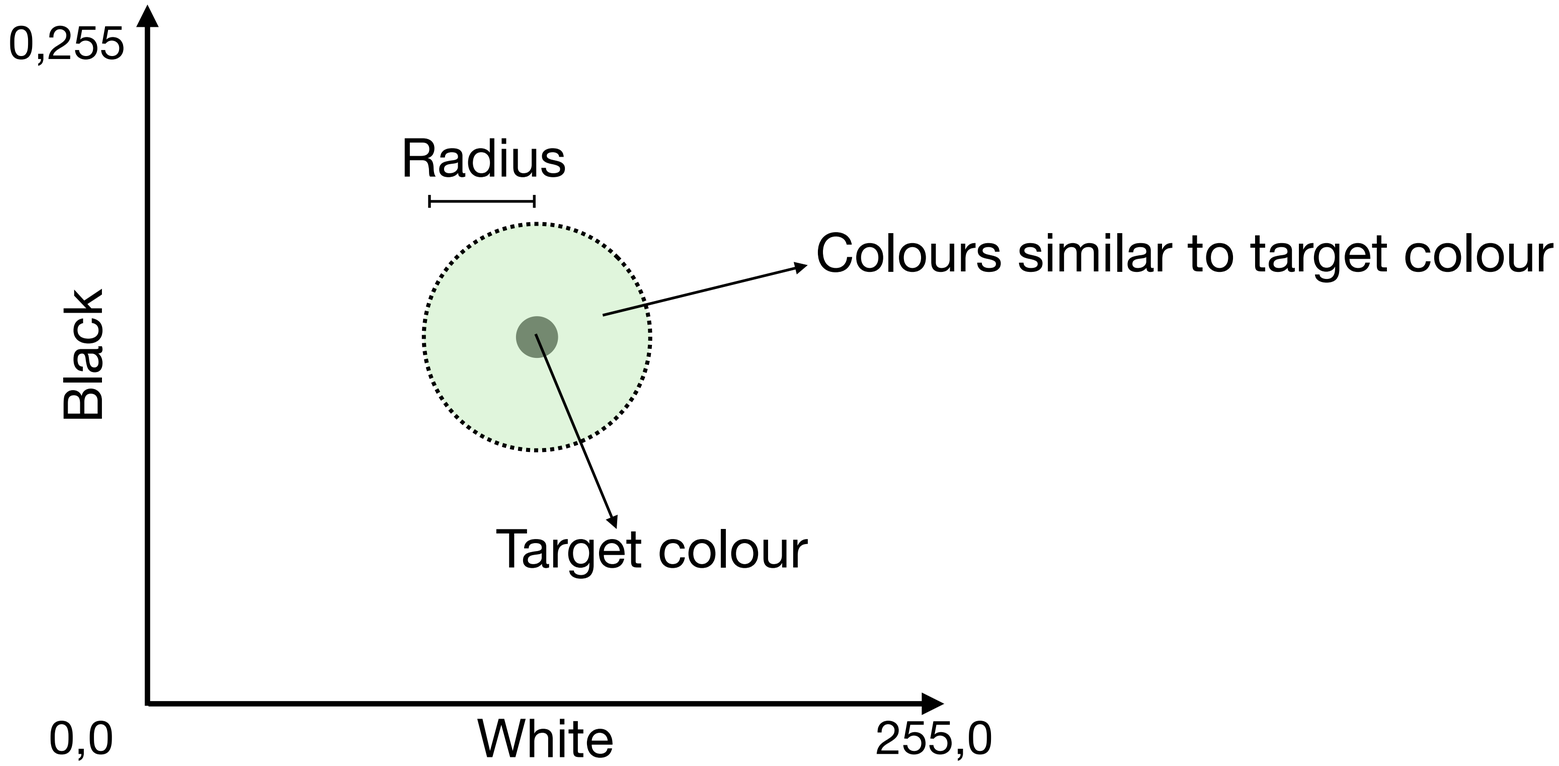
Text and Image Query

A photo of a dog

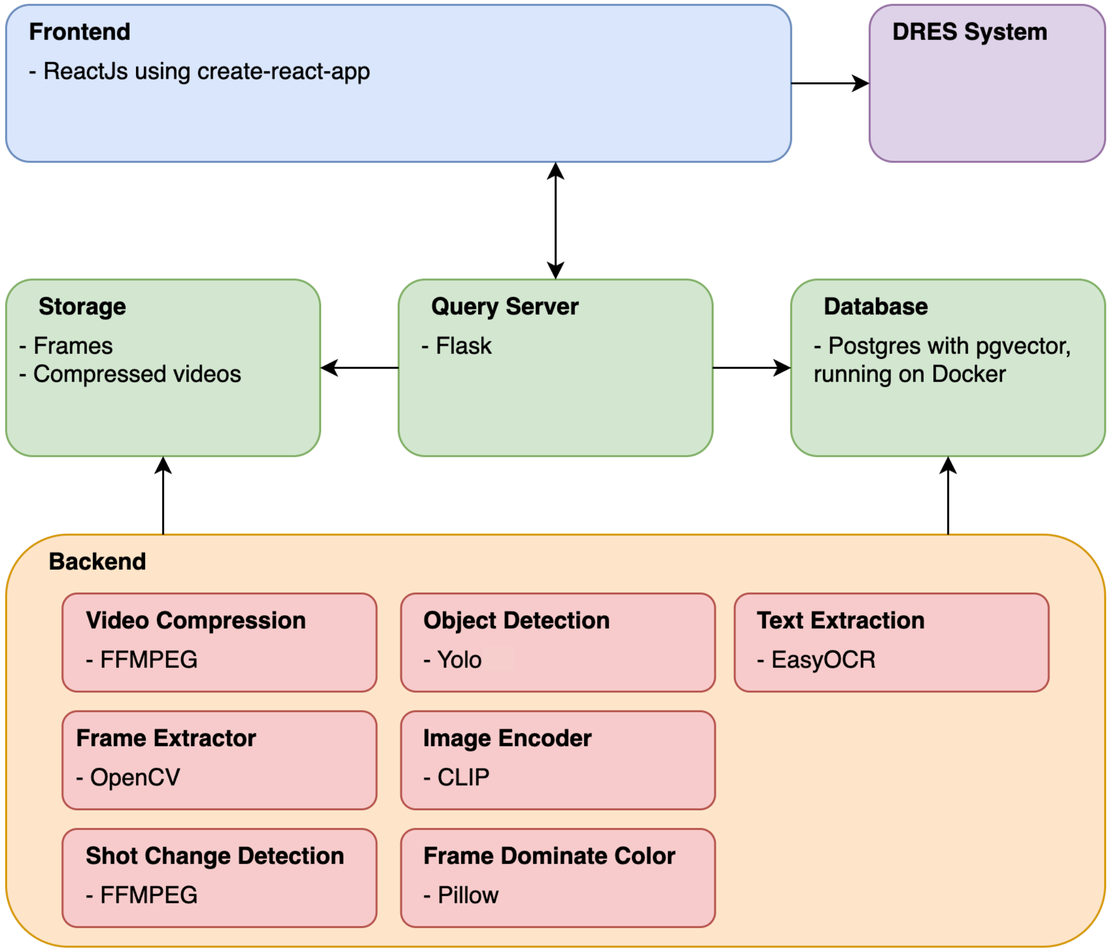




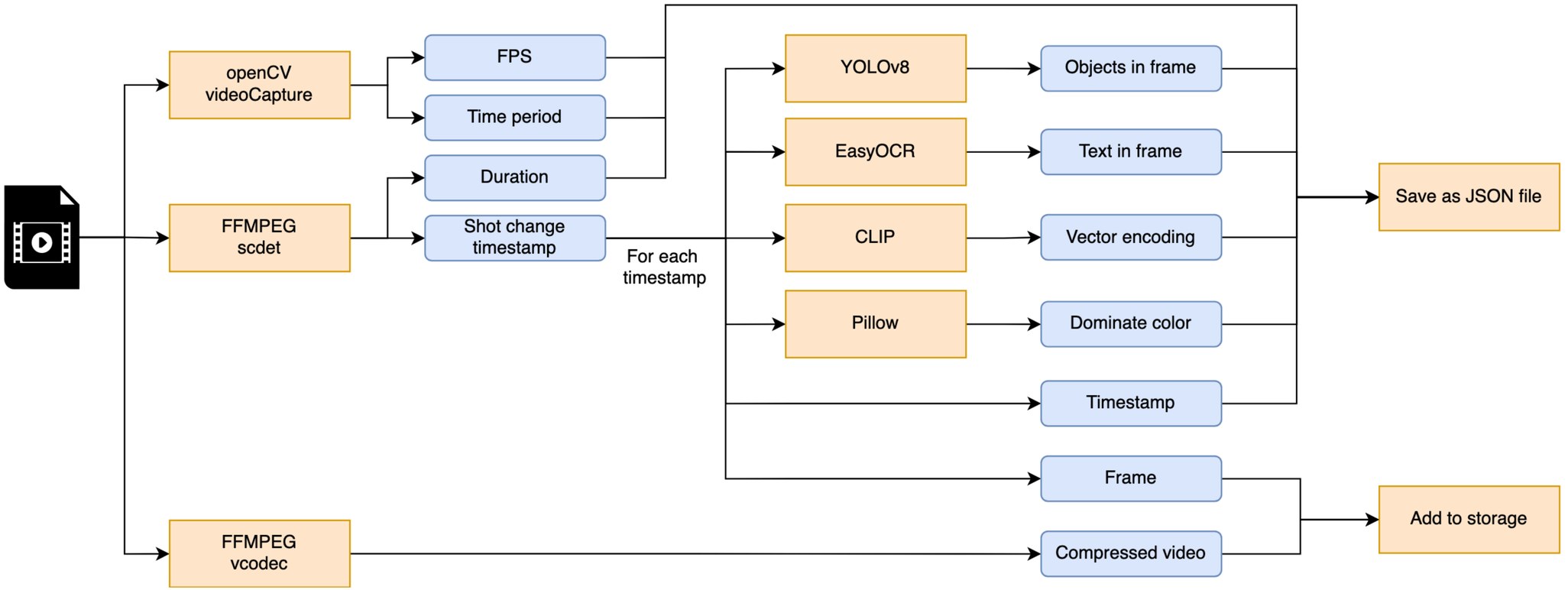
Color Query


0,255
Radius
Colours similar to target colour
Black
Target colour
0,0
White
255,0

Overview

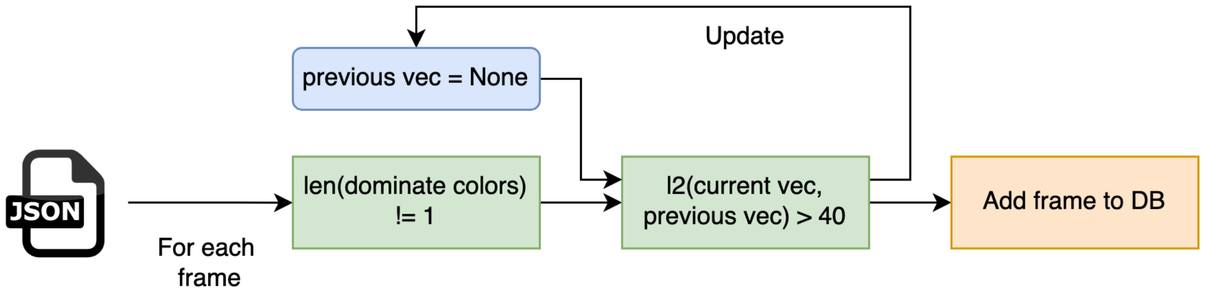

Frontend
- ReactJs using create-react-app
DRES System
Storage
- Frames
- Compressed videos
Query Server
- Flask
Database
- Postgres with pgvector, running on Docker
Backend
Video Compression
- FFMPEG
Object Detection
- Yolo
Text Extraction
- EasyOCR
Frame Extractor
- OpenCV
Image Encoder
- CLIP
Shot Change Detection
- FFMPEG
Frame Dominate Color
- Pillow

Data Extraction

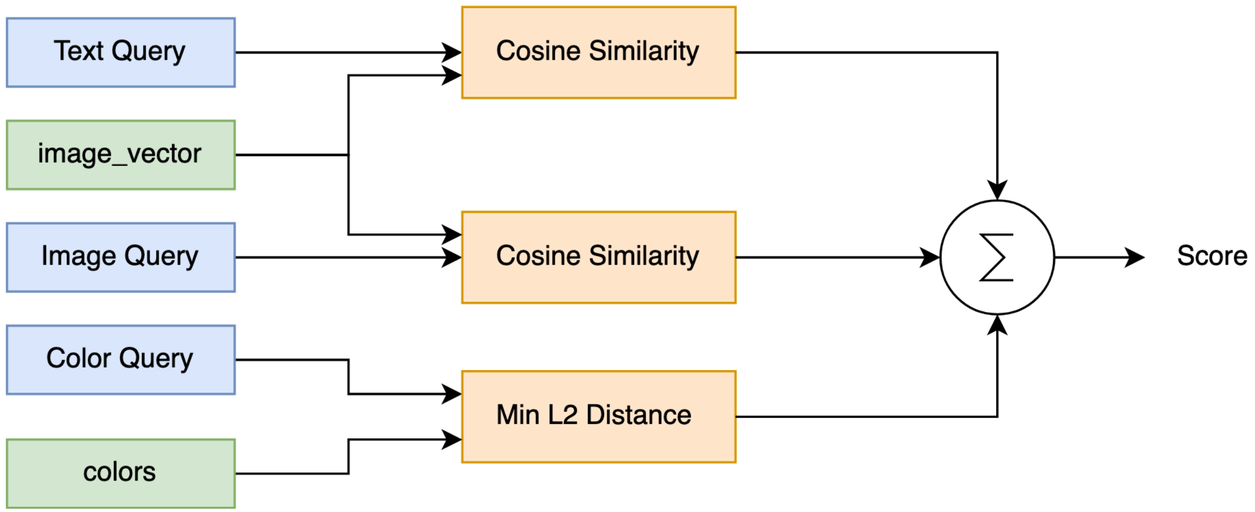


Inserting into DB

id text PRIMARY KEY
video_id text
frame_id text
timestamp float
objects text[]
colors int[][]
image_vector vector(768)
text text[]



Scoring of Segments



Settings

Parameter	Type	Description
Max Results	Integer	Limits the number of frames returned by the database query
Max Text Similarity	Float	Maximum threshold for cosine similarity between the query text and the frame
Color Radius	Integer	
Max Image Similarity	Float	Maximum threshold for cosine similarity between the query image and the frame
Contains	any, all, only	Defines how to look for objects in the frame any: the frame contains any of the defined objects all: the frame contains all of the defined objects, the frame may contain more only: the frame contains all and only the defined objects.

Limitations and Improvements

- OpenAI's CLIP model is limited to English language, text query in any other language will lead to unexpected results. This limitation can be avoided with the use of a translation into English before encoding the text query.
- The objects that can be detected are limited to the YOLO classes. This limitation can be avoided with the use of a different object detection model such as YOLO World, but this approach introduces new problems. As now the user has to define the objects.
- Currently, the system only utilises visual element of the video. It would also be useful for the user if they can perform a search based on the audio content of the video as well. A speech to text model can be used to convert the spoken words into text.

Demo