Interest scenes retrieval in long duration videos using image to text codification

Padilla Luis Edgar Abidán[1] & Pinto Avendaño David Eduardo[1] & Cerino Jiménez Rigoberto[1] & López Cortés Francisco José[1] & Reyes Peralta Alberto Esteban[1]

(1) Facultad De Ciencias de la Computación, Benemérita universidad Autónoma de Puebla, México

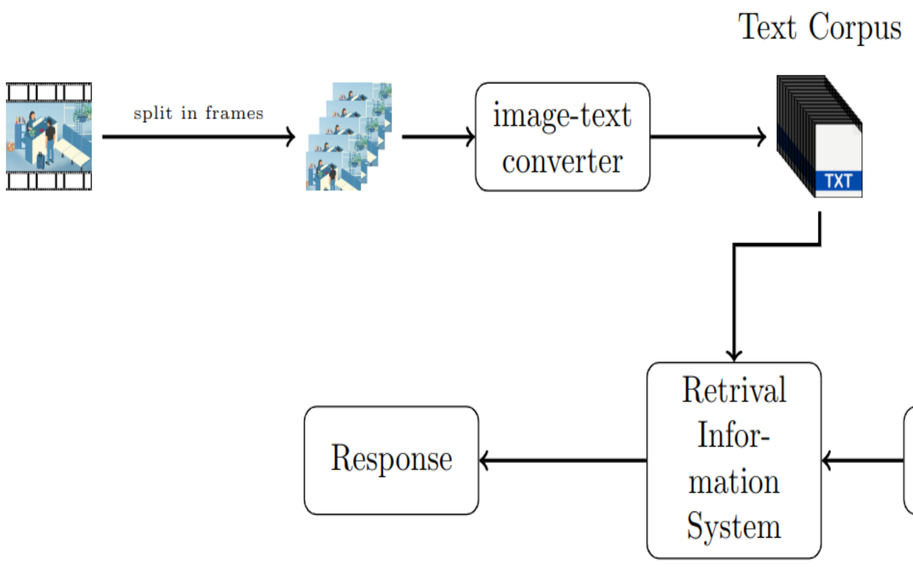
edgar.padillal@alumno.buap.com.mx; davideduardopinto@gmail.com; cerinorigoberto@hotmail.com; lopcorp.z@gmail.com; alberto.reyesp@alumno.buap.mx.

ID paper 69

1. Introduction:

The widespread use of digital cameras results in a significant volume of multimedia files. Extracting information poses challenges, tackled by machine learning techniques such as deep learning [1]. Despite advancements in artificial intelligence, manually searching scenes in lengthy videos remains time-consuming and labor-intensive [2]. An approach based on text information retrieval is proposed to identify scenes in long videos, adapting the algorithm introduced by Luo et al [3] for concealing information in images, thereby creating a system with reduced computational resource requirements.

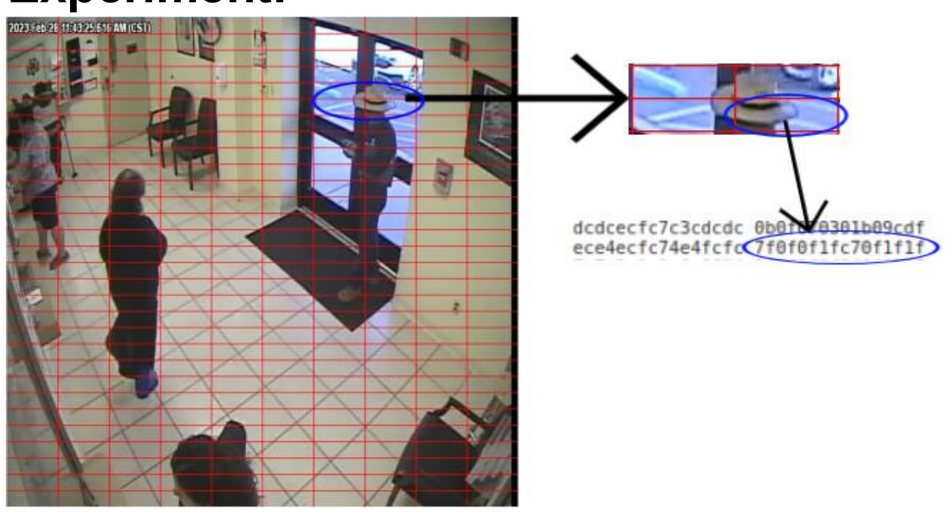
2. Methodology:



3. Image-to-Text Conversion Algorithm:

1. Video frame extraction.
   - Extract images at regular intervals ***t*** from the video, denoted as I_t.
2. Image division.
   - Each I_t is divided into ***m x n*** blocks.
   - Blocks are noted as $B_t = \{B_1, B_2, \ldots, B_{mn}\}$.
3. Mapping blocks to characters.
   - Assign a character string α to each block b_i in B_t using a mapping function ***f***, where $f(b_i) = \alpha$.
4. Mapping Techniques.
   - Grayscale Averaging.
   - Image to Hash Conversion.
5. Text corpus formation.

4. Experiment:



If we consider the string "7f0f0f1fc70f1f1f" to be spelled correctly, using fault-tolerant retrieval with the Levenshtein algorithm [4], it is possible to recover similar images.

Frame	Posicion	Distance
Frame_13	(5,7)	0
Frame_12	(5,7)	0
Frame_14	(5,7)	1
Frame_11	(5,7)	1
Frame_15	(5,7)	2
Frame_16	(5,7)	2
Frame_1176	(10,1)	3
Frame_19	(5,7)	3
Frame_1223	(10,1)	3
Frame_18	(5,7)	3
Frame_17	(5,7)	3

Table: Frame recovered in this experiment when applying fault-tolerant retrieval. In this experiment, an accuracy of 0.81 is achieved

5. Conclusions.

Our study showcases the feasibility of locating portions of an object in an image via information retrieval. However, our findings suggest that it's achievable to detect entire objects rather than just segments, eliminating the necessity for deep learning. These results propose an efficient method for retrieving scenes of interest in long-duration videos without requiring extensive computational resources.

References:

[1] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(5):2687–2704, 2021.

[2] Niels Haering, Péter L Venetianer, and Alan Lipton. The evolution of video surveillance: an overview. Machine Vision and Applications, 19(5-6):279–290, 2008

[3] Yuanjing Luo, Jiaohua Qin, Xuyu Xiang, Yun Tan, Qiang Liu, and Lingyun Xiang. Coverless real-time image information hiding based on image block matching and dense convolutional network. Journal of Real-Time Image Processing, 17:125–135, 2020.

[4] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10, pages 707–710. Soviet Union, 1966.