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

Edgar Abidán Padilla Luis, David Pinto, Rigoberto Cerino Jiménez, Francisco José López Cortés, Alberto Esteban Reyes Peralta, Beatriz Beltrán

Facultad De Ciencias de la Computación, Benemérita universidad Autónoma de Puebla, Avenida San Claudio, Heroica Puebla de Zaragoza, 72570, Puebla, México.

*Corresponding author(s). E-mail(s): david.pinto@correo.buap.mx;
Contributing authors: edgar.abidan.pl@gmail.com;
cerino_rigoberto@hotmail.com; lopcorp.z@gmail.com;
alberto.reyesp@alumno.buap.mx; bbeltranmtz@gmail.com;

Abstract

This article presents an approach for retrieving scenes of interest in long-duration videos through image-to-text encoding. Unlike conventional approaches that often involve the use of neural networks, this method proposes a technique that avoids the use of these complex structures in order to reduce computational resource consumption. Through experiments, the feasibility and effectiveness of this technique are demonstrated, concluding that it is feasible to employ it for multimedia information retrieval, offering an efficient and economical alternative for this task.

Keywords: Information retrieval, scene identification, long-duration videos, image-to-text encoding.

1 Introduction

In recent years, due to the development and creation of better technological devices, there has been an increase in the use of digital cameras, such as cell phones, surveillance systems, specialized cameras, among others, resulting in a large volume of multimedia files. Extracting information from these massive data volumes presents various challenges, which vary depending on the type of information to be retrieved (video, images, audio, or text). Some of the techniques used to obtain information belong

to the field of *machine learning*, with deep learning standing out, as it is capable of automatically extracting features from data [7]. However, despite the growth in artificial intelligence studies, many public and private organizations still rely on human resources to perform specific tasks such as searching for scenes of interest in long-duration videos. This method of obtaining information is slow and consumes a large number of human-hours, as the process of identifying scenes in videos is difficult and tedious, especially when dealing with large volumes of data. In contrast, it is observed that image information retrieval techniques help reduce human costs in this task [10].

Currently, the advancement of deep learning provides us with tools to develop multimedia information retrieval systems. However, if these types of techniques are to be implemented, the following points must be considered: a large corpus is needed to train deep learning models; training these models often takes too long; a specialized model is required; and computational resources are high (RAM, GPU, processor, and disk space).

This document proposes an approach based on text information retrieval to identify scenes of interest in long-duration videos. These techniques tend to require fewer computational resources, making them ideal for adaptation on many servers that store videos.

2 State of the Art

Multimedia Information Retrieval Systems (MIRS) are closely linked to technological advancements. There are various techniques applied to these systems. For example, the Bag-of-Visual-Words method involves finding global descriptors of images by creating histograms of high dimensionality. This technique offers the ease of creating inverted indices and vector spaces to measure their distance from other images, but it tends to be less accurate in searches [27]. It has been shown that global descriptors do not capture relevant characteristics well, and local descriptors tend to yield better results [5]. In the past decade, the concept of Vector of Locally Aggregated Descriptors (VLAD) was introduced. It extracts regions using invariant detectors (algorithms for extracting characteristics from an image, robust against geometric transformations) and then characterizes them by SIFT descriptors (Scale-Invariant Feature Transform). The results are classified into clusters, which for information retrieval are associated with a vocabulary [13]. These types of systems are known as Content-Based Image Retrieval (CBIR) systems and are used in various fields, such as medical applications [4]. In recent years, methods for obtaining image descriptors have been improved. For example, Alsmadi in [1] obtains color, shape, and texture information, and by combining this information, presents results with good accuracies. Some methods additionally create hash codes (bit strings) from the constructed descriptors, which are used to index images and apply IR algorithms based on these indices [17].

However, the rise of artificial intelligence, specifically the use of Convolutional Neural Networks (CNNs), has shown that descriptors can be found almost automatically [25]. Hash codes are used to simplify and accelerate searches; currently, CNNs are trained to obtain these codes [11], and studies are beginning to be conducted with

other architectures like transformers to generate these codes [8]. It is worth mentioning that the use of CNNs requires a dataset to match images with text. CNNs have shown the ability to associate images with text and text with images by projecting the text and the image into a single feature subspace [23, 9, 12].

There are various studies focused on the medical field that use CNNs to create specialized Multimedia Information Retrieval Systems (MSIRS). Shamna and Aziz train a magnetic resonance classification model using CNNs and, through a similarity comparison, retrieve related documents that will assist the physician in making a diagnosis [20]. In the same vein, Zhang et al. employ neural networks to classify resonances and match them with text related to medical diagnosis, thus providing specialists not only with retrieved images but also with potential diagnoses [26].

In the fashion domain, Whu and Gao introduce the first dataset to support the advancement of image retrieval systems for fashion [24], leading to the development of the first image retrieval models referencing this database [6, 19].

On the other hand, geolocation has gained great importance, from the automation of unmanned vehicles to tourist use. Tang et al. present an information retrieval model that does not rely on a global satellite navigation system. They use CNNs to predict images of the environment and, through a RIM system, retrieve possible geographic areas in which they are located [22].

In terms of security, video surveillance is of utmost importance. Prathiba introduces a multimodal retrieval system using a promising clustering algorithm for human-computer interaction. This algorithm extracts features from the frames comprising the video, applies the nearest neighbor-based algorithm, and based on this, calculates the distance between frames to retrieve the most similar images [18].

A quick review of the state of the art highlights that neural networks lead the way in tasks related to multimedia information retrieval [2, 3, 8, 9, 11, 12, 16, 18, 20, 21, 22, 23, 24, 25, 26].

3 Methodology

The methodology proposed in this document is illustrated in Fig. 1. This methodology consists of two main components: firstly, a module is included to generate the text corpus, which is carried out through an image-to-text conversion algorithm; secondly, a text information retrieval system is implemented, allowing text queries to be made and relevant documents associated with the query to be obtained as a response.

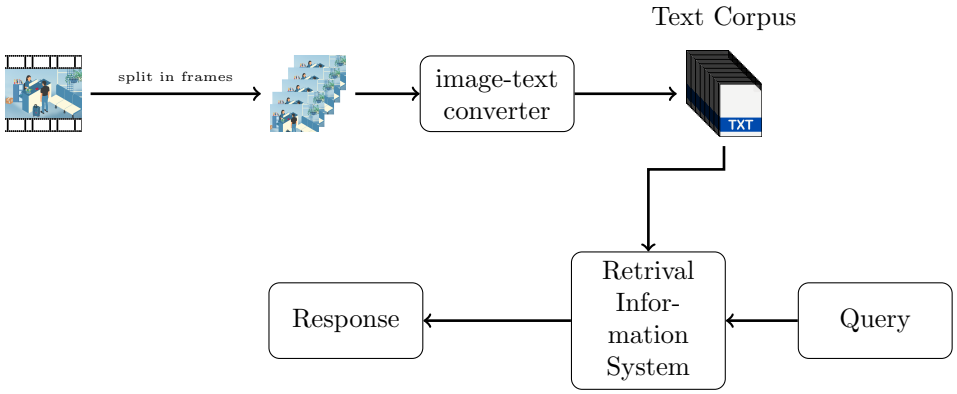


Fig. 1: Proposed methodology for identifying scenes of interest in long-duration videos using an image-to-text encoding module.

3.1 Image-to-Text Conversion Algorithm

After obtaining the video frames, a modification to the steps for extracting feature sequences presented by Luo et al. [15] is applied. This algorithm is used to hide information in images; however, a text corpus is created using the following steps: from the video, images are extracted at regular intervals t, referred to as I_t. Each I_t is divided into $m \times n$ blocks, where each block is part of the set $B_t = \{B_1, B_2, \ldots, B_{mn}\}$. A character string α is assigned to each $b_i \in B_t$ through some mapping f, such that $f(b_i) = \alpha$. This document presents two mappings: conversion to characters by averaging grayscale and image-to-hash code conversion. Thus, each image is represented by a text file, and the generated text files form the corpus that feeds the information retrieval system.

4 Experiments

For the following experiments, a 5-hour video recording an office at a tax agency in the United States was downloaded from the internet.

4.1 First Experiment: Conversion to Characters by Averaging Grayscale

Based on the idea presented in section 3 for obtaining textual representations, the following process is carried out to obtain a text corpus:

1. A frame is obtained every 1.5 seconds.
2. A partition of each frame is obtained as follows:
   (a) Each image, I_i, is divided into N rows and M columns, forming a matrix of sub-images.
   (b) Each sub-image, $S_{n,m}$, is divided into P elements.

3. Each element, P_p, undergoes a mapping f, where this mapping transforms each element into grayscale values, then the average of the grayscale colors is obtained, and this average is associated with one of the 26 letters of the English alphabet.
4. The letters are joined to form a character string.
5. The document will have N rows with M character strings.

(a) Frame 13 of the video.

(b) The text file associated with frame 13 will consist of 30 lines.

(c) Each line will consist of 10 words.

(d) Each word will contain 9 letters.

Fig. 2: Partition of frame 13 obtained from the video.

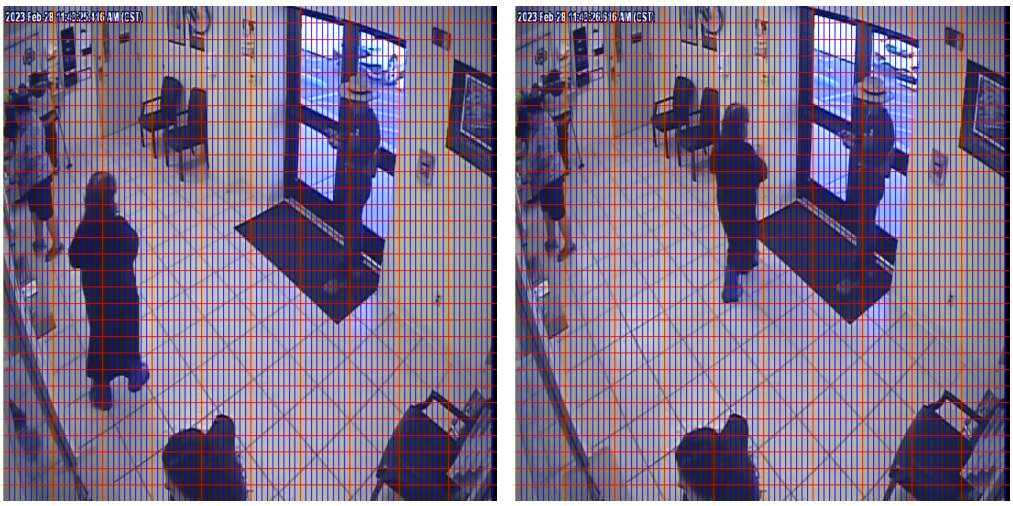
For this experiment, the parameters are as follows: $N = 30$, $M = 10$, and $P = 9$. In Fig. 2, the process for partitioning frame 13 (Fig. 2a) obtained from the video is shown. For this purpose, the frame is divided into 30 rows (Fig. 2b) and 10 columns (Fig. 2c). Subsequently, each sub-image is divided into 9 elements (Fig. 2d). This division can be interpreted as a text document consisting of 30 lines, with each line containing 10 words composed of 9 letters.

Once the division process is complete, the mapping f is applied, thus obtaining a textual representation for each sub-image and a text document by concatenating all the representations following steps 5 and 6 presented in this section. The document obtained after applying this process is shown in Fig. 3.

```
ikkjklllm kmjllmlll jmojklknn opnmkklil nkprpppqq qrrqqpjfg lnnmlkmnn opppppoo oooonnnmm mmllkkkia
ijkllllno nkhhpopmm hlmkjlmoo oppmmknjl nnrspppo mrrrsrkfg syyyyywsh ggnmkjkmn ppooooiik lmmmlllja
jkkkjlmnq oiedpnkmm ikmkkmmoo ppqooopkk mssspppqp ossssrkfg rvvxyyxqh gjpsswxur oopppoljk knnmmmlja
klmljlmmm mkefnoool lmmmmmloo pppqqqqqr rsrsqqqrr qssssrkef psssrrrlg gkompqsws tqpppppoo oghiklmka
klmkleefk kkgfkppok lmmmmmmop ppiijnqqq qrrsqqqrr prsssrkef qrrrsrrli llmmmmmnr rqqqqpppp mddiihgga
jffefggjk kknoppppm mmmmmmmmn nkeeehnff gsssqqqrs ssssrqjde rssrssskk jjjlqrsrr qqqqqqqpp jefikjjha
fdeegkkii hlopppppm mmmmmmlhk igddfihee erssqqqrs sssrpohee egjlnpshf fggglrrrr rrrqqqqqq gfihikiga
dcegikmhd imooppppn lmmmmmmhi gifffgfee eqrsrqqrr rssrqphdg spnkhggff ggfffkrrr rrrrrrqqo ffiiihhfa
hghjhjlid hnmoooooo nooooool hkghgfggi gorsrqqrr rsssrqhdh uuuuuplff fffffinpr rrrqpqqqn ifgggihga
giigfhjlj hmkmoooop pnkkkmppp nolmlhkjk jmrrrqqqr rrrsrqfdj stttttnff ffffgheem rrrokknqq qqnljhgfa
jigffghgg knooppqqq meeeefnqq ppppplnnn noqrrpppq qrrrqofdg mqsssskef eeffottpq rrrollqqq qqqqppnka
jiheecddf noopppqqq hddddiqqq qrrqqqqqq qqqqqqqqq pqqonlgfe edeilogee eeejssssr rrrrrrrrr qqqqqppna
ikjkkfdde moppqqqph dddddegkq rrrrrrrrr rrrrrrrrr pmkhfeffh iihfdddde eemrrssrr rrrrrrrrr qqqqpppna
jkllmkhlj joppqrrid ddddddddee iqsrrrsss rrrrrrnhf fffffffff fffhiigfe eegknrrrr rrrrrrrrr qqqqqppna
nillligjl hkqqqrrkd ddddddddd gqsssssss sssssssp lhfffffff fffffeghe eeedddmqr rrrrrrrrq qqqqqppna
llhhkmllk mprrrqrrl ddddddddg sssssssss sssssssss ssqmiffff fffffeee eegjihqqr rrrrrrrrq qqqqpppna
hmjiknmno opprssrqr fddddddei sssssssss sssssstsss ssrrrrnjf fffffffee efeeelqqq rrrrrrrqq qqqppppma
gimopommo oopqqrrrq ofddddddk sssssrst ttttsrss sssssrrr njfffffee eefilpqqq qqrrqqpoq qqpppppma
iginmklll opopqqqqq pidddddl rsssssss sssssssss ssssrsss rrrokgffg kossrpppp qqqqqqqpq qppppooma
jigiokjik mooopppqqn fddddddn srrsssss sssssssss srrrssss srrqqropr ssrrqpppp pppppqqq ppppooola
kihginnpo ppppooppn iededdddq sssrrssss sssrrrsss ssssrrrrr rsrrrqqqr qqqqqpppq pppopppppp pppooonla
lkljginpp pooooopp pnddddegq srrssrrsr rsrsssrrs sssrrrrrr rsrrrrrqq qrrrqqqqp oopppppoo ooooonmia
klllkhiop pppppppoo oonliefgl qrrrrrrrr sssssssr qrrssssr rrrrrrrrr rrqqqqppp qpppoppoo oonnnmlka
klkkllhhn ppppoooon nmmmmkkon oqqrssssss srssrrrrs srrrrssss rrrrrrrrr rrrqqpqqq qqpppkijm onnmjjkia
jkkklmlhh mpoopppoo nnoonmnno poprrssss srqnprsss sssrrrrrr ssssrrqqq rqqqrqqqp plhfglonk iiiihggfa
kkkkllmmi gkpppppoo nnmnooooo pppppnnmn qngfffqss ssrrrrrrr sssssrrqq qqqqqqqql gfeeeegil lkiefihha
lkkkklmmm jgioonnno ooonnoooo oolfeeeee eeffffpqr rrsssssr rqrrrrrrr rrqppqohf eeeeeeedd efgkkllja
llkjjkkll lkghlmmmn nooonnnnn oofeedddd eeeeefhpr qrsssssss rqqqqrrrr rqqqqmhfe eeeeeeddd eillkkkja
lllkjjjkk kkkjhlmmm mnnnnnnoo oplfddddd ddddefiq rrqqrrrrr rrrrrqqqr qqppofeee eeeeedddd iihghijga
klllkjjjk kkkkihjlm mmnooooon opoiifedd ddddddefm qrrqppqrr rrrrrrrqp ppppjknig eddddddddd deghkiffa
```

Fig. 3: Textual representation obtained from frame 13.

After setting frame 13 as the scene of interest, a manual search was conducted in the nearby frames. As a result, it was found that the scenes from frame 11 to frame 20 are very similar. Observing Fig. 4, it is inferred that the only significant change occurring from frame 11 to frame 20 is the movement of a person.



(a) Frame 11 of the video. (b) Frame 20 of the video.

Fig. 4: Frames close to frame 13.

In frame 13, focus is placed on the hat of the person near the door, specifically on the hat band. Fig. 5 shows this focus, and the associated string for this element is **jjjlqrsrr**.

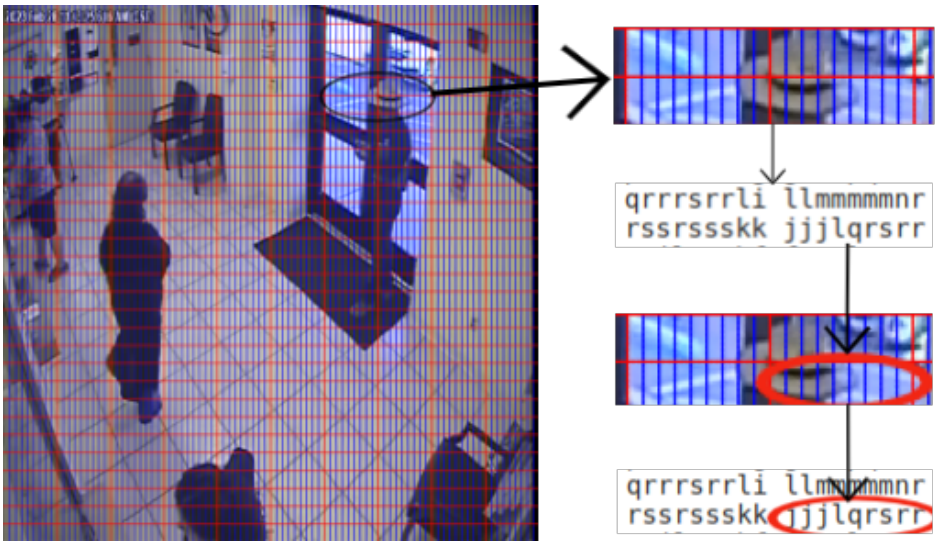


Fig. 5: Analyzing the hat band present in frame 13.

When performing a search in the information retrieval system, it is found that the only file containing the string **jjjlqrsrr** is frame 13, and this string is found on line 5 and column 7. If a search is conducted for the tokens representing this band, it is noticed that this band appears in the same position in frames 11 and 20 (the extreme frames of the manual search). The strings in that position are:

- Frame 11 -> **jjjlqrs**s**r**.
- Frame 20 -> **j**kkkoq**s**s**r**.

It is observed that the only difference between the string obtained from frame 11 is the letter 's' (highlighted in red), while in frame 20 there is a greater difference in the strings.

If it is considered that the string **jjjlqrsrr** obtained from frame 13 is a word and that this word is correctly spelled, it can be considered that the strings obtained in frames 11 and 20 are misspelled words, and their correction should be the string **jjjlqrsrr**. In this case, we could apply an algorithm present in fault-tolerant information retrieval, which calculates the Levenshtein distance [14]. For example, the Levenshtein distance between the strings **jjjlqrs**s**r** and **jjjlqrsrr** is 1, as only the letter 's' (highlighted in red) needs to be changed to 'r' in the first string to obtain the second string.

File	Position	Distance
frame_13.txt	(5, 7)	0
frame_11.txt	(5, 7)	1
frame_12.txt	(5, 7)	
frame_967.txt	(17, 1)	3
frame_1106.txt	(20, 0)	
frame_1114.txt	(20, 0)	
frame_1113.txt	(20, 0)	
frame_1109.txt	(20, 0)	
frame_1111.txt	(20, 0)	
frame_1115.txt	(20, 0)	
frame_1110.txt	(20, 0)	
frame_1112.txt	(20, 0)	
frame_1107.txt	(20, 0)	
frame_1108.txt	(20, 0)	
frame_501.txt	(18, 4)	
frame_14.txt	(5, 7)	
frame_862.txt	(19, 4)	
frame_500.txt	(20, 4)	
frame_503.txt	(25, 4)	

Table 1: Strings with Levenshtein distance less than or equal to 3.

All documents containing tokens with a Levenshtein distance less than or equal to 3 from the token **jjjlqrsrr** were retrieved, resulting in 19 tokens. Table 1 shows the results.

Out of the 19 results, only 4 documents are within the desired range.

4.2 Second Experiment: Conversion to Hash Code

For the second experiment, the following steps are followed:

1. A frame is obtained every 1.5 seconds.
2. Each frame is partitioned as follows: Each image, I_i, is divided into N rows and M columns, forming a matrix of sub-images.
3. Each sub-image, $S_{i,j}$, is mapped using function f to obtain a hash code.

Fig. 6 shows the partition of frame 13 into 30 rows and 10 columns.

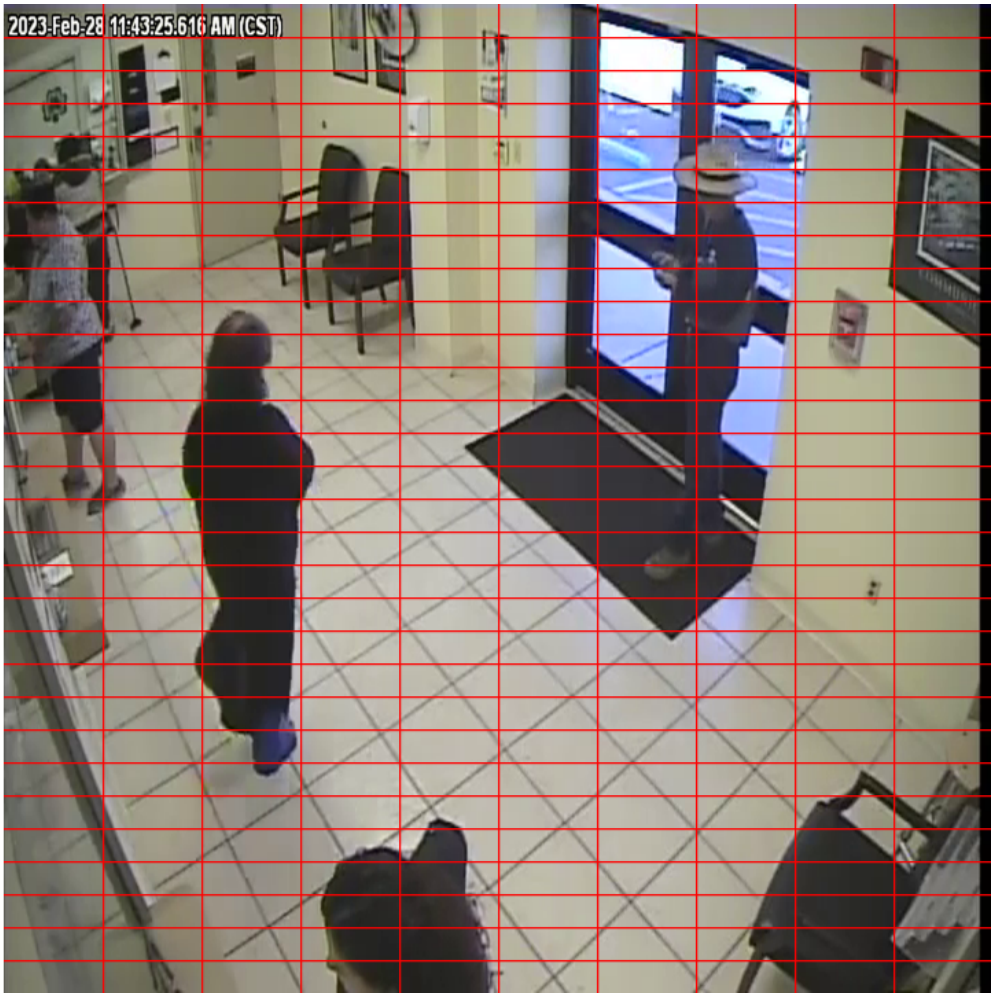


Fig. 6: Partition of frame 13 in the second experiment.

Fig. 7 shows the text document associated with frame 13.

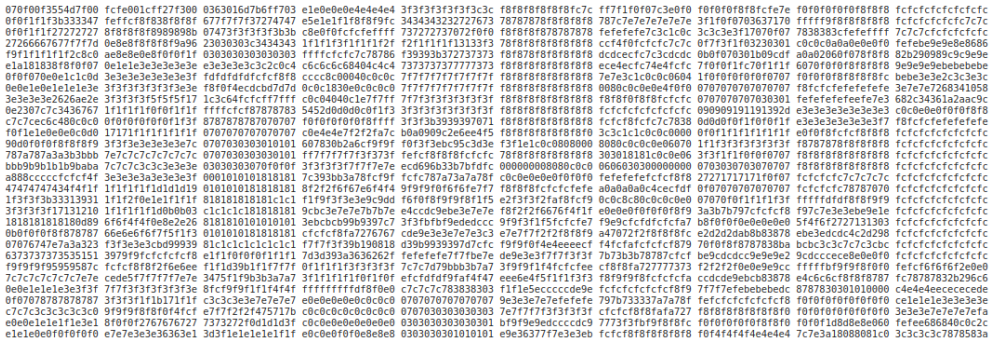


Fig. 7: Text associated with frame 13 in the second experiment.

Again, the section of frame 13 with a person wearing a hat is analyzed, focusing on the hash code representing the hat's stripe. Fig. 8 shows this image segment to be analyzed.

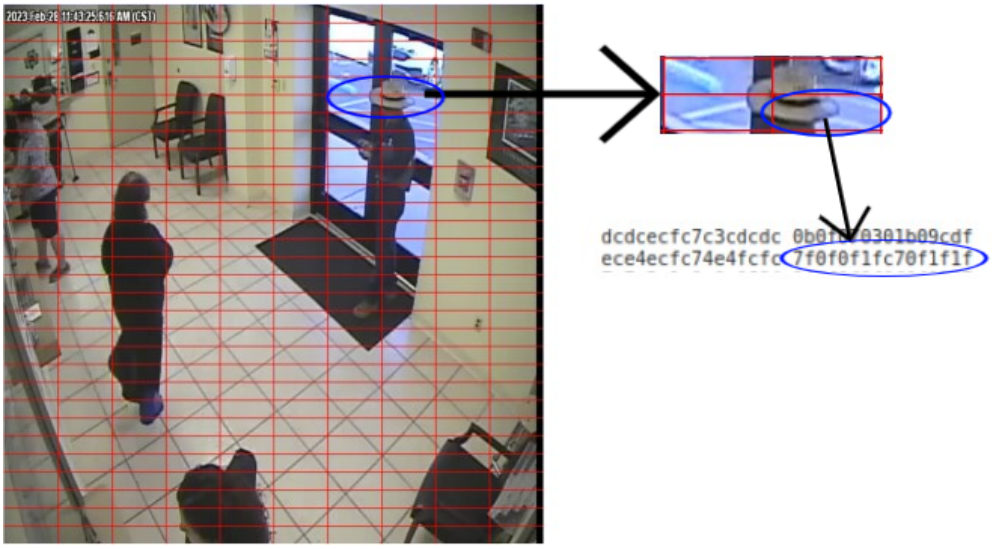


Fig. 8: Analyzing frame 13 with hash code.

Searching for the string **7f0f0f1fc70f1f1f** in the information retrieval system retrieves two documents: frame 13 and frame 12. Similar to the previous experiment, the hat stripe appears in that position from frame 11 to frame 20. The corresponding hash codes are as follows:

- Frame 11 -> **5f0f0f1fc70f1f1f**.
- Frame 20 -> **7e0f0f0fc7670f1f**.

Again, if we consider that the strings in frames 11 and 20 are misspelled, we can apply the Levenshtein algorithm. Searching for tokens with a distance less than or equal to 3 from the token **7f0f0f1fc70f1f1f** retrieves 11 results, which are presented in Table 2.

File	Position	Distance
frame_13.txt	(5, 7)	0
frame_12.txt	(5, 7)	
frame_14.txt	(5, 7)	1
frame_11.txt	(5, 7)	
frame_15.txt	(5, 7)	2
frame_16.txt	(5, 7)	
frame_1176.txt	(10, 1)	3
frame_19.txt	(5, 7)	
frame_1223.txt	(10, 1)	
frame_18.txt	(5, 7)	
frame_17.txt	(5, 7)	

Table 2: Documents retrieved by applying fault-tolerant information retrieval.

Out of the 11 results, 9 fall within the manually found range.

5 Discussion

It is important to consider the following points regarding resource usage and time. For the first experiment:

- 1,237 text files were obtained.
- The process of obtaining the corpus took approximately 15 minutes and 40 seconds, meaning each image took 0.75 seconds to process.
- The total weight of the text files is approximately 3.7 MB.

For the second experiment, the following information is available:

- 1,237 text files were obtained.
- The process took approximately 14 minutes and 32 seconds, approximately 0.70 seconds per frame.
- The total weight of the text files is approximately 6.3 MB.

This data indicates that the use of computational resources is very low compared to the use of neural networks since it is not necessary to store large amounts of images or use a GPU for training neural network models. Additionally, the required time is less since not only is the construction of a specialized dataset for the task avoided, but also the training time and adjustment of deep learning models.

When analyzing the results presented in Table 1 (referring to the first experiment where the average grayscale is used as mapping), it is observed that 4 out of the 19 recovered frames correspond to the scene in the video where the hat's stripe appears.

On the other hand, in Table 2 (referring to the second experiment which applies a Hash code as mapping), it is observed that 9 out of the 10 recovered frames correspond to the scene of interest. This suggests that converting image to text using Hash code leads to better performance in the retrieval system.

By using an evaluation metric, it is possible to obtain a more concrete data to assess these two experiments. The precision metric is used to compare their results, and it is given by the following formula:

$$Accuracy = \frac{|\text{relevant documents} \cap \text{recovered documents}|}{\text{recovered documents}}$$

Considering that there are 10 relevant documents (from frame 11 to frame 20) and taking into account the results from Tables 1 and 2, for the first experiment we have a precision of:

$$\text{Accuracy} = \frac{4}{19} = 0.21$$

and for the second experiment we have a precision of:

$$\text{Accuracy} = \frac{9}{11} = 0.81$$

This indicates that using Hash code as mapping yields better results compared to using the average grayscale as mapping.

6 Conclusions

Currently, the development of artificial intelligence greatly contributes to the advancement of technologies. However, it is important to note that not all organizations and individuals have access to the computational resources required for these algorithms, as well as the time invested in this area. Therefore, it is vital to continue researching alternative techniques.

This article demonstrates that it is possible to find parts of an object of interest in an image through information retrieval. However, with the results presented here, it is considered possible to find objects in their entirety and not just a section of them, without the need for deep learning. With the results generated through this research, a possible efficient method for retrieving scenes of interest in long-duration videos is opened without the need for high computational resources.

References

[1] Mutasem K Alsmadi. Content-based image retrieval using color, shape and texture descriptors and features. *Arabian Journal for Science and Engineering*, 45(4):3317–3330, 2020.

[2] Hanan Butt, Muhammad Raheel Raza, Muhammad Javed Ramzan, Muhammad Junaid Ali, and Muhammad Haris. Attention-based cnn-rnn arabic text recognition from natural scene images. 2021.

[3] Jorge Luis Suaña Chambi, Juan Carlos Gutiérrez Cáceres, and Cesar Armando Beltrán Castañón. Densenet 3d for violent action recognition in surveillance video sequences. In *2022 41st International Conference of the Chilean Computer Science Society (SCCC)*, pages 1–8, 2022.

[4] Filipe Coelho and Cristina Ribeiro. Evaluation of global descriptors for multimedia retrieval in medical applications. In *2010 Workshops on Database and Expert Systems Applications*, pages 127–131. IEEE, 2010.

[5] Thomas Deselaers, Daniel Keysers, and Hermann Ney. Features for image retrieval: an experimental comparison. *Information retrieval*, 11:77–107, 2008.

[6] Eric Dodds, Jack Culpepper, and Gaurav Srivastava. Training and challenging models for text-guided fashion image retrieval. *arXiv preprint arXiv:2204.11004*, 2022.

[7] 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.

[8] Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu. Vision transformer hashing for image retrieval. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[9] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[10] 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.

[11] Lirong Han, Peng Li, Xiao Bai, Christos Grecos, Xiaoyu Zhang, and Peng Ren. Cohesion intensive deep hashing for remote sensing image retrieval. *Remote Sensing*, 12(1):101, 2019.

[12] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. *arXiv preprint arXiv:2306.07196*, 2023.

[13] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[14] 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.

[15] 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.

[16] Alonso Medina Cortes, Magdalena Saldaña Pérez, Humberto Sossa Azuela, Miguel Torres Ruiz, and Marco Moreno Ibarra. An application of deep neural network for robbery evidence using face recognition approach. In Anna Visvizi, Miltiadis D. Lytras, and Naif R. Aljohani, editors, *Research and Innovation Forum 2020*, pages 23–36, Cham, 2021. Springer International Publishing.

[17] Şaban Öztürk. Stacked auto-encoder based tagging with deep features for content-based medical image retrieval. *Expert Systems with Applications*, 161:113693, 2020.

[18] T. Prathiba and R. S. S. Kumari. Content based video retrieval system based on multimodal feature grouping by kfcm clustering algorithm to promote human–computer interaction. *Journal of Ambient Intelligence and Humanized Computing, 12, 6215-6229.*, 2021.

[19] K. Saito, K. Sohn, X. Zhang, C. L. Li, C. Y. Lee, K. Saenko, and T. Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. *n Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19305-19314).*, 2023.

[20] NV Shamna and B Aziz Musthafa. Feature extraction method using hog with ltp for content-based medical image retrieval. *International journal of electrical and computer engineering systems*, 14(3):267–275, 2023.

[21] G. Tahmasebzadeh, E. Kacupaj, E. Müller-Budack, S. Hakimov, J. Lehmann, and R. Ewerth. Geowine: Geolocation based wiki, image, news and event retrieval. *In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2565-2569).*, 2021.

[22] J. Tang, C. Gong, F. Guo, Z. Yang, and Z. Wu. Automatic geo-localization framework without gnss data. *IET Image Processing, 16(8), 2180-2195.*, 2022.

[23] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021.

[24] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317, June 2021.

[25] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 53–61, 2015.

[26] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[27] Qiuzhan Zhou, Cheng Wang, Pingping Liu, Qingliang Li, Yeran Wang, and Shuozhang Chen. Distribution entropy boosted vlad for image retrieval. *Entropy*, 18(8):311, 2016.