Automating the Annotation of Medical Images in Capsule Endoscopy Through Convolutional Neural Networks and CBIR

Rodrigo Fernandes[1,2(✉)], Marta Salgado[3], Ishak Paçal[4], and António Cunha[1,2]

[1] UTAD—University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal
acunha@utad.pt
[2] INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
rodrigo.c.fernandes@inesctec.pt
[3] Centro Hospitalar Universitário de Santo António, 4099-001 Porto, Portugal
martasalgado.gastro@chporto.min-saude.pt
[4] Igdir University, Iğdır, Turkey
ishak.pacal@igdir.edu.tr

Abstract. This research addresses the significant challenge of automating the annotation of medical images, with a focus on capsule endoscopy videos. The study introduces a novel approach that synergistically combines Deep Learning and Content-Based Image Retrieval (CBIR) techniques to streamline the annotation process. Two pre-trained Convolutional Neural Networks (CNNs), MobileNet and VGG16, were employed to extract and compare visual features from medical images. The methodology underwent rigorous validation using various performance metrics such as accuracy, AUC, precision, and recall. The MobileNet model demonstrated exceptional performance with a test accuracy of 98.4%, an AUC of 99.9%, a precision of 98.2%, and a recall of 98.6%.

On the other hand, the VGG16 model achieved a test accuracy of 95.4%, an AUC of 99.2%, a precision of 97.3%, and a recall of 93.5%. These results indicate the high efficacy of the proposed method in the automated annotation of medical images, establishing it as a promising tool for medical applications. The study also highlights potential avenues for future research, including expanding the image retrieval scope to encompass entire endoscopy video databases.

Keywords: Automatic Medical Image Annotation · Convolutional Neural Networks · Content-Based Image Retrieval

1 Introduction

Capsule endoscopy, a recent innovation in gastroenterology, offers a non-invasive visualisation of the gastrointestinal tract, paving the way for improved diagnoses and interventions [1]. However, this advancement brings challenges. The vast amount of video data each examination generates demands extensive manual interpretation, introducing risks of omissions and errors.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 195–207, 2024.
https://doi.org/10.1007/978-3-031-60665-6_15

Artificial intelligence, particularly machine learning, presents a promising solution to these challenges. CNNs, having been applied in various medical contexts from image diagnosis [2, 3] to genomic analysis [4, 5], can automate the identification and categorisation of critical areas in capsule endoscopy, optimising analysis time and accuracy [6].

Our research group has been dedicated to advancements in wireless capsule endoscopy and its automated analysis. In [7], utilising the DenseNet-161 model, we tackled the challenge of manual analysis, achieving significant precision and recall rates in lesion detection. Our work in [8] provided a comprehensive overview of endoscopic capsule technology's evolution, emphasising its advantages over traditional endoscopic procedures. Meanwhile, [9] explored the challenges of abnormality detection in endoscopy videos. We demonstrated effective classification even with limited video capsule endoscopy data by leveraging deep learning and transfer learning.

The efficacy of CNNs largely rests on the availability of vast annotated datasets. Acquiring such datasets in many medical areas, including gastroenterology, poses difficulties due to ethical dilemmas, patient confidentiality, and the data's specialised nature.

The central challenge addressed in this research is the time-consuming and potentially inconsistent nature of manual medical image annotation, particularly in capsule endoscopy. Manual annotation demands expertise and is often subjective, leading to professional variability. Ethical and privacy concerns further constrain the availability of manually annotated datasets.

Traditional content-based image retrieval (CBIR) methods serve as a cornerstone in the quest to search for images within databases based on their inherent visual attributes like colour, shape, and texture, reliant on low-level features, and struggle to grasp the detailed nuances of medical images. This necessitates a novel method to automate the annotation process, integrating deep learning with CBIR to bridge this semantic gap.

The main objective of this study is to develop an automated method for annotating medical images, focusing specifically on capsule endoscopy videos. The research aims to extract and compare visual features from medical images by taking advantage of the capabilities of pre-trained Convolutional Neural Networks (CNN), such as MobileNet and VGG16. Combining deep learning techniques with CBIR methods, this study seeks to eliminate the need for manually annotated datasets, often scarce in medical contexts due to ethical and privacy restrictions. Through rigorous validation involving various performance metrics such as accuracy, AUC, precision and retrieval, the study aims to establish the effectiveness of the proposed methodology for the automatic annotation of medical images.

2 Related Work

Over recent years, CBIR has witnessed exponential growth in theoretical frameworks and practical applications [10, 11]. Despite these advances, the state-of-the-art in CBIR is far from perfect and presents limitations. Some of the most relevant contributions on this topic are highlighted below.

One of the most challenging barriers in the evolution of CBIR systems is the semantic gap disconnect between low-level visual descriptors and the high-level semantic meaning

that humans associate with images [12]. Various strategies have been proposed to mitigate this issue. Djeraba [13], for instance, championed a knowledge-content-based retrieval system that leverages insights gleaned from image repositories, offering a fresh avenue for CBIR systems.

Douze et al. [14] explored the idea of using attribute vectors for image retrieval to further bridge the semantic gap. Their work yielded performance metrics on par with contemporary state-of-the-art methods, suggesting that attribute-based techniques could be valuable to CBIR systems.

In medical imaging, CBIR encounters unique hurdles. Medical images encapsulate spatial and intricate structural data, often poorly represented by traditional, low-level CBIR methods [15]. This highlights an urgent need for CBIR frameworks that can incorporate high-level structural nuances and the spatial relationships between different regions of an image.

The incorporation of machine learning techniques into CBIR has been a game-changing development. Ali & Sharma [16] developed a CBIR system combining feature extraction with machine learning, aiming to improve retrieval accuracy and efficacy.

Efficient indexing mechanisms and relevance feedback are indispensable for optimising CBIR performance. In this context, Jeyasekhar & Mostefai [17] expounded on the importance of state-of-the-art indexing methodologies, such as R-trees and KDB trees, in augmenting the retrieval efficiency of CBIR systems.

Despite the progress of CBIR, the method continues to grapple with issues related to scalability, efficiency, and accuracy. Ouni et al. [18] recently proposed a novel CBIR framework that merges semantic segmentation networks with swift spatial binary encoding, aiming to achieve rapid retrieval at a reduced computational cost.

Despite significant advances in the field of CBIR, there are still gaps that need attention. While many works have focused on mitigating the semantic gap problem, the ideal combination of techniques that can effectively overcome this challenge is still a topic of investigation. Additionally, the application of CBIR in the context of medical imaging, and more specifically capsule endoscopy, presents unique challenges due to the complexity and variety of the data. Automating the annotation of these images is crucial to improving clinical efficiency and diagnostic accuracy. Our work stands out by introducing an innovative approach that combines deep learning and CBIR techniques using convolutional neural networks for the automated annotation of capsule endoscopy images.

3 Methodology

The study's methodology unfolds in several phases, as illustrated in the provided image (Fig. 1):

The images extracted from the original capsule endoscopy dataset undergo a meticulous 'Frame Selection' process to identify pertinent frames based on specific criteria as described in the "Frame Selection Process" section. Subsequently, these frames are adapted and structured during the 'Data Preprocessing' phase, optimising them for enhanced machine learning model performance. Leveraging Convolutional Neural Networks (CNNs), the classifier discerns between similar and non-similar frames, utilising

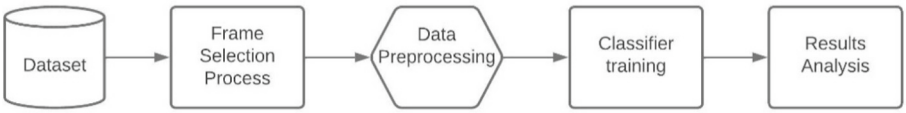

Fig. 1. Flowchart of the experiment.

intrinsic visual features and inter-frame distances. The classifier's proficiency is then rigorously assessed using performance metrics such as accuracy, AUC, precision, and recall, offering a holistic evaluation of the model's capabilities.

3.1 Dataset

This study used a dataset made up of 320 capsule endoscopy videos in GVF format, acquired from a private medical source. All the videos feature patients with various abnormalities detected in the gastrointestinal tract, such as polyps, bleeding, ulcers, among others. These abnormalities were duly noted by medical experts during the review of the videos. It is important to mention that although the images are annotated, they are not categorised into specific classes. The dataset has been anonymised to respect patient privacy and includes 2896 annotated frames with a resolution of 320×320 pixels. To guarantee the quality of the data, the videos went through cleaning and validation processes. During pre-processing, the images were cleaned to remove a black band introduced during extraction, and they were also resized to 224×224 pixels to fit pre-trained networks. In addition, a validation stage was carried out to ensure the correct correspondence between the annotated frames and the originals, enabling accurate annotation of abnormalities.

3.2 Frame Selection Process

For each video in our dataset, we focus on frames annotated by medical experts, referred to as 'target frames'. A pre-trained CNN model first processes these target frames to extract feature vectors. Subsequently, every frame in a video is individually processed to create its feature vector.

We define an interval of 100 frames before and after each target frame to establish a probable correspondence zone. Frames outside this interval are classified as 'not similar'. Using Euclidean distance as a similarity metric, we compare each frame's feature vector with the target frame's.

This generates a ranking of each class's top 10 most similar frames ('similar' or 'not similar'). The procedure is repeated for each target frame across all videos in the dataset. This process is exemplified in Fig. 2 for a more straightforward perception.

Frames Preprocessing
To efficiently integrate the frames with the Convolutional Neural Networks (CNN) models used in this study, it was necessary to implement a preprocessing stage focused on resizing the images. This step is essential, as the CNN models were originally trained

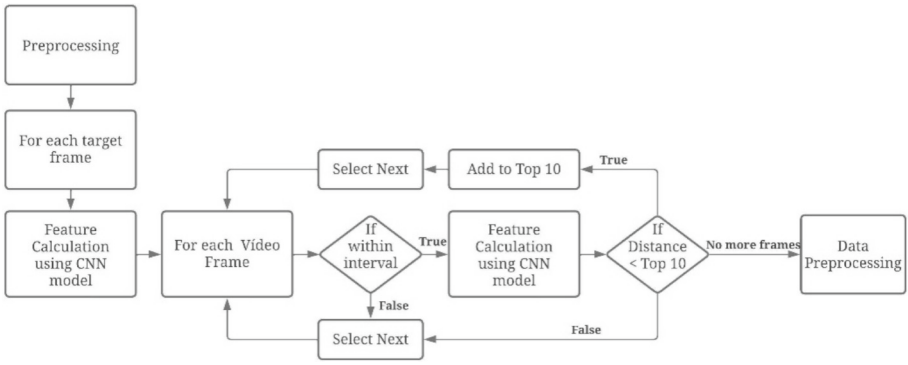

Fig. 2. Flowchart of the proposed frame selection algorithm.

with inputs of specific dimensions. We, therefore, adapted each frame to a standard size of 224 × 224 pixels, thus ensuring compatibility with the neural network architectures.

CNN Models

The previous study used the 25 pre-trained CNN models in the TensorFlow library as feature extractors for localising frames in wireless endoscopic (WCE) capsule videos. These models, which include MobileNet, MobileNetV2, MobileNetV3, ResNet50, ResNet101, ResNet152, ResNet50V2, ResNet101V2, ResNet152V2, ResNet200V2, VGG16, VGG19, DenseNet121, DenseNet169, DenseNet201, InceptionV3, InceptionResNetV2, NASNetMobile, NASNetLarge, EfficientNetB0 to EfficientNetB7, and Xception, have different architectures, providing unique extraction characteristics. Through rigorous preliminary evaluations, considering both accuracy and computational efficiency, eight models were selected for in-depth experimentation: MobileNet, ResNet152v2, VGG19, VGG16, DenseNet121, InceptionResNetv2, ResNet50v2 and ResNet101v2.For the current study, we focused solely on the pre-trained MobileNet model, which demonstrated superior performance in our previous research.

Similarity Metric

A similarity metric quantifies the similarity or dissimilarity between two objects or datasets. In this project, we use the Euclidean distance as the similarity metric.

The Euclidean distance is a widely used metric for calculating the distance between two points in an Euclidean space. It is calculated as the square root of the sum of the squares of the differences between the coordinates of the points. The Euclidean distance Eq. 1 for the two points **p** and **q**, in a space of n dimensions, is given by:

$$dist(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

where p_i and q_i represent the coordinates of the points **p** and **q** in the dimension i, respectively.

The Euclidean distance measures the magnitude of the vector connecting the two points and is used to evaluate their closeness or similarity. The smaller the Euclidean distance, the more similar the points are.

This similarity metric allows us to quantify the distance between pictures with anomalies and identify patterns or similarities, which is critical for analysing and classifying abnormalities in the GI tract.

3.3 Data Preprocessing

The Data Preprocessing section serves multiple purposes. First, it should be noted that the resizing of frames to 224 × 224 pixels was already completed in a previous stage. During this phase, we applied random rotations (0, 90, 180, or 270 degrees) to all images to augment the data. The goal of the data preprocessing step is to classify image pairs into two categories: 'similar' and 'not similar', as illustrated in the following (Fig. 3).

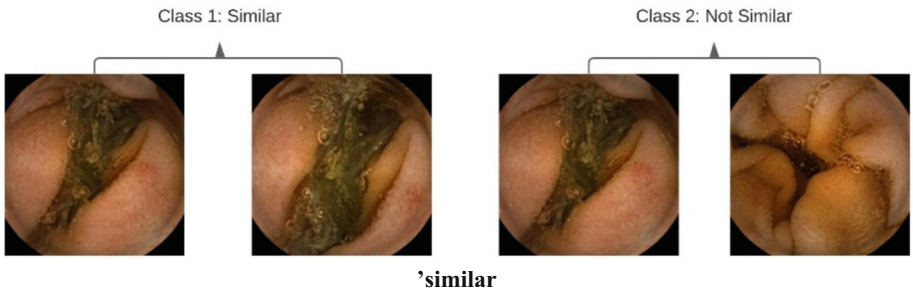

Fig. 3. Data Preprocessing Workflow.

The dataset at this stage contains 57,565 image pairs, comprising 28,885 in the 'similar' category (Class 1) and 28,680 in the 'not similar' category (Class 2). In Class 1, each target frame is paired with one of the top 10 most similar frames, based on the interval of 100 frames defined in the Frame Selection Process section. Conversely, in Class 2, each target frame is paired with one of the top 10 most similar frames but from an interval of 100 frames away from the target frame. This classification process is repeated for each of the top 10 most similar frames for all target frames across the dataset. The dataset was then split into 70%, 15%, and 15% for training, validation, and testing. This classification process defines the classes for the classifier's training.

3.4 Classifier Training

The Classifier Training section outlines a meticulous process for processing each 'image pair. This process is shown in Fig. 4.' Initially, each image in the pair is individually fed into a pre-trained Convolutional Neural Network (CNN) model. Stripped of its original classification layers, this model serves as a feature extractor and generates a feature vector for each image.

After extraction, we concatenate the feature vectors of the two images, creating a composite vector that embodies the image pair's features. Next, this composite vector is flattened, converting it into a one-dimensional array, making it more manageable for the upcoming neural network layers.

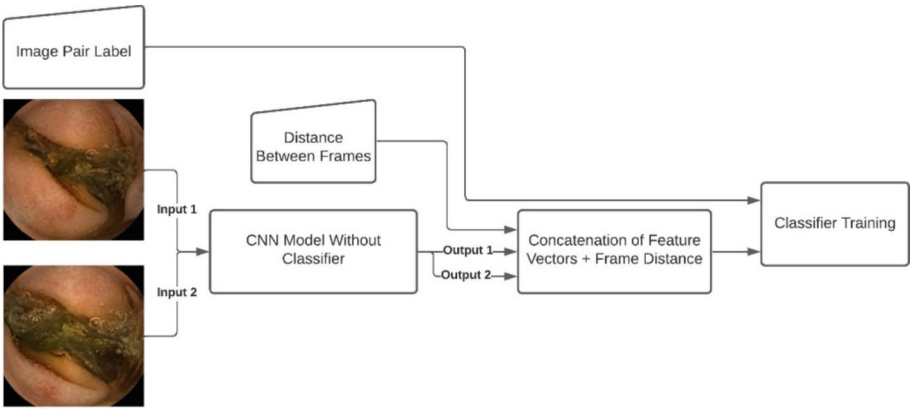

Fig. 4. Classifier Training Flowchart

An additional metric, the frame distance between the two images, is added to the beginning of this one-dimensional vector. Expressed as the difference in the number of frames between the two images (for example, a difference of 10 frames would be represented as 999–989 = 10), this serves as an extra feature for the classifier.

This enriched vector, which now includes both the intrinsic features of the images and the relative frame distance between them, is then provided to the classifier for training, along with a corresponding label indicating their 'similarity' or 'non-similarity.' This comprehensive method ensures the classifier is trained with a rich, multidimensional representation of each image pair, maximising the model's effectiveness and accuracy in the classification tasks.

It's important to note that the classifier's output represents how well the model discerns the degree of similarity sufficient for categorising an image pair within the interval of interest for similarity. This output is vital for the learning objective, empowering the classifier to distinguish between similar and non-similar image pairs effectively.

CNN Models

In this subsection, we conduct experiments with both CNN models, VGG16 [19] and MobileNet [20], to provide a solid basis for comparison. These two models were selected from the eight we tested in our previous study based on their exceptional performance in preliminary tests. Both have been stripped of their original classifier layers and are used exclusively for feature extraction. Using both networks allows for a more robust comparative analysis, reinforcing the validity of our findings.

The VGG16 model, known for its deep and robust architecture, offers the advantage of capturing high-level features, making it highly effective for complex classification tasks. On the other hand, the MobileNet model is known for its computational efficiency, making it an ideal choice for applications that require real-time processing or computational resource limitations. Both models serve as robust feature extractors and complement each other in different aspects, making the analysis more comprehensive.

Training

This subsection presents a detailed description of the classifier's training process, including parameter choices and the model's architecture. We employed the TensorFlow library for code development.

Parameters. Several parameters were adjusted to optimise training:

- Image dimensions: To ensure compatibility with the foundational models, the images were adjusted to a size of 224 × 224 pixels.
- Batch size: We used a batch size of 32 to balance computational efficiency and gradient quality.
- Number of colour channels: The images are processed in RGB format, resulting in 3 colour channels.

Base Model

The base model chosen can be MobileNet or VGG16, both with weights pre-trained on the ImageNet dataset. The layers of this model have been frozen to prevent training during this phase, allowing us to use the features learned in previous tasks.

Additional Layers and Concatenation

After processing by the convolutional layers of the base model, the outputs corresponding to the two input images are concatenated. The result is then flattened to form a one-dimensional vector, adding the distance between the image frames (in number of frames) at the beginning of the vector.

Dense Layers and Activation

The concatenated and flattened vector is passed through a dense layer of 256 units with ReLU activation, which is used to learn more complex representations. Subsequently, a dense layer with a single unit and sigmoidal activation is used for binary classification.

Compilation and Metrics

The model is compiled using the binary cross entropy loss function and the Adam optimiser. Several metrics are monitored during training to evaluate performance, including accuracy, AUC(Area Under the ROC Curve), precision and recall.

Training

The training step is performed over 100 epochs and incorporates an early stopping mechanism to closely monitor the validation loss. If the validation loss does not improve for ten consecutive epochs, training is stopped, and the model weights are restored to the epoch state with the lowest validation loss. We use a custom data generator to provide batches of data to the model during training.

4 Results and Discussion

This section presents the experimental results obtained by applying Convolutional Neural Networks (CNN) in content-based image retrieval (CBIR) for medical images. The "Results" subsection provides a quantitative analysis of the performance metrics achieved by the models, while the "Discussion" delves into the implications and complexities of these results. Together, they offer a cohesive understanding of the study's findings and their relevance in the broader context of medical image analysis.

4.1 Results

This section details the experimental results obtained using the MobileNet and VGG16 models. Both models were evaluated based on five crucial metrics: Test Loss, Test Accuracy, Test AUC, Test Precision and Test Recall.

Table 1 shows the results obtained from training the model as described in the classifier training section:

Table 1. Models result with the distance between frames.

	MobileNet	VGG16
Loss	0.042	0.123
Accuracy	0.984	0.954
AUC	0.999	0.992
Precision	0.982	0.973
Recall	0.986	0.935

The results in Table 1 indicate that MobileNet outperformed VGG16 in all the metrics evaluated. MobileNet recorded a loss of 0.042, an accuracy of 98.4%, an AUC of 0.999, a precision of 98.2% and a recall of 98.6%. On the other hand, VGG16 showed a loss of 0.123, accuracy of 95.4%, AUC of 0.992, precision of 97.3% and recall of 93.5%.

To ensure that the model was not relying excessively on the distance between frames metric, a test was conducted without supplying this metric to the model. Table 2 shows the results of this test:

Table 2. Models result without distance between frames.

	MobileNet	VGG16
Loss	0.581	0.641
Accuracy	0.705	0.6*80*
AUC	0.715	0.700
Precision	0.717	0.697
Recall	0.810	0.793

Analysing Table 2, we see that MobileNet obtained a loss of 0.581, accuracy of 70.5%, AUC of 0.715, precision of 71.7% and recall of 81.0%. In turn, VGG16 showed a loss of 0.641, accuracy of 68.0%, AUC of 0.700, precision of 69.7% and recall of 79.3%. These results indicate a reduction in performance compared to the results where the distance metric was provided but still demonstrate the models' ability to identify similarities between frames based on the extracted visual characteristics.

4.2 Discussion

The main aim of this study was to develop an automated method for identifying similar frames within a specific range using content-based image retrieval (CBIR) techniques. This is a significant step towards reducing reliance on manually annotated datasets, often scarce in clinical settings due to ethical and privacy concerns.

The results show that by using convolutional neural networks (CNNs) such as MobileNet and VGG16 for visual feature extraction, it is possible to achieve high accuracy in identifying similar frames. Notably, the MobileNet model outperformed VGG16 in all the metrics evaluated, highlighting its effectiveness and efficiency in CBIR.

A crucial observation from the results is the impact of the distance between frames metric. When this information was removed during the model's training and testing, a performance reduction was observed, but the metrics remained at acceptable levels. This suggests that although distance information contributes to the model's accuracy, the visual characteristics extracted by the CNN also play a crucial role in determining the similarity between frames.

However, it is worth emphasising that the reduction in performance without the distance metric highlights the importance of combining multiple features and metrics for robust and accurate classification. Combining various features can be the key to achieving reliable and clinically relevant results in medical scenarios where accuracy is critical.

Analysing the following image provides insight into the complexity of identifying similar frames in endoscopy videos (Fig. 5).

The "target frame" represents a specific example of an image in search of matches. We have ten images from each class to the right of this target image. The images in Class 1, identified as similar, were selected based on the smallest Euclidean distance to the reference frame within a specific range. However, not all images in Class 1 manifest the pathology present in the target image, even if they show remarkable visual similarities in terms of texture, colour and shape. This contradiction has significant implications. Without the aid of distance metrics, the model may face challenges when trying to discern between visually similar frames but differ in clinical terms.

On the other hand, images in Class 2, although categorised as different, are not radically different in appearance. Some of these images display visual structures remarkably similar to the target image, even though the overall colour tone may be lighter. This subtle feature can confuse the model, especially when trying to differentiate between frames based on visual characteristics alone, without additional distance information.

Thus, visualising this arrangement of images reiterates the intricate nature of the problem and the need for a model that can consider multiple features and metrics to make sound classification decisions.

4.3 Limitations and Future Work

One limitation is the size and diversity of the data set used. Although we have achieved promising results, exploring how this approach behaves with more extensive and diverse datasets is necessary. In addition, the CBIR technique still needs to be compared

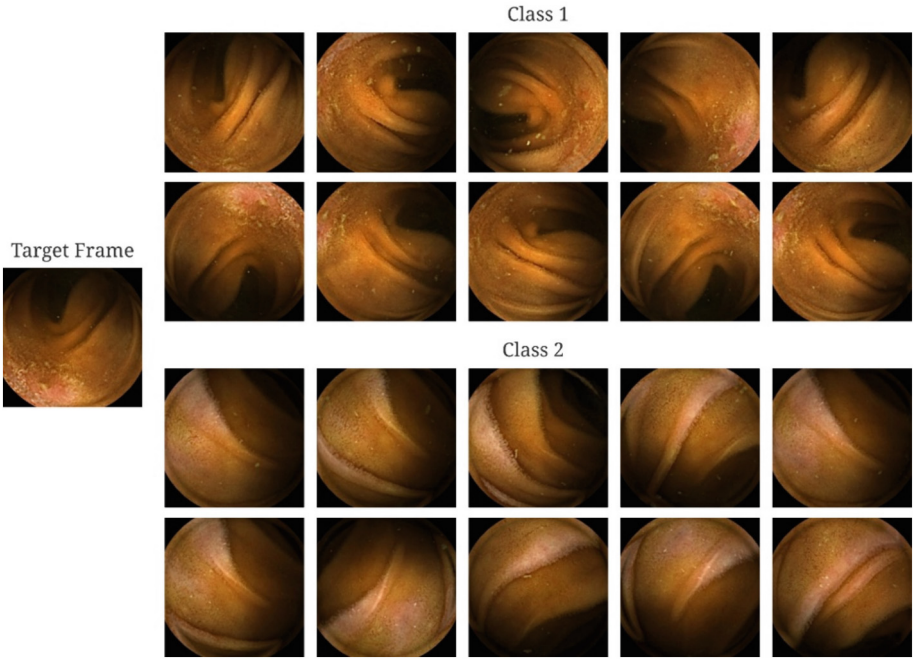

Fig. 5. Comparison between target frame and both class matches.

directly with traditional annotation methods to evaluate its effectiveness and efficiency comprehensively.

In future research, we plan to expand the application of this approach to search for similar images throughout the video, not just in a specific time interval. The aim will be to identify frames that may represent the same medical condition, even if these frames have not been previously annotated. This can reveal undetected cases of similar diseases, thus providing an even more robust tool for automated annotation and diagnosis.

Future research could also explore different network architectures and similarity metrics to improve the method's effectiveness further. Evaluating this method in a natural clinical setting would also be helpful to validate its applicability and effectiveness.

5 Conclusion

This study introduces a novel technique for identifying similar frames in endoscopic capsule videos using content-based image retrieval (CBIR). The approach aims to reduce reliance on manually annotated datasets, a challenge in medical environments due to ethical and privacy concerns. Image features were extracted using the CNN models MobileNet and VGG16, and a classifier was trained, with MobileNet outperforming VGG16 on all metrics.

The implications of this work are vast for automated annotation in medicine. Future plans include expanding this technique to identify similar images throughout the video, which could enhance the detection of unannotated cases of similar medical conditions.

Acknowledgements. This work is financed by National Funds through the Portuguese funding agency,FCT Fundacäo para a Ciéncia e a Tecnologia, within project PTDC/EEIEEE/5557/2020. Co funded by the European Union (grant number 101095359) and supported by the UK Research and Innovation (grant number 10058099). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

1. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature **405**(6785), 417 (2000). https://doi.org/10.1038/35013140
2. Kermany, D., Goldbaum, M., Cai, W., Valentim, C., Liang, H., Baxter, S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122-1131.e9 (2018). https://doi.org/10.1016/j.cell.2018.02.010
3. Yamashita, R., Nishio, M., Gian, R., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights Imaging **9**(4), 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9
4. Ching, T., Himmelstein, D., Beaulieu-Jones, B., Kalinin, A., Brian, T., Way, G., et al.: Opportunities and obstacles for deep learning in biology and medicine. J. Roy. Soc. Interf. **15**(141), 20170387 (2018). https://doi.org/10.1098/rsif.2017.0387
5. Strobelt, H., Gehrmann, S., Pfister, H., Rushton, G.: LSTMVis: a tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Trans. Vis. Comput. Graph. **24**(1), 667–676 (2018). https://doi.org/10.1109/tvcg.2017.2744158
6. Xie, S., Yu, Z., Lv, Z.: Multi-disease prediction based on deep learning: a survey. Comput. Model. Eng. Sci. **128**(2), 489–522 (2021). https://doi.org/10.32604/cmes.2021.016728
7. Lesions multiclass classification in endoscopic capsule frames. Proc. Comput. Sci. **164**, 637–645 (2019). https://doi.org/10.1016/j.procs.2019.12.230
8. Libório, A., Couto, S., Cunha, A., Coelho, P.: Endoscopy—Brief historical survey, developments and therapeutics. In: 2011 IEEE 1st International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–4, November 2011. https://doi.org/10.1109/SeGAH.2011.6165440
9. Fonseca, F., Nunes, B., Salgado, M., Cunha, A.: Abnormality classification in small datasets of capsule endoscopy images. Proc. Comput. Sci. **196**, 469–476 (2022). https://doi.org/10.1016/j.procs.2021.12.038
10. Rui, Y., Huang, T.S., Chang, S.-F.: Image retrieval: current techniques, promising directions, and open issues. J. Vis. Commun. Image Represent. **10**(1), 39–62 (1999)
11. Sugamya, K., et al.: CBIR using SIFT & FDCT with relevance feedback mechanism. Int. J. Innov. Technol. Explor. Eng. **8**(11), 1103–1108 (2019). https://doi.org/10.35940/ijitee.j1193.0981119
12. Khodaskar, A., Ladhake, A.: New-fangled alignment of ontologies for content based semantic image retrieval
13. Djeraba, C.: Association and content-based retrieval. IEEE Trans. Knowl. Data Eng. **15**(1), 118–135 (2003)
14. Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. In: CVPR 2011. IEEE (2011)
15. Sharma, H., et al.: Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics. Diagn. Pathol. **7**(1), 1–20 (2012)

16. Ali, A., Sharma, S.: Content based image retrieval using feature extraction with machine learning. In: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE (2017)
17. Jeyasekhar, S., Mostefai, S.: Towards effective relevance feedback methods in content-based image retrieval systems. Int. J. Innov. Manag. Technol. **5**(1) (2014)
18. Ouni, A., Chateau, T., Royer, E., Chevaldonné, M., Dhome, M.: A new CBIR model using semantic segmentation and fast spatial binary encoding. In: Nguyen, N.T., Manolopoulos, Y., Chbeir, R., Kozierkiewicz, A., Trawiński, B. (eds.) ICCCI 2022. LNCS, vol. 13501, pp. 437–449. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16014-1_35
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, 10 April 2015. https://doi.org/10.48550/arXiv.1409.1556
20. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications, 16 April 2017. http://arxiv.org/abs/1704.04861. Accessed 25 Sept 2023