Image-to-Video Retrieval by ResNet50

Chomtip Pornpanomchai*

[1]*Faculty of Information and communication technology, Mahidol University, Phuthamonthon 4 road, Salaya, Nakhorn Pathom, Thailand, 73170. E-mail: chomtip.por@mahidol.ac.th*

Abstract: The objective of this research is to create a computer system which can retrieve a video clip by using only a single image. The developed system is called “Image-to-Video Retrieval System (I2VRS)”. The system employs the convolutional neural networks called “ResNet50”, which is a toolbox in MATLAB software to retrieve the video clip dataset. The ResNet50 is one of the powerful CNN to recognize an image in the image processing technique. The I2VRS creates its own dataset called I2VRS dataset, which consists of 101 video clips and each video clip contains 1,000 video frames. All video clips are filmed around 60 s. each in the .MP4 file-format. The system also tests an un-training dataset with 100 images, which are directly taken with a mobile phone in the .JPEG file-format. The mean average precision (mAP) of the system is 0.9825, with the training dataset time being 5,668.7 s. The average access time to retrieve a video clip is 1.5726 s. per image.

Keywords: Confusion Matrix, Convolutional Neural Networks, Image-to-Video Retrieval, Mean Average Precision (mAP), ResNet50

1. INTRODUCTION

According to the fast development of mobile-phone camera, internet bandwidth, internet speed, disk capacity and social media applications are make people upload and download with a huge of video clips daily. Qiu, G. estimated more than 3 billion images with 700,000 h. of videos transmitted daily across the network [1]. Suppose, people want to know about someone, some place, some food, some machine instruction etc. They can manually retrieve an interesting video from the internet by their mobile phone or computer system. It is a time consuming and many human errors when they browse internet for retrieving interesting video. Therefore, many researchers and scientists try to develop computer system to help people to retrieving an interesting video. There are three main methods to retrieve the social network video clips, which are a context-based, an image-based, and a short video-based [2]. Each video retrieval technique has the following details.

1.1 Content-Based Video Retrieval (CBVR)

The CBVR method queries video dataset by two techniques, which are an annotation-based-image-retrieval (ABIR) and a content-based-image-retrieval. The first method uses text-mode to retrieve the video in a dataset. The second method uses contents in video frame, which are color, shape, texture to retrieve the video in a dataset [3]. Many researchers conducted on many experiments with many methods, which have the following brief details.

1.1.1 Annotation -Based-Video Retrieval (ABIR)

Gao and Xu presented the text-to-video retrieval in two datasets, which are the Charades-STA and Activity-Caption. The system designed a video-conditioned sentence generator to produce a suitable sentence match to the video. Then, the image-sentence-embedding-space part generated relevant text and video score for querying video clip [4]. Dong et al. proposed dual encoding network jointing both video and natural language queries into powerful dense representation. The experiment conducted on three video datasets, which are MSR-VTT, TRECVID 2016, and 2017 AVS [5]. Yasin et al. presented semantic video retrieval by using person name and the FaceNet to retrieve video clip based on the person-name query. Their experiments conducted on 50 videos from YouTube [6].

1.1.2 Content-Based-Video Retrieval (CBVR)

Tseytlin and Makarov illustrated a content-based-video retrieval by employing the temporally informative representative image (TIRI) for weighting and extracting video frames. After that, the system used modified image fingerprinting measure called “Quadrant rHash vectors” to retrieve the video. The experiment conducted on 519 videos with 27 hours [7]. Li and Ma employed scale-invariant feature transform (SIFT) to extract locale features of video-frames and Fisher vector (FV) to describe global features of video-clips. The research used a convolutional neural network to train and retrieve the video and the experiment conducted on 3 datasets, which are Flickr60K, ImageNet and Standford12V [8]. Mallick and Mukhopadhyay employed color co-occurrence feature (CCF) to represent the video-pixels and used graph-based matching to query videos. The experiment conducted on HMDB51 and UCF11 datasets [9]. Amayo et al. introduced a visual content-based retrieval system called “VISIONE”. The system can query video by keywords, object-location, color-location and visual-example. The system analyzed on the video browser showdown (VBS) dataset [10].

1.2 Image-Based Video Retrieval (IBVR)

The IBVR technique treats every video frames as an individual image and indexing them for matching between a query-image and an individual image. By using IBVR technique, people no need to create any keywords, any video description for query interesting video. Many researchers applied various image-to-video retrieval techniques, which have the following brief details.

Liu proposed a motion-assisted activity proposal-based image-to-video retrieval (MAP-IVR) approach to query a video. The developing system tested on Thumos14 and ActivityNet datasets [11]. Zhang et al. combined convolutional neural network (CNN) with bag of visual word (BoVW) to represent video-frame features. They proposed visual weighted inverted index (VWII) to retrieve the YouTube and Sports-1M datasets [12]. Yuan et al. proposed hash-center concept to create similar data pairs scattered in the Hamming space with the sufficient mutual distance between an image and video frame. The experiment conducted on five datasets, which are ImageNet, MS COCO, NUS_WIDE, UCP101 and HMDB51 [13]. Araujo and Girod proposed an asymmetric comparison technique for Fisher vectors and systematically explore query with varying amounts of cluster. They constructed two dataset Asym-QCD and Asym-DCQ to conduct the experiment [14]. Song et al. illustrated a quantization-based hashing (QBH) to create hash-code and query video by apply the Hamming distance method. The experiment conducted on three datasets, which are SIFT1M, GIST1M and SIFT1B [15].

1.3 Video-based video retrieval (VBVR)

Normally, the VBVR technique tries to retrieve near-duplicate video in the video datasets. The VBVR aims to find the most similar content from the both videos. Many researchers applied many techniques for VBVR, as the following brief details.

Jo et al. illustrated the ISO/IEC MPEG standardizing called “compact descriptors for video analysis (CDVA)”. They applied temporal nested invariance pooling (TNIP) to compare the most similar videos for video-to video retrieval [16]. Liu et al. applied the bag-of-words (BOW) model to measure the video-frames similarity and used relative edit distance similarity (REDS) to filter non-near duplicate video for video-to-video query [17]. Yu et al. showed the surgical video-to-video retrieval in the operation room. They used two hashing function models to train and retrieve Cholec80 and CEV64 surgical dataset [18].

According to the previous video retrieval researches, many researchers employed various datasets namely, MSR-VTT, TRECVID 2016, 2017AVS, HMDB51, UCF11, UCF101, YouTube, Sports-1M, Thumos14, ActivityNet ImageNet, MS COCO, NUS_WIDE, SIFT1M, GIST1M and SIFT1B, Cholec80 and CEV64, etc. Moreover, they applied many methods namely, SIFT, BoVW, QBH, MAP-IVR, CNN, etc. The convolutional neural network (CNN) is one technique, which researchers employed for training and matching the video [2][5][10][11][12]. There are many types of CNN and no consensus, which CNN is the best one. Deshpande illustrated GoogleNet, AlexNet, VGG16, ResNet50 and Inception V.3 to classify 2,900 brain tumor images. The accuracy rate of the training dataset for

each CNN model is shown in Table 1 and the ResNet50 is shown to have the highest accuracy [19]. Based on this research, the I2VRS adopts the ResNet50 to conduct the experiment.

Normally, the ResNet50 consists of three main components, which are feature extraction, classification and output classification component, as shown in Figure 1. The operations of the ResNet50 have the following steps. First, getting an input-image, then a feature extraction component repeats in convolution and pooling layer until it extracts all the input-image features. Second, the fully connection component uses all features for neural network input-layer and classify input-image to show to an output-layer. Finally, the output classification component displays the most similar between input-image and image in dataset.

Table 1 CNN accuracy comparison

CNN Model	accuracy rate (%)
GoogleNet	81.67
AlexNet	91.84
VGG16	93.06
ResNet50	**98.14**
Inception V.3	90.79

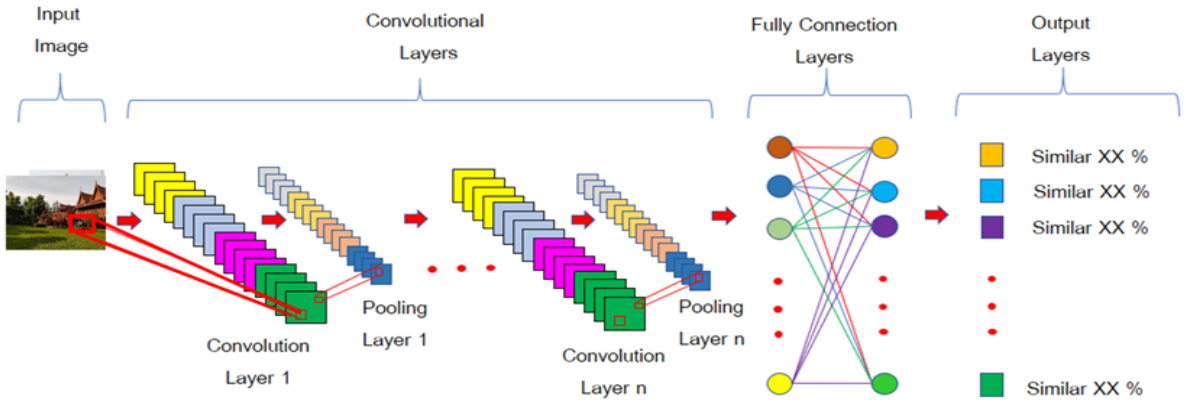

Figure 1 The convolutional neural network architecture

Based on the previous video retrieval researches, the image-to-video retrieval is a simple and easy for people to query video clip because it is easy for them to take a picture by their mobile phone and use that mobile phone to query the most similar between an input-image and a video clip in internet. The contribution of this research is to develop a computer system call “I2VRS”, which can get an input-image and search the most similar between an input-image and an output-video clip from a dataset. Finally, the I2VRS can play and pause the video clip. The architecture and methods of I2VRS has the following details.

2. MATERIEL AND METHODS

The I2VRS was developed on the following computer hardware and software. The Intel(R) Core (™) i7-1195G7 CPU @ 2.90 GHz with 16 GB RAM was used as the central processing unit and Windows 11 was the operating system. MATLAB R2020b with license number 40598465 was the developing software. The digital cameras used in this research were Xiaomi, Redmi 8 to take all the video clips. To avoid the copyright problem, this research creates its own video dataset by taking 101 campus-life videos around Mahidol University, Salaya campus and each video-clip contains more than 1,000 video-flames for conducting the experiment.

2.1 Conceptual Diagram

The I2VRS conceptual diagram starts with a user capturing an interesting image by using a simple mobile phone

camera. Then the image is submitted to the computer system for retrieving a relevant video clip. The I2VRS retrieves the video clip by matching an image with all video-frames in a training-dataset. Finally, the system displays the retrieval result to the user, as shown in Figure 2.

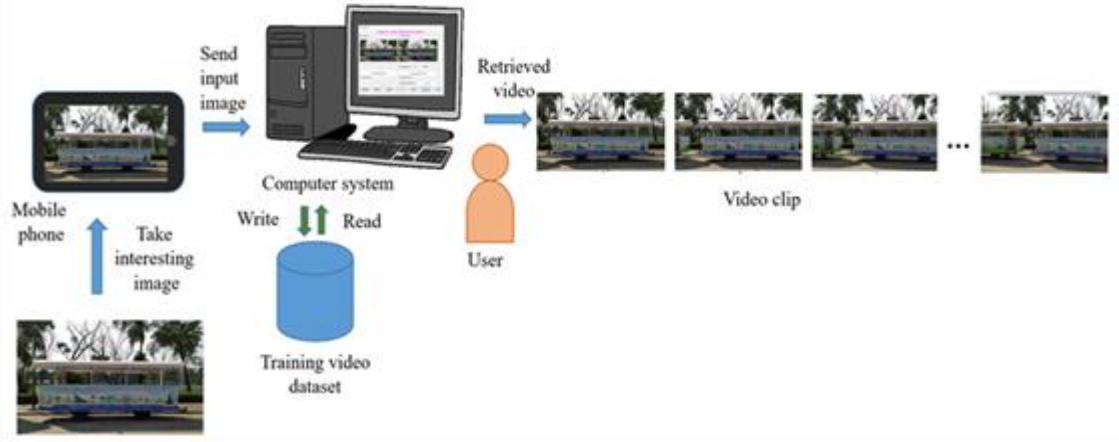

Figure 2 The I2VRS conceptual diagram

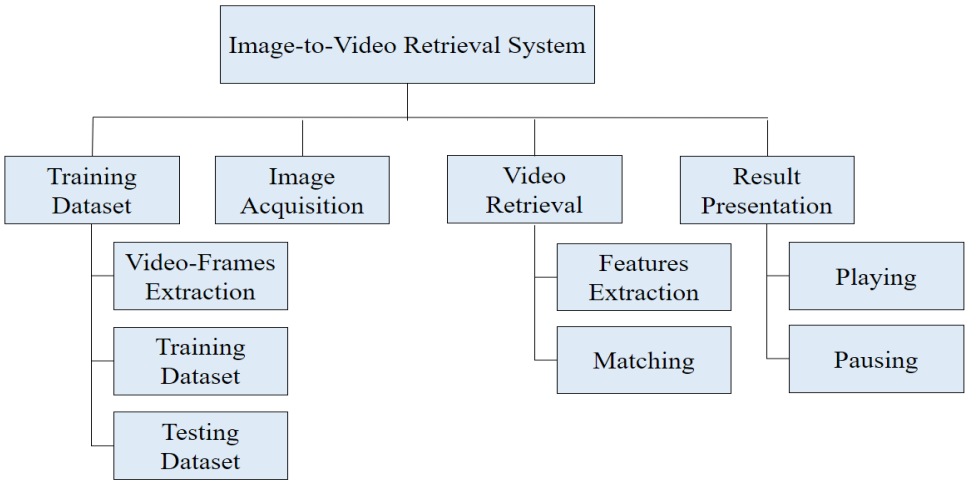

Figure 3 The I2VRS structure chart

2.2 System Structure Chart

The I2VRS structure chart consists of four main modules, namely: 1) training dataset module, 2) image acquisition module, 3) video retrieval module, and 4) result presentation module, as shown in Figure 3. The result presentation module contains two submodules, which are 1) play a video and 2) pause a video submodule. Each I2VRS module has the following details.

2.2.1 Training Dataset Module

Regarding the I2VRS research, the I2VRS employed ResNet50 to train and test dataset. The ResNet50 is famous and commonly used to implement image recognition without human supervision [20]. The I2VRS consists of 101 video clips and each video clip taking around Salaya campus, Mahidol university. The video is taken in the full HD system with the resolution of 1920 x 1080 x 3 (pixel-width x pixel-height x plane) in the .MP4 file format. Therefore, the size of extracting video-frame is 1080 x 1920 x 3 pixels. Normally, the video clip contains 32 frames/second. There are 1,920 video-frames if the system captures every frames in a minute video-clip (32 x 60 = 1,920). The I2VRS captured only 1,000 video-frames in each video-clip. The I2VRS employed 101 video-clips with

1,000 frames each. Totally, the I2VRS conducted the experiment with 101,000 video-frames. The I2VRS randomly selected 800 video-frames for a training dataset and the remaining 200 video-frames for an un-training dataset. The total number of training dataset is 80,800 frames (101 x 800) and un-training dataset is 20,200 frames (101 x 200). The training dataset used for training the ResNet50 and the un-training dataset used for validating the ResNet50 of this system. The training dataset module consists of three sub-modules, which are video-frames extraction, training and testing sub-modules. Each sub-module has the following details.

2.2.1.1 Video-frame extraction sub-module

The I2VRS extracted the video-frame by capturing and saving every frames in the video clip in a JPEG format. The system stored 1,000 extraction JPEG-files in one folder. Therefore, the I2VRS consists the 101 video-clips in one folder and the 101 extraction files in another folder. Both folders are directly mapping.

2.2.1.2 Training Sub-Module

For the 80,800 video-frames, the system also separated into two parts, which are training part and testing part. The training part is randomly selected 80% of video-frames (80,800 x 0.8 = 64,640 frames) for the training part while the remaining 20% of video-frames (80,800 x 0.2 = 16,160 frames) of the testing part. The I2VRS employed 64,640 video frames to train the ResNet50.

2.2.1.3 Testing Sub-Module

The I2VRS employed 16,160 video-frames for evaluating the ResNet50. The MATLAB command to get a testing image is ‘testImage’ and the command to retrieve the video-frame is ‘predictedLabel’ Both commands are in MATLAB ResNet50 tool box.

2.2.2 Image Acquisition Module

There are two ways to retrieve a video clip by using a single image, which are first, directly taking an image from user mobile phone and second, select a video-frame from the dataset. The system is taken the picture with mobile phone camera with the full HD image of size 1920 x 1080 x 3 (pixel x pixel x plane) in three planes. The ResNet50 is resized the full HD image to the ResNet50 suitable size, which is 224 x 224 x 3 pixels. The I2VRS uses ResNet50 image-size to retrieve the most similar video-clips for the answer.

2.2.3 Video Retrieval Module

The I2VRS employs the ResNet50 to retrieve video clip. The structure of the ResNet50 in the I2VRS contains three components, namely 1) feature extraction component, 2) classification connected component and 3) output component, as shown in Figure 4. The video retrieval module consists of two sub-modules, which are features extraction and retrieval sub-module. Each sub-module has the following details.

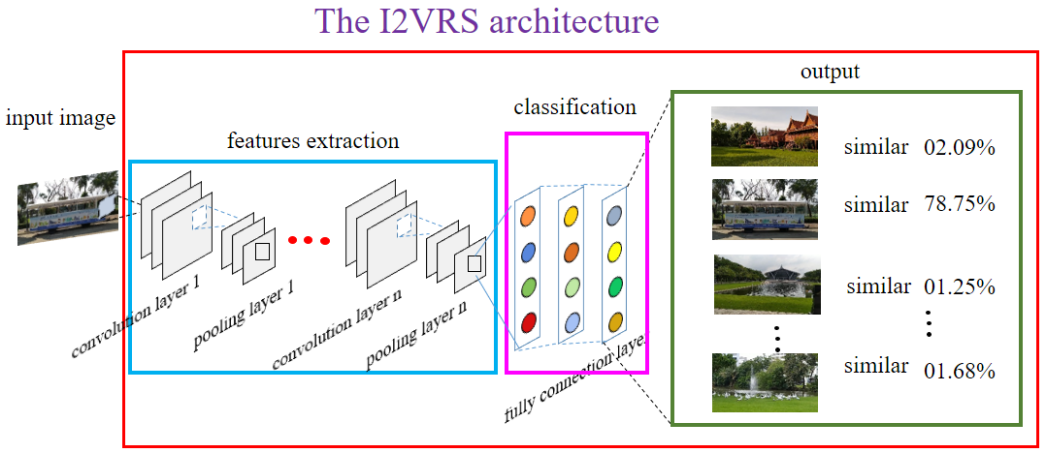

Figure 4 The architecture of ResNet50 in this research

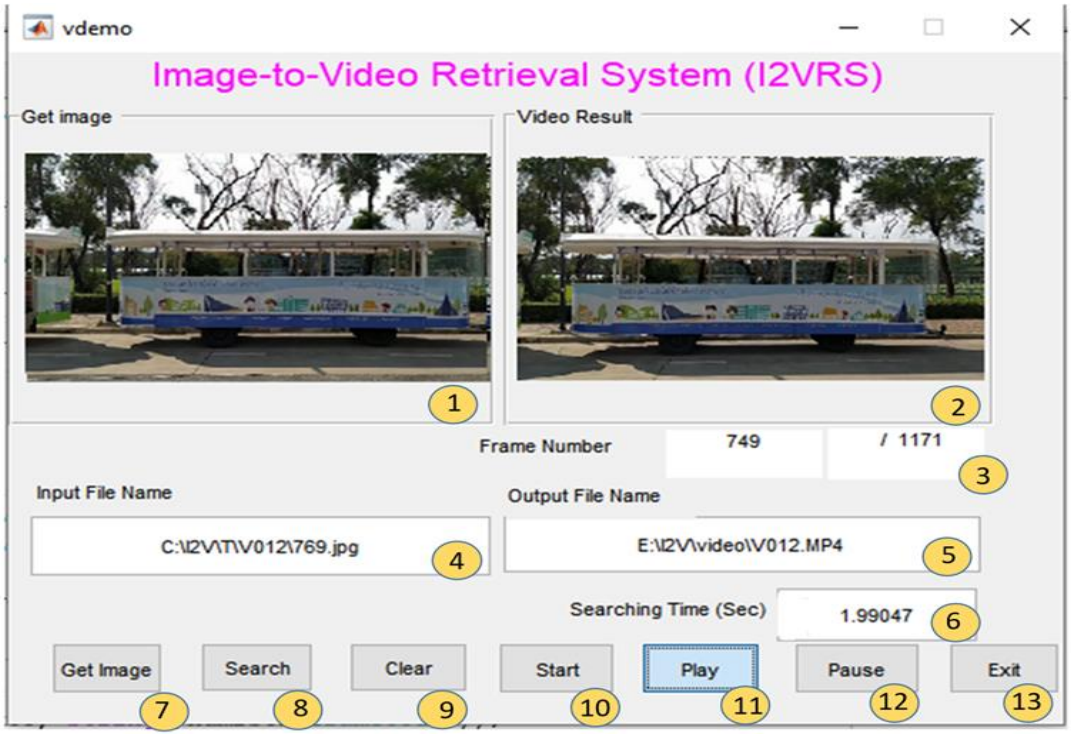

Figure 5 The GUI of the I2VRS

2.2.3.1 Feature Extraction Sub-Module

The I2VRS starts by receiving the input-image from users, after that it sends an input-image to the feature extraction component. The features extraction component consists of two sub-components, which are convolutional layer and pooling layer sub-component. The convolutional layer extracts all of the important features from an image and pooling layer reduces the dimension from the convolutional layer. The ResNet50 repeats the convolutional and pooling layer 50 loops to extract all features from the input-image.

2.2.3.2 Retrieval Sub-Module

The classification component consists of fully connection neural-network. The classification component matches the most similar between input-image and video-frames from the dataset and sends the matching result to the output component. Finally, the output component displays the retrieval video-clip to the user.

2.2.4 Result Presentation Module

This module shows the I2VRS queries result. The I2VRS graphic user interface (GUI) is composed of two display-graphic windows, four display-text boxes and seven push buttons, as shown in Figure 5.

The two display-graphic windows have the following details:

1) In the label 1 of Figure 5, the display-graphic window for showing the input-image.
2) In the label 2 of Figure 5, the display-graphic window for showing the retrieval video-clip.

The four display text boxes have the following details:

1) In the label 3 of Figure 5, the display of the frame number and total frame number of the displaying video clip.
2) In the label 4 of Figure 5, the display of the input path and filename box.
3) In the label 5 of Figure 5, display of the retrieval video filename box.
4) In the label 6 of Figure 5, the display of the average retrieval time box

The seven push buttons have the following details:

1) In the label 7 of Figure 5, the get image button for getting the input image.
2) In the label 8 of Figure 5, the search button for matching input-image with all video-frames in the dataset.
3) In the label 9 of Figure 5, the clear button for clearing all the I2VRS values.
4) In the label 10 of Figure 5, the start button for playing video clip from the starting frame.
5) In the label 11 of Figure 5, the play button for playing video clip from the current frame.
6) In the label 12 of Figure 5, the pause button for pausing video clip from the current frame.
7) In the label 13 of Figure 5, the exit button for exiting the system

3. RESULTS AND DISCUSSIONS

3.1 Experiment Results

The I2VRS employed 80,800 video frames to train the dataset and used 20,200 video frames for validation of the ResNet50. The training time for I2VRS dataset was 5,804.4 s. or 1 h. 36 m. and 100.4 s. The confusion matrix for training the ResNet50 had the true positive (TP) of 10,095 (101 x 101 x 0.9896), false positive (FP) of 106 (101 x 101 x (1-0.9896)), false negative (FN) of 106 and true negative (TN) of 1,019,994 (101 x 101 x 101 - 10,095 -106 - 106), as shown in Table 2. Moreover, the confusion matrix for validation of the ResNet50 had the true positive (TP) of 10,201 (101 x 101 x 1.00), false positive (FP) of 0 (101 x 101 x 0.0), false negative (FN) of 0 and true negative (TN) of 1,020,100 (101 x 101 x 101 - 10201), as shown in Table 3. Finally, this research evaluated the system by using unknown pictures, which were taken in the 100 JPEG files. The unknown pictures were directly taken by using a simple mobile phone, which had the scene related to the video dataset. The experimental result found 97 pictures matching and 3 pictures miss-matching. The confusion matrix for the unknown dataset had TP of 9,895 (101 x 101 x 0.97), FP of 306 (101 x 101 x 0.03), FN of 306 and TN of 1,019,794 (101 x 101 x 101 – 9805 – 306 - 306), as shown in Table 4. The average access time to retrieve a video clip is 1.5726 s. / image.

Table 2 The confusion matrix for training the ResNet50

		Actual Class	
		True	False
Prediction Class	Positive	10,095 (TP)	106 (FP)
	Negative	106 (FN)	1,019,994 (TP)

Remarks: TP = true positive; FP = false positive; FN = false negative; TN = true negative

Table 3 The confusion matrix for ResNet50 validation

		Actual Class	
		True	False
Prediction Class	Positive	10,201 (TP)	0 (FP)
	Negative	0 (FN)	1,020,100 (TP)

Remarks: TP = true positive; FP = false positive; FN = false negative; TN = true negative

Table 4 The confusion matrix for ResNet50 unknown dataset

		Actual Class	
		True	False
Prediction Class	Positive	9895 (TP)	306 (FP)
	Negative	306 (FN)	1,019,794 (TN)

Remarks: TP = true positive; FP = false positive; FN = false negative; TN = true negative

The accuracy graph and loss graph for training ResNet50 are shown in Figure 6 (a) and 6 (b), respectively. The setting parameters for training ResNet50 have an epoch per iteration, 0.01 learning rate and 32 maximum epochs.

Figure 6 The accuracy graph and loss graph for training ResNet50 in the I2VRS

The I2VRS used MATLAB multi-classes confusion-matrix function called “confusion.getMatrix” to calculate the six statistical values to measure I2VRS performance, which are accuracy, precision, recall, F1-score, average precision (AP) and mean of average precision (mAP) value. All statistical values calculated from the confusion matrix, which have true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Each statistical value has the following details [21] [22].

3.1.1 Accuracy Statistical Value

An accuracy value is one of the most common measurements of matching performance and it is defined as the ratio between the correct number of video-frames matching and the total number of video-frames, as shown in Equation 1.

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (1)$$

3.1.2 Precision Statistical Value

A precision value is an ability of an image to identify only the relevant videos within a dataset. The precision value find by using the number of true positives divided by the number of true positives plus the number of false positives, as shown in Equation 2.

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

3.1.3 Recall Statistical Value

A Recall value is an ability of an image to find all the relevant videos within a dataset. The recall value is defined as the number of true positives divided by the number of true positives plus the number of false negatives, as shown in Equation 3.

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

3.1.4 F1-Score Statistical Value

The F1-score considers both the precision and the recall of the test to compute the score. The F1 score can be interpreted as a weighted average of the precision and recall, as shown in Equation 4.

$$\text{F1-score} = 2 \times [(\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})] \quad (4)$$

3.1.5 Average Precision (AP) Statistical Value

The average precision is a measure that combines precision and recall for ranked retrieval results. The relevant video-clips that are ranked higher contribute more to the average than the relevant video-clips that are ranked lower. The average precision score is calculated by using Equation 5.

$$AP = \sum_{k=1}^{N} \text{Precision}(k) \times \Delta \text{Recall}(k) \quad (5)$$

Where N = total number of video-frames to test the dataset

Δ Recall (k) = value change between Recall(k) to Recall(k+1) video-frames

3.1.6 Mean Average Precision (mAP) Statistical Value

The mean average precision uses to measure the accuracy of information retrieval models. The mAP has a value between 0 - 1 with higher scores representing a more accurate model. The mAP can calculated by using Equation 6

$$mAP = = \left(\sum_{i}^{M} APi \right) / M \quad (6)$$

Where M = total number of video-clips to test the datase

The system conducted the experiments by using the different interval between the skipping video-frames, which are skipping every one-frames (1, 3, 5, 7, …), skipping every two-frames (1, 4, 7, …) until skipping every ten-frames [1, 12, 23, …). The statistical measurement values for evaluating the I2VRS are accuracy, precision, recall, F1-score and training time, as shown in Table 5. Moreover, the system is plotted the precision and recall graphs, as shown in Figure 7. The mean average precision (mAP) of the I2VRS is 0.9825, which is calculated by Equation 6. Based on the I2VRS experiment results, the matching and mismatching between image and video are illustrated on Figure 8 (a) and 8 (b), respectively.

Table 5 Comparison performance between different video-frames interval

Video Interval (Frames)	Accuracy	Precision	Recall	F1-score	Training Time (Sec.)
1	0.9896	0.9896	0.9896	0.9896	5,668.7
2	0.9894	0.9894	0.9894	0.9894	5,267.2
3	0.9868	0.9864	0.9868	0.9863	4,938.7
4	0.9865	0.9862	0.9865	0.9860	4,826.5
5	0.9864	0.9847	0.9864	0.9856	4,664.7
6	0.9862	0.9842	0.9862	0.9841	966.8
7	0.9859	0.9856	0.9859	0.9857	785.5
8	0.9855	0.9854	0.9855	0.9853	716.3
9	0.9802	0.9795	0.9802	0.9799	628.2
10	0.9800	0.9764	0.9800	0.9783	589.9

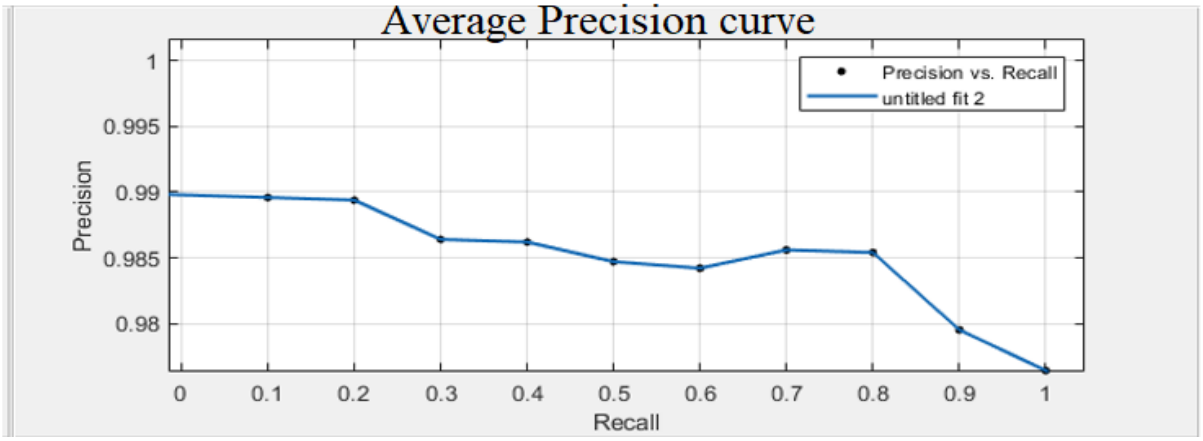

Figure 7 The precision and recall graph

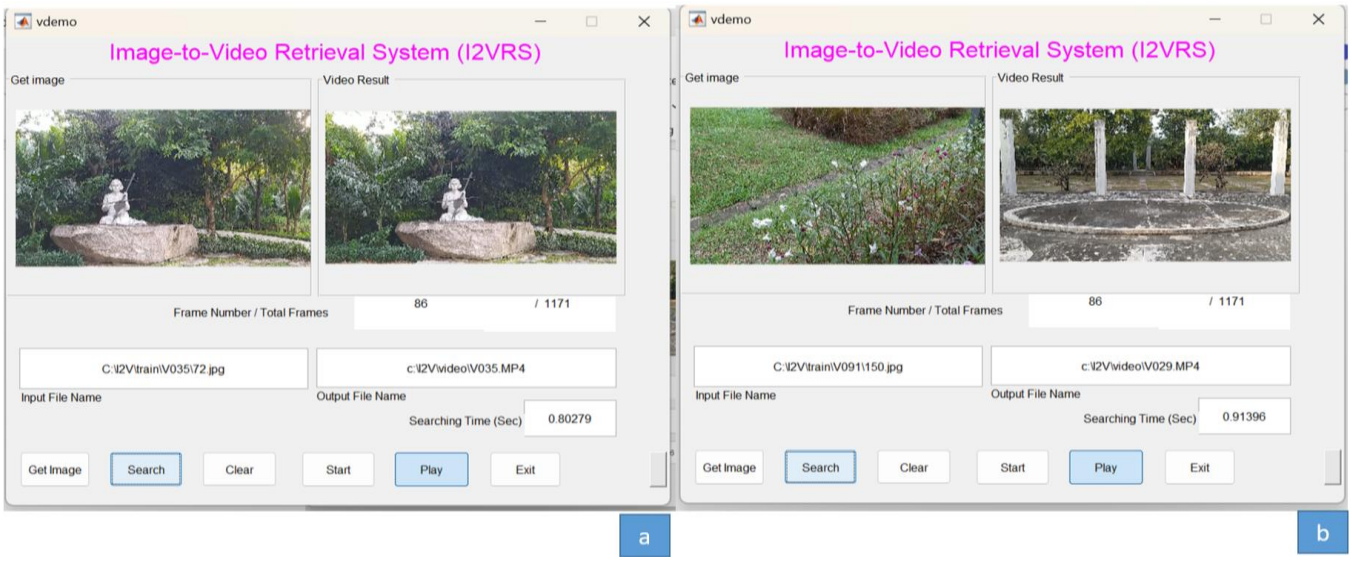

Figure 8 The GUI samples of (a) correct matching result and (b) incorrect matching result

3.2 Comparison Performance Between Convolutional Neural Networks

This research was conducted on 7 different CNN types, namely: AlexNet, GoogleNet, InceptionV3, ResNet18, ResNet50 and ResNet101 and VGG16 with the same video dataset. These 7-CNN types are supported by the MATLAB 2020b version. The comparison of accuracy, precision, recall, F1-score and training times is shown in Table 6. The AlexNet gives the shortest time and VGG16 gives the longest time for training dataset. The ResNet101 give the lowest measurements and the ResNet50 give the highest measurements. Therefore, this research employed the ResNet50 for the I2VRS.

Table 6 Performance and training time comparison.

CNN model	Accuracy	Precision	Recall	F1 score	Training Time
AlexNet	0.7079	0.7812	0.7079	0.7427	1,384.0
GoogleNet	0.9854	0.9853	0.9854	0.9854	5,971.5
InceptionV3	0.9879	0.9873	0.9879	0.9876	7,429.0
ResNet18	0.9852	0.9845	0.9852	0.9848	2,957.6
ResNet50	**0.9896**	**0.9896**	**0.9896**	**0.9896**	5,202.5
ResNet101	0.6552	0.7290	0.6552	0.6901	7,887.3
VGG16	0.7505	0.7666	0.7505	0.7585	14,0270

3.3 Comparison Performance Between Previous Researches

There are many researchers conducted many experiments on video retrieval by many methods with many kinds of dataset, as shown in Table 7. All previous research employed different dataset with different techniques. The state-of-art for the video retrieval research have the mean average precision (mAP) between 0.80 to 0.98.

Table 7 Performance comparison between previous video-retrieval researches.

Author	Year	Dataset Name	mAP	Methods
Liu [17]	2017	CC_WEB_Video, CBCD2011	0.980	Bag of word (BOW) and Relative edit distance similarity (REDS)
Araujo [14]	2018	SI2V-4M, VB-4M	0.760	Relative edit distance similarity (REDS)
Zhang [12]	2019	Youtube8M, Sport-1M	0.800	asymmetric comparison technique for Fisher Vectors
Yuan [13]	2020	UFC101, HMDB51	0.880	Bag of Visual word (BoVW) and AlexNet
Liu [11]	2021	THUMOS14, ActivityNet	0.304	Central similarity Quantization (CSQ)
Mallick [9]	2022	UCF11, HMDB51	0.925	Activity-based image-to-video retrieval (AIVR)
Jo [16]	2022	FIVR-200K	0.870	Color co-occurrence feature (CCF) and Graph-based matching
Current Study	2024	I2VRS	0.985	Compact descriptors for video analysis (CDVA)

3.4 Discussion

The I2VRS fulfills the objective of this research, which is to develop a computer system to retrieve video-clips by using a single image. The system training 80,800 video-frames, which are 101 video-video clips with 800 video-frames per a video-clip. The I2VRS validation with 20,200 video-frames, which are 101 video-clips with 200 video-frames per a video-clip. Moreover, this research also tested with un-training 100 picture, which directly taken by mobile phone. The I2VRS conducted the experiments by employed the ResNet50.The results of the experiment are followed the video-retrieval state-of-art. The I2VRS future works are create bigger dataset and develop the system, which can retrieve video-clip via the internet networks.

3.5 System limitation

There are two limitations of the image-to-video retrieval system, namely: 1) all the systems need to train the dataset before used the system to retrieve the video clips. The training dataset processes are time consuming processes. The huge dataset needed more time to train the dataset more than a small dataset and 2) an unknow image in dataset, the CNN will find the best matching image but it is a wrong answer. The CNN cannot show an unknow image to the GUI.

CONCLUSION

The I2VRS fulfills the objective of this research, which is to develop a computer system for retrieving video clips by using a single image. The system has trained the I2VRS dataset, which are 101 video clips, with 800 video frames per a video clip. The I2VRS performs validation with 20,200 video frames, which are 101 video clips, with 200 video frames per a video clip. Moreover, this research has also tested on un-training 100 JPEG pictures, which were directly taken by a simple mobile phone. The I2VRS has conducted the experiments by employing the ResNet50. The experimental results reveal a state-of-the-art video retrieval performed by the system developed. The future works of this system are to enlarge the I2VRS dataset by adding more interesting spots on the campus or adding outside campus video clips. Moreover, developing the I2VRS to receive and display results via the internet networks.

Acknowledgements

The author would like to thank the Faculty of Information and Communication Technology (ICT), Mahidol University in supporting this research.

REFERENCES

[1] Qiu G. Challenges and opportunities of image and video retrieval. Front Imaging, 2022; 1:951934. http://doi: 10.3389/fimag.2022.951934

[2] Qi, B. Design and implementation of multimodal video retrieval system. Proceedings of the 5th Annual International Conference on Data Science and Business Analytics (ICDSBA), 2021 September 21-24. Changsha, China, IEEE explore; pp.131-134.

[3] Wankhede VA, and Mohod PS. Content-based image retrieval from videos using CBIR and ABIR algorithm. Proceedings of 2015 Global Conference on Communication Technologies (GCCT 2015), 2015 April 23-24. Tamil Nadu, India, IEEE explore; pp. 767-771.

[4] Gao J, Xu C. Learning video moment retrieval without a single annotated video. IEEE Transactions on circuits and systems for video technology, 2022; 32(3): 1646-1657.

[5] Dong J, Li X, Xu C, Ji S, He Y, Yang G, Wang X. Dual encoding for zero example video retrieval. Proceedings of the 2019 computer vision foundation (CVF), 2019 June 16-20. California, USA, IEEE explore; pp. 9346-9355.

[6] Yasin D, Sohail A, Siddiqi I. Semantic Video Retrieval using Deep Learning Techniques. Proceedings of the 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2020 January 14-18. Islamabad, Pakistan, IEEE explore; pp. 338-343.

[7] Li H, Ma Z. An Efficient Approach Based on Image Pixel and Semantic Features towards Video Retrieval. Proceedings of the 2018 2nd International Conference on Imaging, Signal Processing and Communication, 2018 July 20-22. Kuala Lumpur, Malaysia, IEEE explore; pp.46-63.

[8] Tseytlina B, Makarova I. Content based video retrieval system for distorted video queries. Proceedings of the Modeling and Analysis of Complex Systems and Processes (MACSPro'20), 2020 October 22-24. Venice, Italy, CEUR Workshop Proceedings; pp. 1-9.

[9] Mallick AK, Mukhopadhyay S. Video retrieval framework based on color co-occurrence feature of adaptive low rank extracted keyframes and graph pattern matching. Information Processing and Management, 2022; 59(2):102870. http://doi.org/10.1016/j.ipm.2022.102870

[10] Amato G, Bolettieri P, Carrara F, Debole F, Falchi F, Gennaro C, Vadicamo L, Vairo C. 2021. The VISIONE video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. Journal of imaging, 2021; 76 (7): 1-25. http://doi.org/10.3390/jimaging7050076

[11] Liu L, Li J, Niu L, Xu R, Zhang L. Activity image-to-video retrieval by disentangling appearance and motion. Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21). 2021 February 2-9. Virtual Event, AI Magazine and AAAI Press; pp. 2145-2153.

[12] Zhang C, Lin Y, Zhu L, Liu A, Zhang Z, Huang F. 2019. CNN-VWII: An efficient approach for large-scale video retrieval by image queries. Pattern Recognition Letters, 2019; 123: 82-88.

[13] Yuan L, Wang T, Zhang X, Tay FE, Jie Z, Liu W, Feng J. Central similarity quantization for efficient image and video retrieval. Proceedings of the 2020 computer vision foundation (CVF). 2020 June 16-18; Virtual Event, IEEE explore; pp. 3083-3092.

[14] Araujo A, Girod, B. Large-scale video retrieval using image queries. IEEE Transactions on circuits and systems for video technology, 2018; 28(6): 1406-1420.

[15] Song J, Gao L, Liu L, Zhu X, Sebe N. 2018. Quantization-based hashing: a general framework for scalable image and video retrieval, Pattern Recognition, 2018; 75: 175-187.

[16] Jo W, Lim G, Kim J, Yun J, Choi, Y. Exploring the temporal cues to enhance video retrieval on standardized CDVA, IEEE access, 2022; 10: 38975-38981. http://doi/10.1109/access.2022.3165177

[17] Liu H, Zhao Q, Wang H, Lv P, Chen Y. An image-based near-duplicate video retrieval and localization using improved Edit distance, Multimedia Tools Applications, 2017; 76: 24435–24456.

[18] Yu T, Mascagni P, Verde J, Marescaux J, Mutter D, Padoy N. 2022, Live laparoscopic video retrieval with compressed uncertainty. arXiv, 2022; 2203.04301 [eess.IV]: 1-16. http://doi.org/10.48550/arXiv.2203.04301

[19] Deshpande A, Estrela VV, Patavardhan P. The DCT-CNN-ResNet50 architecture to classify brain tumors with super-resolution, convolutional neural network, and the ResNet50. Neuroscience Informatics, 2021; 1(4): 1-8.

[20] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Muthana A, Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big data, 2021; 8(53): 1-74.

[21] Amalia L, Alejandro C, Alejandro M, Ana H. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 2019; 91: 216-231.

[22] Guo Y, Li S, Zhang Z, Li Y, Hu Z, Xin D, Chen Q, Wang J, Zhu R. 2021. Automatic and accurate calculation of rice seed setting rate based on image segmentation and deep learning. Front. Plant Sci, 2021;12:770916. https://doi.org/ 10.3389/fpls.2021.770916

DOI: https://doi.org/10.15379/ijmst.v10i3.3411