Semantic concept based video retrieval using convolutional neural network

Nitin Janwe¹ · Kishor Bhoyar²

Received: 16 November 2019 / Accepted: 7 December 2019 / Published online: 14 December 2019
© Springer Nature Switzerland AG 2019

Abstract

Retrieval of videos efficiently and effectively has become a challenging issue nowadays and dealing with multi-concept videos is the center of focus. The aim of the work presented here is to propose an improved semantic concept-based video retrieval method using a novel ranked intersection filtering technique and a foreground driven concept co-occurrence matrix. In the proposed ranked intersection filtering technique, an intersection of ranked concept probability scores is taken from key-frames associated with a query shot to identify concepts to be used in retrieval. Convolutional neural network is used as a baseline. The proposed method is implemented using a classifier built with a fusion of asymmetrically trained deep CNNs to deal with data imbalance problem, a novel foreground driven concept co-occurrence matrix to exploit concept co-occurrence information and a ranked intersection filtering approach. Performance is evaluated by a measure, mean average precision on TRECVID multi-label dataset. The results are compared with state-of-the-art other existing methods in its class and shown its superiority.

Keywords Video concept detection · Semantic concept-based video retrieval · Multi-label classification · Convolutional neural network · Deep learning · Concept co-occurrence matrix · Foreground driven concept co-occurrence matrix

1 Introduction

Due to the technology advancement in video capture, its storage and possible transmission, and extremely affordable costing of these devices has made big contribution in the explosive growth of video collections on the internet, which leads to need for efficient and effective access and retrieval of video data. Concept-based video retrieval is a way to facilitate video access. Providing concept level access to video data requires, indexing techniques which indexes videos on semantic concepts. For better access and retrieval of videos, effective indexing and retrieval techniques are necessary. The effectiveness of a video retrieval algorithm depends on the accuracy with which videos are accessed from dataset. The most important

factors of any video retrieval system is (1) the effectiveness of a concept detector, (2) post classification improvement of concept probabilities and (3) effectively dealing multiple key-frames of a shot for effective access from dataset.

The work in this paper is an extension of [1]. Here, we addressed the issue of semantic video retrieval and proposed and implemented a novel approach for retrieval with improved performance. In [1], we have shown that an efficient concept classifier can be implemented using deep convolutional neural network (CNN) by extracting semantic features from video. In this work, the same approach is used for implementing a classifier. The main contribution of this paper is a video retrieval framework based on concept co-occurrence information and proposed Ranked

✉ Nitin Janwe, nitinj_janwe@yahoo.com; Kishor Bhoyar, kbbhoyar@yahoo.com | ¹Department of CSE, Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, India. ²Department of IT, Yeshwantrao Chavan College of Engineering, Nagpur, India.

Intersection Filtering (RIF) approach for efficient video retrieval.

The paper is organized as follows: related and existing work is summarized in Sect. 2. Proposed method with detailed framework and implementation of CNN classifier its architecture, concept of asymmetrically trained deep CNNs, FDCCM and Ranked intersection filtering approach is described in Sect. 3. Experimental results and performance evaluation measure is given in Sect. 4 and lastly, Sect. 5 concludes the paper.

2 Related work

In semantic concept-based video retrieval system, retrieval of videos is a last step after initial steps of shot boundary detection, extraction of key-frames and concept detection.

Video search and retrieval process can be effectively carried out on the indexed database. A good survey on concept-based video retrieval is presented by Snoek and Worring [2]. Feng and Bhanu [3] discussed the concept and use of concept co-occurrence patterns for image annotation and retrieval. Kuo et al. [4] presented the work on the use of deep convolutional neural network for image retrieval. Podlesnaya and Podlesnyy [5] and [6] exhibits the work on deep learning based video indexing and retrieval. Kikuchi et al. [7] presented the work on video semantic indexing using object detection-derived features. Awad et al. [8] discussed a 6-year retrospective on TRECvid semantic indexing of video. The use of co-occurrence information can be used to classify videos is presented in [9]. Multi-label image classification with a probabilistic label enhancement model is discussed by [10]. Donahue et al. [11] presented the work on use of deep learning for visual recognition. A discussion on the effectiveness of semantic concepts for improved video retrieval is given by [12]. Many robust concept-based video retrieval methods have been presented in [13–16]. Effective recognition of objects in videos is important for effective object-based video concept detection and retrieval. Visser et al. [17] and [18, 19] exhibited works on video retrieval based on concepts using detected objects. Concept-based image retrieval approaches could be useful for video retrieval methods [20, 21]. Mezaris et al. [22] exploited automatically extracted video semantics for improved interactive video retrieval. Shirahama et al. [23], and [24, 25] implemented robust video retrieval system. The work [26] presents selection of concepts and concept detectors for video search.

3 Proposed method

The main objective of the proposed system is to retrieve relevant video shots from a dataset of videos when a set of key-frames are given as input to the system. To achieve the above objective, it is needed to have a concept classifier which will detect semantic concepts from input keyframes and on the basis of detected concepts relevant shots are retrieved from dataset indexed on semantic concepts. The proposed framework at basic level therefore consists of two modules namely (1) Training module and (2) Testing module through which it does its designated objective. In training module, a supervised training is used to implement a classifier with training key-frames dataset and in testing module, relevant key-frames and underlying shots are obtained by using a trained classifier. The framework of the proposed system is shown in Fig. 1. The proposed system also highlights the detailed steps of concept detection, score refinement and detecting common concepts from input key-frames for efficient retrieval.

The proposed framework retrieves the key-frames and the underlying shots using following modules—(1) Multi-class-classifier for performing video concept detection (2) Score refinement using Foreground Driven Concept Cooccurrence Matrix (FDCCM) to improve concept detection rate and (3) Use of proposed novel RIF approach to short-list common concepts from input key-frames for efficient video shots retrieval.

A classifier for the proposed method is built using a fusion of asymmetrically trained deep CNNs followed by score refinement using FDCCM matrix. FDCCM matrix is obtained from Concept Cooccurrence Matrix (CCM) which is obtained by averaging two CCMs, namely CCM_{local} and CCM_{global} as given by Eq. 1. CCM_{local} is implemented using local Trecvid dataset whereas CCM_{global} is obtained from random image dataset. Random image dataset is a collection of images other than Trecvid dataset which are obtained by retrieving from internet using Google Images.

$$CCM = Avg(CCM_{local}, CCM_{global}) \quad (1)$$

A fusion of classifiers and its implementation, idea of asymmetric training and its performance gain has been discussed in detail in next subsections. The concepts detected for a key-frame/s will serve as an index for corresponding video shot and will be helpful and useful in retrieval process. As this index is based on semantic concepts, it is called semantic index. The entire dataset will be maintained based on semantic index.

In testing module, key-frames of a query shot are given as input and the aim is to search for top-k most semantically relevant shots from dataset matching to the input. If the input is through query by example then key-frames are

Fig. 1 Concept based video indexing and retrieval using FDCCM and RIF

extracted from a query shot. After applying concept detection step for each individual key-frame, common concepts among these key-frames are identified using RIF method. Using a set of common concepts, a database indexed on semantic concepts is searched for finding most relevant key-frames and associated shots. The performance is evaluated by a parameter Precision.

3.1 Building CNN classifier

The proposed method implements a classifier using CNN as a baseline. The architecture of the CNN used for building classifier is shown in Fig. 2. The proposed CNN architecture is comprised of 7 layers. The input

key-frames are presented to the first convolutional layer and output is taken from the last Softmax layer. A key-frame of size 352×288 (with three color channels), are initially resized to 88×72 and then converted to gray-scale (single channel) is presented as input. The aim of resizing and gray-scale conversion is to reduce memory requirement and time complexity. The gray-scale output is convolved with 40 different first layer square filters, each of size 6×6 using a stride of 2 in both x and y directions. The resulting feature maps are then: (1) pooled (max within 2×2 regions, using stride 2), (2) contrast normalized across feature maps and (3) passed through a rectified linear function, to give 40 different 17×21 element feature maps. Similar operations

Fig. 2 Proposed CNN architecture used to build concept classifier

are repeated in layer 2. The last two layers are fully connected, taking features from the top convolutional layer as input in vector form. C-way Softmax is the last layer of the architecture where C being the number of classes which is 36 in our case.

A fusion of multiple asymmetrically trained CNNs, a novel approach proposed by us in [1], is used for classifier implementation. The approach of asymmetric training has been very effective while dealing with imbalanced datasets. Imbalanced dataset has training samples considerably different in number for each class. In imbalanced dataset, the classes with larger training sample population can be considered as strong classes while with lower sample population are weak classes. It has been observed that, while classifier training, strong classes tend to learn quickly than weak classes and exhibits improved detection rate during initial stage of training. If the same classifier is used to train both strong and weak classes, it becomes difficult to achieve better detection rate for both categories of classes. Therefore, it was proposed in [1] to use an approach of a fusion of asymmetric trained deep CNNs to build a classifier to deal with imbalanced dataset samples. Here, two separate CNNs with the same architecture are trained independently (asymmetrically) with the same dataset to classify strong and weak classes respectively as shown in Fig. 3. The output scores from these separate classifiers are then fused to get the final detection rate. To separate strong and weak classes, the dataset concepts are divided into two Groups using Global Thresholding Method. The first will hold Group-1 concepts with larger population and other group will hold Group-2 concepts of smaller population. Once the classifier output is obtained, it is refined using novel FDCCM, which is obtained exploiting concept co-occurrence relationship among concepts and is discussed next in detail.

3.2 Foreground driven concept co-occurrence matrix

The FDCCM concept, an approach proposed by us in our previous work [1], is discussed in this section. In a process of detecting semantic concepts, concept's visual co-existence helps in providing semantic information. In video shot, concepts co-exist with each other. Current methods use concept pairs for deriving semantic concepts from shot. If *Road-Bus* is a semantic pair, then the presence of concept *Road* increases the confidence of *Bus* by its co-existence, as *Bus* runs on a *Road*. A concept co-occurrence matrix (CCM) needs to be implemented to maintain concept co-occurrence data. The CCM we compute is a combination of two CCMs computed for local and global levels respectively. The local level CCM is derived by considering prelabelled dataset and their concept visual co-existence. The global level CCM is derived from images retrieved using Google Images. Final CCM is computed by averaging CCMs at local and global levels as given by Eq. 1.

In a concepts list, we notice two concepts types: *foreground* or actor concepts (such as *moving Car, Ball*) and *background* or passive concepts (such as *Road* and *Crowd* in football ground). Once the CCM is prepared, we derive FDCCM from it. FDCCM is a matrix of foreground and background concepts. It is prepared in such a way that, when we pass on any foreground concept, it returns us a list of background concepts co-exists with it. Its use in proposed method is for refinement of concept scores and to get improved detection rate.

3.3 Proposed ranked intersection filtering approach

In the video retrieval system, a query video shot is given as an input to the system and in the output, it retrieves and ranks all the relevant shots from the indexed dataset. A shot segmentation process segments a video into

Fig. 3 A fusion of asymmetrically trained CNNs

multiple shots. As shown in Fig. 1, one of the shots is given as an input to the system. It is then subjected to the key-frame extraction process, which extracts representative key-frame/s for input shot. The key-frames are then submitted to a multi-concept classifier which detects a list of concepts for all submitted key-frames. The concept lists of all key-frames of a shot are given to proposed RIF technique for finding common concepts which belong to all key-frames of a shot. If concept lists are considered as concept sets, then the intersection operation finds common concepts from key-frames. The common concepts in key-frames of a shot are assumed to represent salient content of a shot which is a necessity for effective video retrieval. The concepts thus identified are ranked on their probability values. Figure 4 presents RIF method which shows the extraction of common concepts using intersection operation. In Fig. 4, a shot of three key-frames namely KF-1, KF-2 and KF-3 is given as an input. KF-1 consists of concepts *Crowd*, *Vegetation*, *People*, *Airplane* and *Sky*, KF-2 consists of concepts *Airplane*, *Sky*, *Mountain* and *Vegetation* whereas KF-3 contains *Car*, *Road*, *People*, *Sky* and *Airplane* concepts. After applying proposed RIF method, the concepts which are common in all KFs i.e. *Airplane* and *Sky* are extracted. The extracted common concepts have different probability scores for different KFs. Here, we have to find out a single probability value for each extracted concept. We therefore, take average of probability scores from KFs for each concept. They are then ranked on their final probability values. Hence, this method is named as RIF.

Fig. 4 The proposed ranked intersection filtering method (KF—key-frame)

4 Experimental setup

The dataset and the measure used for performance evaluation for the proposed method is discussed in this section.

4.1 Video dataset

In the experimentation, for training and testing CNN classifier, TRECVID2007 development dataset has been used which consists of 110 video clips preprocessed into approximately 19 thousand shots and 6 lakh 64 thousand key-frames. The dataset is preprocessed to shots and corresponding key-frames dataset, using some shot detection and key-frame extraction techniques before classifier training. We divided the dataset into two partitions; namely Training and Testing as given in Table 1. The training dataset is chosen using 17,114 randomly selected positive key-frames from first 90 videos while test dataset is formed from 9352 randomly selected positive key-frames of next 20 videos for testing. The dataset has 36 concepts in its concepts vocabulary, whose details are available on the TRECVID website [27]. NIST [28] provides ground-truth dataset. The experimentation is performed using the ground-truth dataset.

4.2 Evaluation measure

The measure used to evaluate performance of proposed video retrieval method is Precision (P). It is defined as the ratio between the relevant shots retrieved by the method (Hit) from the total relevant and non-relevant (False-hit) shots retrieved in the output.

In our experimentation, we have considered evaluation of performance for top-6 retrieved shots in the retrieved output. As in our method, every shot is represented by representative key-frames, hence in our experimentation, relevant and non-relevant key-frames will be considered instead of relevant and non-relevant video shots. The computation of P is done by using Eq. 2,

$$P = \frac{H}{(H + F)} \quad (2)$$

where H is total number of relevant key-frames (Hit) from total retrieved frames and, F is total number of non-relevant key-frames (False-hit). The MatConvNet [29], an open source library for Matlab is used to implement CNN.

Table 1 TRECVID dataset and details of its partitions

Dataset	Partition	Partition name	# of videos	# of key-frames
TRECVID development dataset	Partition-I	Training dataset	90	17,114
	Partition-II	Testing dataset	20	9352

4.3 Experimental results

Figure 5 presents the video retrieval results for two different input instances of sample key-frames. In the first instance, a shot consists of a single key-frame (key-frame no. 9199, in a test dataset of 9352 key-frames) is presented as an input to the proposed system. Column 1 shows the visual appearance of the input key-frame, Column 2 gives the ground-truth concept ids for the input key-frame which are 4 (for Building), 6 (Car), 21 (Outdoor), 23 (Person), 32 (Urban), and 34 (Walking_Running). Ground-truth concepts are known concepts for a sample key-frame. Upon presenting input key-frame to the proposed system, a classifier outputs probability scores for all 36 concepts, then it is subjected to score refinement process using FDCCM, the refined scores are then ranked as per their values and then top 'n' scores are matched with the ground-truth concepts (evaluation measure) and then finally matched concepts are identified as detected concepts. Column 3 in Fig. 5 shows the detected concepts for a input key-frame. For first case, the concepts detected are 21 (Outdoor), 32 (Urban), 4 (Building), 27 (Sky) and 6 (Car). Column 4 gives us the common detected concepts from the presented input key-frames after applying proposed RIF method on them. Since in the first case, there is a single frame, the output of RIF method is the same i.e. concepts (21,32,4,27 and 6). Using these detected concepts and their index, the dataset is searched and top six most relevant keyframes are retrieved (key-frame no. 7437, 7645, 8010, 8092, 8106 and 8114, from training dataset) which are shown in next columns. In second instance, since two key-frames belongs to

the same shot, two key-frames are presented as input (key-frame no. 5311 and 5312), their ground-truth concepts are 17,21,23,24,33 for frame-1 (5311) and 6,17,21,23,33,34 for frame-2 (5312) respectively. Their individual detected concepts are 6,26,32,31 and 24 for frame-1 and 23,24,10 for frame-2. Post application of RIF method, common concept detected is only 24 (Police_Security) and hence using this concept dataset is searched and retrieves the key-frames having *Police_Security*, the key concept. The top six retrieved key-frames are shown in next columns. The performance of the proposed work is given in Table 2. For both input examples, out of six retrieved key-frames, the number of Hits are 5 and False-hits are 1, therefore Precision for both cases computes to $5/6 = 0.83$. The MAP for entire Trecvid testing dataset is computed which comes out to 0.544. The performance comparison of the proposed method with state-of-the-art other methods in the category is given in Table 3. The MAP for the proposed method which is 0.544 is lot superior than other contemporary methods in the class, Statistical_Active_Learning [30], MAP = 0.235; CRMActive [31], MAP = 0.260 and pLSA [32], MAP = 0.390.

Table 2 Sample test key-frames retrieval performance

S. no.	Key-frame/s Id	Key concept	H	F	Precision (P)
1.	9199	Car and person	5	1	0.83 (83%)
2.	5311, 5312	Police_Security, Person, Face, Crowd, Outdoor	5	1	0.83 (83%)

Key-frame	Ground-truth concepts	Detected concepts	Ranked Detected Concepts	Retrieved Key-frames for shots
4,6,21,23,32,34	21,32,4,27,6	21,32,4,27,6
9199	6	4		7437	7645	8010	8092	8106	8114
Key Fr-1: 17,21,23,24,33	6,26,32,31,24	24
5311	5			10526	10531	10536	10537	10562	10566
Key Fr-2: 6,17,21,23,33,34	23,24,10							
5312	6								

Concepts:

4: Building

6: Car

10: Crowd

17: Military

21: Outdoor

23: Person

24: Police_Security

26: Road

27: Sky

32: Urban

33: Vegetation

34: Walking_Running

Fig. 5 Retrieval results using proposed algorithm

Table 3 Comparison of performance of proposed and other existing methods

Method	Dataset used	Precision (MAP)
Proposed method	TRECVID 07	0.544
Statistical_Active_Learning [30]	TRECVID 07	0.235
CRMAActive [31]	TRECVID 07	0.260
pLSA [32]	TRECVID 07	0.390

5 Conclusion

This work has introduced a novel method for semantic concept-based video indexing and retrieval using a state-of-the-art classifier built using a fusion of asymmetrically trained deep CNNs to deal with imbalance dataset combined with FDCCM, and a novel RIF approach. Its evaluation on TRECVID dataset using Precision parameter shows that, proposed method is lot superior than other contemporary methods in the class.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Janwe NJ, Bhojar KK (2018) Multi-label semantic concept detection in videos using fusion of asymmetrically trained deep convolutional neural network and foreground driven concept co-occurrence matrix. *Appl Intell* 48(8):2047–2066
- Snoek CGM, Worring M (2009) Concept-based video retrieval. *Found Trends Inf Retr* 2(4):215–322
- Feng L, Bhanu B (2016) Semantic concept co-occurrence patterns for image annotation and retrieval. *IEEE Trans Pattern Anal Mach Intell* 38(4):785–799
- Kuo CH, Chou YH, Chang PC (2016) Using deep convolutional neural networks for image retrieval. *Soc Imaging Sci Technol*. <https://doi.org/10.2352/issn.2470-1173.2016.2.vipc-231>
- Podlesnaya A, Podlesnyy S (2016) Deep learning based semantic video indexing and retrieval. [arXiv:1601.07754\[cs.LG\]](https://arxiv.org/abs/1601.07754)
- McCormac J, Handa A, Davison A, Leutenegger S (2016) SemanticFusion: dense 3D semantic mapping with convolutional neural networks. [arXiv:1609.05130v2\[cs.CV\]](https://arxiv.org/abs/1609.05130)
- Kikuchi K, Ueki K, Ogawa T, Kobayashi T (2016) Video semantic indexing using object detection-derived features. In: *Proceedings 24th European signal processing conference (EUSIPCO)*, Budapest, Hungary, pp 1288–1292
- Awad G, Snoek CGM, Smeaton AF, Quénot G (2016) TRECVID semantic indexing of video: a 6-year retrospective. *ITE Trans Media Technol Appl* 4(3):187–208
- Modiri S, Amir A, Zamir R, Shah M (2014) Video classification using semantic concept co-occurrences. <https://doi.org/10.1109/cvpr.2014.324>
- Li X, Zhao F, Guo Y (2014) Multi-label image classification with a probabilistic label enhancement model. In: *UAI'14 proceedings of the thirtieth conference on uncertainty in artificial intelligence*, pp 430–439
- Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: a deep convolutional activation feature for generic visual recognition. In: *Proceedings of the international conference on machine learning, ICML, Beijing, China*, pp 647–655
- Hauptmann A, Yan R, Lin WH, Christel M, Wactlar H (2007) Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Trans Multimed* 9(5):958–966
- Memar S, Suriani AL (2013) An integrated semantic-based approach in concept based video retrieval. *Multimed Tools Appl* 64:77–95
- Xu CS, Wang JJ, Lu HQ (2008) A novel framework for semantic annotation and personalized retrieval of sports video. *IEEE Trans Multimed* 10(3):421–436
- Dalton J, Allan J, Mirajkar P (2013) Zero-shot video retrieval using content and concepts. In: *22nd international conference on information & knowledge management. ACM, New York, NY, USA*, pp 1857–1860
- Aly R, de Franciska JDH, Apers PMG (2012) Simulating the future of concept-based video retrieval under improved detector performance. *Multimed Tools Appl* 60:203–231
- Visser R, Sebe N, Bakker EM (2002) Object recognition for video retrieval. In: *International conference on image and video retrieval*, pp 262–270
- Browne P, Smeaton AF (2005) Video retrieval using dialogue, keyframe similarity and video objects. In: *IEEE international conference on image processing. IEEE*, pp 1208–1211
- Jiang Y, Ngo C, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: *6th ACM international conference on image and video retrieval. ACM, Amsterdam, The Netherlands*, pp 494–501
- Carneiro G, Chan AB, Moreno PJ, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans Pattern Anal Mach Intell* 29(3):394–410
- Siddiquie B, Feris RS, Davis LS (2011) Image ranking and retrieval based on multi-attribute queries. In: *IEEE conference on computer vision and pattern recognition*, pp 801–808
- Mezaris V, Sidiropoulos P, Kompatsiaris I (2011) Improving interactive video retrieval by exploiting automatically-extracted video structural semantics. In: *Fifth IEEE international conference on semantic computing. IEEE*, pp 224–227
- Shirahama K, Kumabuchi K, Uehara K (2012) Video retrieval by managing uncertainty in concept detection using Dempster–Shafer theory. In: *Fourth international conferences on advances in multimedia*, pp 71–74
- Lin L, Chen C, Shyu ML (2011) Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE MultiMed* 18(3):32–43
- Dong X, Chang SF (2017) Visual event recognition in news video using kernel methods with multi-level temporal alignment. In: *IEEE international conference on computer vision and pattern recognition, Minneapolis*
- Wei XY, Ngo CW (2008) Selection of concept detectors for video search by ontology-enriched semantic spaces. *IEEE Trans Multimed* 10(6):1085–1096
- TRECVID. <http://www-nlpir.nist.gov/projects/trecvid/>
- NIST. <http://www.nist.gov>

29. Vedaldi A, Lenc K (2015) MatConvNet: convolutional neural networks for MATLAB. In: International conference on multimedia. ACM, pp 689–692
30. Zha ZJ, Wang M, Zheng YT, Yang Y, Hong R (2012) Interactive Video indexing with statistical active learning. *IEEE Trans Multimed* 14(1):17–27
31. Chatterjee M, Leuski A (2015) CRMAActive: an active learning based approach for effective video annotation and retrieval. In: ICMR'15. ACM
32. Beltran R, Pla F (2015) Incremental probabilistic latent semantic analysis for video retrieval. *J Image Vis Comput* 38:1–48

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.