Pacific Graphics (2018)
H. Fu, A. Ghosh, and J. Kopf (Editors)

Short Paper

A Deep Learned Method for Video Indexing and Retrieval

X. Men[1] and F. Zhou[1] and X.Li [1]

[1]Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications,Beijing,China

Abstract

In this paper, we proposed a deep neural network based method for content based video retrieval. Our approach leveraged the deep neural network to generate the semantic information and introduced the graph-based storage structure to establish the video indices. We devised the Inception-Single Shot Multibox Detector (ISSD) and RI3D model to extract spatial semantic information (objects) and extract temporal semantic information (actions). Our ISSD model achieved a mAP of 26.7% on MS COCO dataset, increasing 3.2% over the original SSD model, while the RI3D model achieved a top-1 accuracy of 97.7% on dataset UCF-101. And we also introduced the graph structure to build the video index with the temporal and spatial semantic information. Our experiment results showed that the deep learned semantic information is highly effective for video indexing and retrieval.

CCS Concepts

•***Computing methodologies*** → ***Visual content-based indexing and retrieval;***

1. Introduction

With an increasing number of videos are generated nowadays, it is essential to index and retrieve the video data efficiently to combat the information explosion. The common method of video indexing and retrieval is usually achieved by manual annotations. Content-based means retrieving by analysing the contents of the videos rather than the metadata such as tags and descriptions associated with the video. Content-based video semantic search aims at searching the high-level semantic concepts of the video content. A concept can be presented as video, audio, or semantic tag of objects, people, actions and scenes in the video content [ALN*17]. Our work focused on Zero-example (0Ex) Video Search (AVS) [ALN*17] which is basically text-to-video search. In this scenario the end user may be interested in the semantic information such as objects, actions in the video and queries are given in text format. Such video retrieval task depends heavily on the accuracy of interpreting the semantic content of videos. The general method is annotating and indexing videos with semantic information by manual work during offline processing, and then search videos with relevant semantic information matching query expression [LNZN17, LZdBN16]. In our work, we devised an improved object detection model ISSD and RI3D model to generate the semantic information for the 0Ex task. Our structure is shown in Figure 1. First, we employ the ISSD model and RI3D model to extract semantic concept from the input video stream, the output contains both temporal and spatial semantic information. Then we employ WordNet [Mil95] to introduce structure based representation. Finally we built the indices for the video with the semantic information output by the deep neural network. WordNet [Mil95] provides the graph-based storage structure and support of fuzzy query. We showed that AVS task can be accomplished by using the semantic information extracted from video by the deep neural network. Our contributions are:(1)We presented a video indexing and retrieval method based on deep neural network.(2)We devised the ISSD model and RI3D model for the 0Ex task.(3)Our work introduced the graph-based structure to support fuzzy query in video retrieval.

2. Related Work

A number of studies have been proposed on the video retrieval and related subjects [JYM*15, LAE*16, JMYH14]. Many well-established methods for video searching and retrieving have to rely on manual annotation to understand the high level semantic information of the video to process the video in semantic level [PP16]. Some other approaches [JYM*15, JMYH14, GM14] extract the key frame and the low level features to tackle this problem. Some methods based on machine learning was also proposed [LAE*16, GM14]. These methods trained a classifier with both low-level and high-level features so that the final decision can be obtained through the fusion of the individual classification results. For example, Gkalelis et al. [GM14] demonstrated a representation for linear support vector machine(SVM) by subclass discriminant analysis. Habibian et al. [HMS14] proposed to index videos by composite concepts that are trained by combining the labeled data of individual concepts.

Inspired by the deep neural network breakthroughs in the image

DOI: 10.2312/pg.20181287

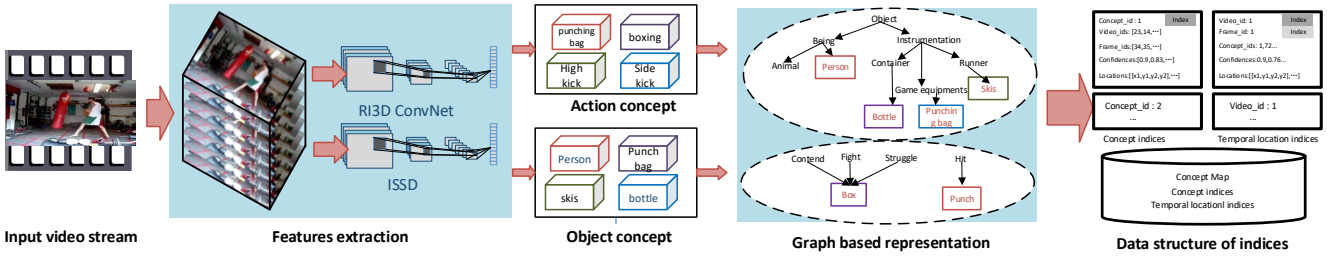

Figure 1: *Schematic illustration of our overall architecture.*

classification and object detection task [LAE*16,CZ17,QWY*16], we introduced the deep neural network to extract semantic information. We started by detecting objects and actions that occurs in the video and then index video according to the semantic information output by the deep neural network. The graph-based storage structure was introduced to store the semantic information, by which we could build the semantic relationship between the extracted semantic information and support more complicated searching queries.

3. Video indexing and retrieval

3.1. Obtain spatial information

To extract the spatial information better, we designed the ISSD object detection model based on Single Shot Multibox Detector(SSD) [LAE*16]. The original SSD performance on the detection of small objects is not satisfying, which means the extracted spatial information is quiet limited. To solve this problem, we introduced the Inception [IS15] structure to replace the convolutional layers of the original SSD. Compared to the original 3×3 Convolution kernel, the Inception structure stacks the 1×1, 3×3 and 5×5 convolution kernels, which indicates the receptive field is a fusion of several receptive fields. The 1×1 convolutional kernel could keep more information of small objects. But this brings about some training problems, such as over-fitting and vanishing gradient. So we also introduced the residual network in our Inception block. The Inception block is shown in Figure 2. Considering the tradeoff between speed and performance, we introduced the VGG16 structure for the basic feature extraction. The size of the extra layers is the same as the [LAE*16] described. Consequently, the scope of the receptive field is expanded, which improve the sensitivity of the network to small objects. The structure of ISSD model is shown in Figure 3.

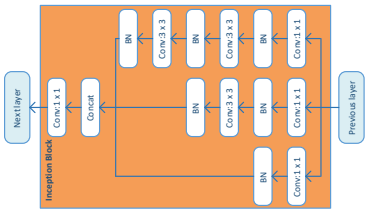

Figure 2: *The Inception Block*

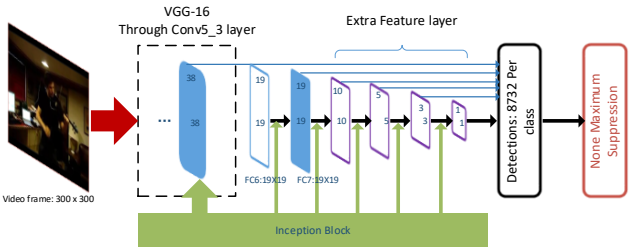

Figure 3: *the ISSD structure*

3.2. Obtain temporal information

The 3D convolutional kernel and residual structure were applied on the GoogLeNet to recognize the action of the video content. We called our network RI3D network. Specifically, the convolutional layer is similar with Inception block shown in Figure 2. We replaced the 2D convolutional layers with 3D convolutional kernels whose depth size and depth strides are set to 3. The 3D convolutional layer makes the network capable of extracting features of multiple frames. We also incorporated the Two-Stream model introduced in [SZ14] which provides the fusion of feature extracted from RGB image and optical flow image.

We employ the TV-L1 Algorithm [ZPB07] to compute optical flow. The optical flow values are truncated to the range [-20, 20] and then rescaled between -1 and 1. The RGB frames are resized with bilinear interpolation, preserving aspect ratio so that the smallest dimension is 256 pixels. The pixel values are also rescaled between -1 and 1. our experiment shows that these operation lead to a action recognition accuracy.

3.3. Graph oriented indexing

The graph-based structure was introduced for building video indices. We used WordNet [Mil95] lexical database. The WordNet database is a graph-based representation of words (synsets) connected with linguistic relations. The graph-based structure could establish the relationships between concepts, which enable us to transform the keyword into different words without loss of the accuracy (e.g. search by the keyword animal, we could return dog, cat, etc. We store the concepts as a graph

$$\begin{aligned} G_{video} &= (N, E) \\ N &= (N_{action}, N_{object}) \\ E &= (E_{hypernym}, E_{hyponym}, E_{derivation}) \end{aligned} \tag{1}$$

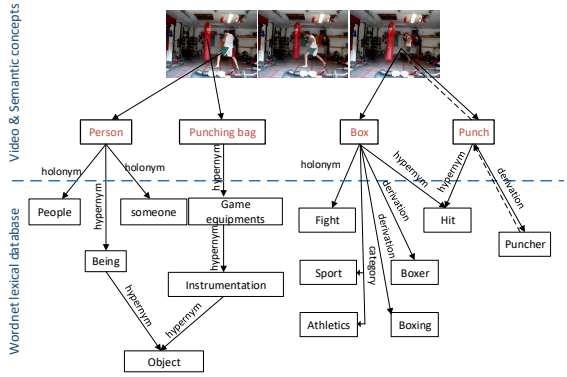

Figure 4: *graph-based representation of video semantic concepts.*

Figure 4 illustrates the graph representation of a video contains object punch bag and action boxing. The dash line outlines the search route of the keyword puncher, which match the extracted semantic concept punch by the derivation related form.

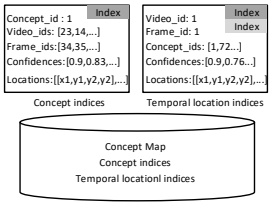

Figure 5: *Data structure of indices for storage*

The semantic information extracted by the models was used to establish indices. The index structure is designed as the Figure 5 shows. Indices bridge the gap between the concepts and video. And the graph-based storage and linguistic relations are kept by the WordNet. In implementation, we set up two tables and several indices to accelerate the query speed of different query statements, Specifically, we selected the objects whose confidence score is more than 0.8 and the top-1 action class to establish the indices.

4. Experiments

In our work, the ISSD model was trained on the MS COCO dataset and the RI3D model was trained on Kinetics and UCF-101 dataset. The Xavier [GB10] method was introduced to initialize all parameters in our neural network and the Adam optimizer was employed to train our network. Both of the network are initialized with ImageNet weight initialization. The batch size was set to 32 and IoU was set to 0.5. The video stream was decoded as many RGB frames and fed into the model. The input size is 300 pixels$\times$300 pixels.

Table 1 shows the ISSD'S performance on COCO and Table 2 shows the Top-1 accuracy RI3D model on UCF-101. We obtained a mAP@[0.5:.05:.95] of 26.7% on COCO and an accuracy of 97.7% on UCF-101. Compared to the original SSD model, the precision and recall in detecting small objects are both largely improved,

Table 1: *COCO test-dev2015 detection results.*

Model	mAP	AP^{50}	AP^S	AP^M	AP^L	AR^S	AR^M	AR^L
Fast [BLZBG16]	20.5	39.9	4.1	20.0	35.8	7.3	32.1	52.0
Faster [LMB*14]	24.2	45.3	7.7	26.4	37.1	12.0	38.5	54.4
YOLOv2 [LAE*16]	21.6	44.0	5.0	22.4	35.5	9.8	36.5	54.4
SSD300 [LAE*16]	23.2	41.2	5.3	23.2	39.6	9.6	37.6	56.5
ISSD300	**26.7**	**42.6**	**8.4**	**25.9**	**40.3**	**14.4**	**42.5**	**59.7**

Table 2: *Accuracy on UCF-101 dataset*

Methods	Accuracy
Two-Stream [SZ14]	88.0
Dynamic Image Networks + IDT [BFG*16]	89.1
Two-Stream Fusion + IDT [FPZ16]	93.5
C3D one network [FPZ16] Sports 1M pre-training	82.3
RGB-RI3D, Kinetics pre-training	**95.6**
Flow-RI3D, Kinetics pre-training	**96.9**
Two-Stream RI3D, Kinetics pre-training	**97.7**

while the AP increases 3.1% and the AR increases 4.8%. We explained this by the capability of the inception block on feature extraction, and the different convolution kernel made the feature map keep more details. And the accuracy of action recognition accuracy also achieves a significant improvement.

The 0Ex task was conducted on UCF-101 dataset. Figure 7 shows some sample query results. Human effort was introduced to count true/false positives. We took 100 queries for the experiment evaluation. Figure 6 shows the precision distribution. Our experiments on UCF-101 dataset by keyword search achieved an average precision of 76.2%.

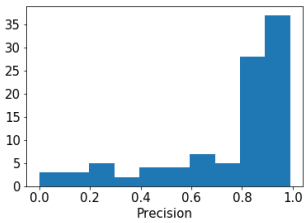

Figure 6: *retrieval precision distribution on UCF-101 dataset*

5. Conclusion

In this work, we demonstrated that deep neural network is suitable and capable of extracting semantic information of videos. We devised the RI3D model to extract temporal semantic information, which enable us to search video by the action occurred in the video. Our ISSD model proved that the fusion of receptive fields helps to keep more information and leads to an improvement on object detection which we used to extract semantic concepts. We also leveraged the graph-base storage to support fuzzy query which proves to be an aided method for video retrieval. We demonstrated that AVS task could be accomplished with a high performance by the deep

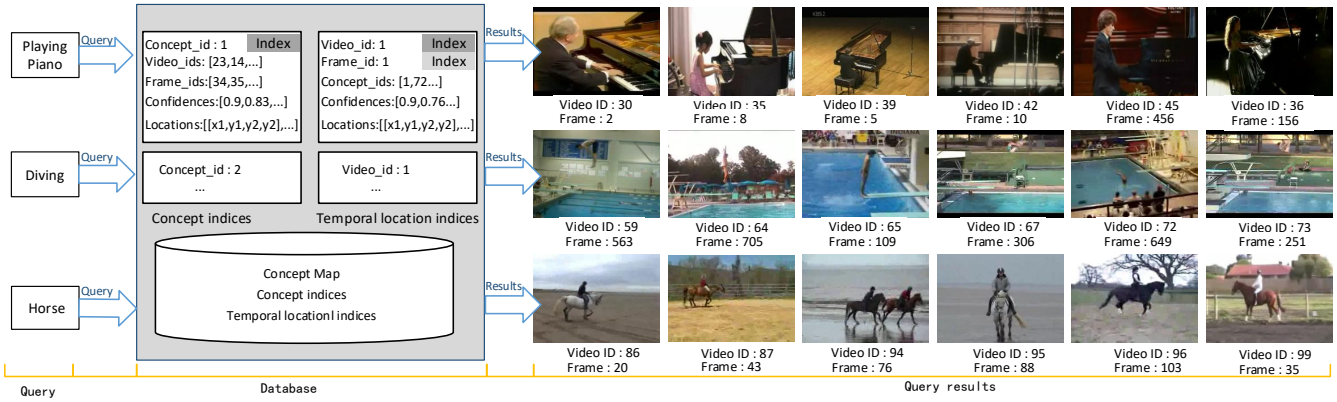

Figure 7: *0Ex Video retrieval result sample on UCF-101 dataset*

learned involved method. Future work may leverage well designed model as a base component for video retrieval task.

Acknowledgements

This works is supported by Beijing Municipal Commission of Education (Co-constructing Program)

References

[ALN*17] AWAD G., LE D.-D., NGO C.-W., NGUYEN V.-T., QUÉNOT G., SNOEK C., SATOH S.: Video indexing, search, detection, and description with focus on trecvid. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval* (2017), ACM, pp. 3–4. 1

[BFG*16] BILEN H., FERNANDO B., GAVVES E., VEDALDI A., GOULD S.: Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 3034–3042. 3

[BLZBG16] BELL S., LAWRENCE ZITNICK C., BALA K., GIRSHICK R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2874–2883. 3

[CZ17] CARREIRA J., ZISSERMAN A.: Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 4724–4733. 2

[FPZ16] FEICHTENHOFER C., PINZ A., ZISSERMAN A.: Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 1933–1941. 3

[GB10] GLOROT X., BENGIO Y.: Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (2010), pp. 249–256. 3

[GM14] GKALELIS N., MEZARIS V.: Video event detection using generalized subclass discriminant analysis and linear support vector machines. In *Proceedings of international conference on multimedia retrieval* (2014), ACM, p. 25. 1

[HMS14] HABIBIAN A., MENSINK T., SNOEK C. G.: Composite concept discovery for zero-shot video event detection. In *Proceedings of International Conference on Multimedia Retrieval* (2014), ACM, p. 17. 1

[IS15] IOFFE S., SZEGEDY C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015). 2

[JMYH14] JIANG L., MITAMURA T., YU S.-I., HAUPTMANN A. G.: Zero-example event search using multimodal pseudo relevance feedback. In *Proceedings of International Conference on Multimedia Retrieval* (2014), ACM, p. 297. 1

[JYM*15] JIANG L., YU S.-I., MENG D., YANG Y., MITAMURA T., HAUPTMANN A. G.: Fast and accurate content-based semantic search in 100m internet videos. In *Proceedings of the 23rd ACM international conference on Multimedia* (2015), ACM, pp. 49–58. 1

[LAE*16] LIU W., ANGUELOV D., ERHAN D., SZEGEDY C., REED S., FU C.-Y., BERG A. C.: Ssd: Single shot multibox detector. In *European conference on computer vision* (2016), Springer, pp. 21–37. 1, 2, 3

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *European conference on computer vision* (2014), Springer, pp. 740–755. 3

[LNZN17] LU Y.-J., NGUYEN P. A., ZHANG H., NGO C.-W.: Concept-based interactive search system. In *International Conference on Multimedia Modeling* (2017), Springer, pp. 463–468. 1

[LZdBN16] LU Y.-J., ZHANG H., DE BOER M., NGO C.-W.: Event detection with zero example: select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (2016), ACM, pp. 127–134. 1

[Mil95] MILLER G. A.: Wordnet: a lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41. 1, 2

[PP16] PODLESNAYA A., PODLESNYY S.: Deep learning based semantic video indexing and retrieval. In *Proceedings of SAI Intelligent Systems Conference* (2016), Springer, pp. 359–372. 1

[QWY*16] QIU J., WANG J., YAO S., GUO K., LI B., ZHOU E., YU J., TANG T., XU N., SONG S., ET AL.: Going deeper with embedded fpga platform for convolutional neural network. In *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (2016), ACM, pp. 26–35. 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (2014), pp. 568–576. 2, 3

[ZPB07] ZACH C., POCK T., BISCHOF H.: A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium* (2007), Springer, pp. 214–223. 2