Analytical Review on Semantic Query Tag Based Video Retrieval

Saba Parveen Bougdadi
PG Scholar, Department of Computer Science and Engineering
B.L.D.E.A's Dr.P.G.Halakatti College of Engineering and Technology,
Vijayapur, Karnataka, India.

Prof. Suvarna L.Kattimani
Assistant Professor, Department of Computer Science and Engineering
B.L.D.E.A's Dr.P.G.Halakatti College of Engineering and Technology,
Vijayapur, Karnataka, India.

Abstract- Content-based video retrieval in unlimited web videos is hard problem due to limited set of vocabulary and accuracy and it will create "semantic query gap". To overcome "semantic query gap" using continuous word space. Continuous word space allows fast video computation with low latency. Continuous word space bridges the "semantic query gap". And it is the core technique used for video retrieval for semantic content in a continuous word space, which leads to neatly packed together video representation. Continuous word space uses dot product to retrieve fast videos from web.

I. INTRODUCTION

It is challenging to extract video from unlimited web video and semantic content extraction is difficult. Semantic query tag based video extract meaningful concepts such as actions and object from videos. Semantic query tag based video retrieval introduces three methods. First one is Concept space which represents the point in a space. Concept space detect the particular concepts present in the detector of the concept. Second one is Dictionary space in dictionary space each concept is linked with another concept in the detector bank which are semantically related to each other. And the core technique is continuous word space which offers mapping of the words.

In continuous word space each word is linked with another word. Such word may not be necessarily present in the dictionary. Continuous word space offer mapping of related concepts.Continuous word space map query concepts to the closest concepts in the detector bank, consider an example "pizza" may be mapped to "food" in absence of a pizza detector. And combine their scores to fill the" semantic query gap". Continuous word space allows single tag or set of tags for semantic query tag video retrieval. In Continuous word space Semantic query tags are map to video and it uses dot product to get fast results of the related concepts. Since mapping of concepts have many advantages, every concept in a detector bank is related to some other concepts. And get fast result of the entered query, where user query consist of one or more tags, each tag is mapped to corresponding continuous word space. For example user query for "Animal" , then response in video for related tags such as "dog", or "cat", will be displayed.

Fisher vector and late fusion are two methods of video retrieving, Fisher vector define as it is a statistical capturing distribution of a set of vector, usually set of local image descript-tors. And Late fusion is semantic video analysis is used to understand human expression through language. Continuous word space is used to retrieve video by attaching semantic content in a continuous word space, this scheme maps query concept with the related concepts in the detector bank. And

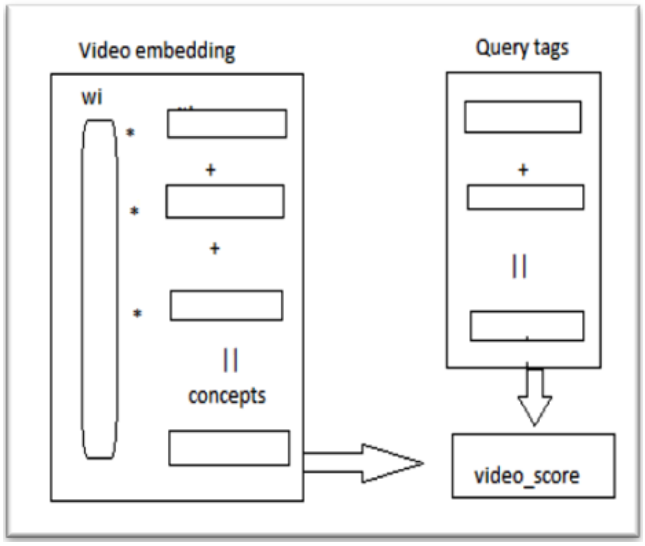

Figure1: Framework of query and video attaching

In continuous word space vector, concepts consist of c1, c2, ... ,cn video and query tag combine and generate related output to that query, they maps each meaningful concepts ci to its corresponding continuous word representation vector v(ci) videos are And this is achieved by using dot product similarity measure show in figure1. In continuous word space first is to train the set of images to detect the particular concepts. Compare to Concept space and Dictionary space, continuous word space gain significant results by retrieving videos in a continuous word space. In continuous word space, they are using the term called concept bank. Concept bank is defining as each concept in the concept bank is treated as node or leaf in the collection of the images known as imagenet. Corresponding particular image name is generated from the collection of the words knows as wordnet. Continuous word space bridges the "semantic query gap ". And allow fast video computation with providing low delay. And this is a main technique for semantic query tag based video retrieval by attaching semantic content in a continuous word space, can fit several thousand videos in a few hundred mega-bytes of memory.

II. LITRARTURE SURVEY

In [1] R. Socher al., introduces the concept of image net. In the collection of images each concept is treated as leaf or node of the image net and the corresponding image name is generated from a collection of word. The new database called "image net" a large set of images built upon the backbone of the word net structure. Collection of images is much larger in diversity, also provide more accurate results compare to current database images constructing a large collection of image database is a challenging task and it uses a method of Amazon Mechanical Turk, where Amazon Mechanical Turk is a online market place for work because of that worker can work at home. This detection of images is useful in object recognition and automatic object clustering.

In [2] P. Kumar al., Presents a self-paced learning for latent variable models which is used for addressing important tasks in machine learning presented that training data is useful to detect the particular action and objects, they uses a technique called self-paced learning. Self-paced learning means teach yourself method of learning that is directed by the learner. Self-paced is any kind of instruction that proceeds based on learner response. They focus on self-paced learning algorithm for latent variable. In statistics latent variable that are not directly observed but statistical model that relates a set of referenced variable. Self-paced learning is used for computer vision application and results increased in growing weakly supervised data.

In [3] K. Soomre et al., review on database consist of user-uploaded videos and their background. It uses bag of word approach. Bag of words is used to represent natural language processing, which increase the performance in terms of human actions. It is challenging to extract the individual human action from dataset. First they have collected images and train detector to identify the human actions from large database. Human actions are categories into human-object interaction, body-motion only, human-human interaction, playing games. Dataset consist of unlimited videos downloaded from web which is challenging to recognize the action. Results are more accurate in dataset using standard bag of words.

In [4] J. Dalton al., review on zero-shot video retrieval using content has been successful at finding videos when the query consist of tens or hundreds of related videos for training models. In zero-shot video retrieval where no training data is provided and query consist only of a text extracted from images in the videos, text recognized in the speech of its audio track, source extracted to build textual representation of semantic video from large external source such as web. Zero-shot video holds both text and semantic video concepts. It uses zero-shot video retrieval technique which requires no training data. Zero-shot video retrieval is used to identify relevant concepts for a text query. Zero-shot video retrieval uses Markov Random Field (MRF) retrieval framework to automatically identify the related concepts.

In [5] C. Sun et al., review on web videos event classification based on fisher vectors. In Computer vision Event recognize is the important topic, Fisher vector is a set of local image descriptor. Local image descriptor is defined as the description of visual feature of the content in image. Which describe characteristics such as the shape, the colour, or the motion also provide graphical representation. This technique is used for classification of unlimited videos by using fisher vectors.

In [6] T. Mikolov al., introduces distributed representation of words, phrases and their compositionality, which is efficient for vector representation to detect semantically related words uses distributed representation method, which improves the quality of vector as well as training speed. Using distributed representation it is possible to learn millions of words and phrases .distributed representation have less complexity.

In [7] R. Socher et al., presents zero-shot learning using cross modal which gives high accuracy on unseen classes and seen classes. It uses zero-shot learning, where no training data is required. Zero-shot learning is able to solve the particular tasks despite not having received any training. This modal works on both seen and unseen classes.

In [8] M. Mazloom al., review on event from video is extracted by using semantic signature. Semantic signature represent-tation is use to capture the event from videos. Semantic signature representation uses late fusion technique. Late fusion is used analyse the semantic video to understand human expression through language. User can enter multiple queries based on that query event is classified. They observe the performance of multiple video queries from event.

In [9] P. Young et al., propose the visual denotation similarity metrics from image description. It uses linguistic semantics to understand human action and expression through graph. Visual denotation similarity metrics generate graph to measure the similarity in the image. Visual denotation similarity metrics contain lexical feature, semantically related feature to get the similarity metrics.

In [10] S. Wu et al., focus on multi-modal fusion concept which uses zero-shot event detection. Zero-shot event detection means it does not required training of data. It only provide textual descriptions in addition to this it uses speech, concept detector to represent video, it uses natural languages. Multi-media zero-shot learning contains video frames, text description concept feature, lexical features finally video is scored by using similarity measure.

In [11] J. van Hout al., focus on calibration for event detection and late fusion is used to understand human expression through language. The calibration for event detection is based on multimedia event detection, multimedia event detection combine with late fusion technique. Calibration and late fusion uses arithmetic fusion scheme to generate the desired outcome.

In [12] R. Socher al., presents the dependency tree recursive neural networks represents image description and their sentence. Recursive neural network is based on dependency tree. In dependency tree recursive neural network offers mapping sentences and their images. Recursive neural network represents compositional sentences vectors, multi-modal representation and image vector representation.

In [13] T. Y. Lin et al., present a dataset to recognize object from scene. They present a statistical analysis of the detected object. The main goal is to understand the visual scene and object. It may be in 2D or 3D. Common object in context focus on image classification and object localization. It will detect particular object by plotting boxes, in that box particular object is present.

In [14] M. Mazloom al., presents an emerging topic for video event retrieval using tag to detect complex event in video. They proposed tag based video retrieval approach from tagged video collection without the need of any training data. In video event retrieval they search for event in videos and it also provide a query event. Result in significant performance gain by using late fusion.

In [15] A. Habibian et al., review on Composite concepts for zero-shot event detection. It does not required training of data. Composite concept uses Boolean logic operators by using AND/ and /OR logic operators. Advantage is that it will optimize the concept, which improves zero-shot detection accuracy.

In [16] L. Jiang al., focus on self-paced re-ranking for multi-media search. Self-paced re-ranking uses mathematical operation. To optimize the problem that can be verified theoretically. And optimize e problem that can be solved by the self-paced learning.

In [17]T. Mitamura et al., propose a technique of event search using multimodal pseudo relevance feedback, which performs task for event retrieval, multiple ranked lists to increase the performance. It offers linear programming which helps to give pseudo relevance feedback.

In [18] S. Guadarrama al., introduces open-vocabulary object retrieval which is extremely useful for robotics application. Open-vocabulary object retrieval uses object retrieval from image to text. Given a phrase e.g.," the sweet potato box", the task is to find the best matching a set of images. Open-vocabulary object retrieval consist of imagenet, also focus on handling open-vocabulary and select the best match based on set of words.

In [19] C. Gan et al., review on zero-shot learning using semantic inter-class relationship in that actions are recognize automatically. Zero-shot learning use to detect actions without training of data. And holds semantic inter-class relationship is measured by continuous word space vectors. This method is fully automatic, result in save human tedious efforts also performance is increased for action detection.

In [20] Y. Yang et al., review on content-based semantic search in web video. propose a scalable solution by using content-based semantic search. Results are fast and accurate in content-based semantic search also maintain the retrieval performance.

In [21] A. Karpathy et al., presents a technique to generate natural language description of images and their regions. This approach holds both datasets of images and their sentence description. Deep visual-semantics for image description is based on convolution neural network. Finally it will generate description of visual data.

Table – 1 .ANAYSIS ON SEMANTIC QUERY TAG BASED VIDEO RETREIVAL

Sl. No	**Area of objective**	**Author**	**Year**	**Major contribution**	**Method Used**
1	Large collection of the	R. Socher *al.*	2009	Useful in object	Amazon

	image database system			recognition ,and automatic object clustering	Mechanical Turk
2	Self-paced learning for latent variable	P. Kumar *al.*	2010	Addresses important tasks in machine learning	Self-paced Learning
3	A dataset consist of several human action from videos in the wild	K. Soomro *et al.*	2012	Recognize human-object interaction, and identify action in playing games	Action recognition using bags of words
4	Zero-shot Video Retrieval of concepts	J. Dalton *al.*	2013	Automatically detect relevant concepts given intext query	Zero-shot video retrieval technique
5	Web Video Event are classified based on Fisher Vectors	C. Sun *et al.*	2013	Fisher vector used to describe the colour. And provide statistical representation.	Fisher Vectors representation
6	Distributed Represent-tation of words, phrases and their composition-nality	T. Mikolov *al.*	2013	It is efficient for vector representation to detect semantically related words	Distributed representation
7	Zero-shot learning using cross-modal	R. Socher *et al.*	2013	Gives high accuracy on unseen classes and seen classes.	Zero-shot learning
8	Event from video is extracted by using Semantic Signatures	M. Mazloom *al.*	2013	Semantic Signature representation use to capture the event from videos.	Late fusion and semantic signature
9	Visual denotation similarity metrics from image description	P. Young *et al.*	2014	It will generate graph to measure the similarity in the image description.	Visual denotations similarity metrics
10	Multi-modal fusion concept uses zero-shot event detection	S. Wu *et al.*	2014	It will generate the description of event	Multi-modal Fusion
11	Calibration for Event detection system	J. van Hout *al.*	2014	Used to understand human expression through languages	Late Fusion and Calibration
12	Grounded Compositional for describing images and their sentences	R. Socher *al.*	2014	Use dependency tree neural network to represent the image.	Dependency tree Recursive Neural Networks (DT-RNN)
13	Common Objects in Context	T.Y. Lin *et al.*	2014	Presents statistical analysis of the object.	Bounding box Representation
14	video event retrieval using tag	M. Mazloom *al.*	2014	Used to detect complex event in Video.	Tag-based Video retrieval
15	Composite Concept for zero-shot event detection	A. Habibian *et al.*	2014	Improves detection accuracy	Zero-shot detection uses Boolean logic operator
16	Self-paced reranking for multi -media search	L. Jiang *al.*	2014	Optimize the problem by using self-paced learning.	Self-paced learning
17	Event search using multimodal pseudo relevance feedback	T. Mitamura *et al.*	2014	Multimodal pseudo relevance feedback performs task for event retrieval.	Multimodal Pseudo relevance feedback
18	Open-Vocabulary object retrieval	S. Guadarrama *al.*	2014	Extremely useful for robotics application	Object retrieval from image to text
19	Zero-shot learning using semantic inter-class relationships	C. Gan*et al.*	2015	Automatically recognize action from zero-shot action.	Zero-shot learning

20	Content-based semantic search in web videos	Y. Yang *et al.*	2015	Maintain the semantic search.	Content-based Semantic Search
21	Deep Visual-Semantic for image description	A. Karpathy *et al.*	2015	Holds datasets of images and their sentence description	Deep Convolutional Neural Networks

III. CONCLUSION

The paper presents a technique of semantic query tag based video retrieval using continuous word space, and also overcome the semantic query gap, Which leads to neatly packed together video representation. Continuous word space gain significant results compare to dictionary space and concept space. Provide low latency. Retrieval performance is improved by training the data.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J.Li, K. Li, and L. FeiFei.Imagenet: A large-scale hierarchical image database. In CVPR, 2009.
[2] P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In NIPS, 2010.
[3] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CRCVTR- 12-01.
[4] J. Dalton, J. Allan, and P. Mirajkar.Zero-shot video retrieval using content and concepts.In CIKM.ACM, 2013.
[5] C. Sun and R. Nevatia.Large-scale web video event classification by use of fisher vectors.In WACV, 2013.
[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean.Distributed representations of words and phrases and their compositionality.In NIPS, 2013.
[7] R. Socher, M. Ganjoo,C. D. Manning, and A. Y. Ng. Zero shot learning through cross-modal transfer. In NIPS, 2013.
[8] M. Mazloom, A. Habibian, and C. G. M. Snoek. Querying for video events by semantic signatures from few examples.In MM. ACM, 2013.
[9] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL,
[10] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In CVPR, 2014
[11] J. van Hout, E. Yeh, D. Koelma, C. G. M. Snoek, C. Sun, R. Nevatia, J. Wong, and G. K. Myers. Late fusion and calibration for multimedia event detection using few examples.In ICASSP, 2014
[12] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y.Ng. Grounded compositional semantics for finding and describing images with sentences.TACL, 2014
[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.
[14] M. Mazloom, X. Li, and C. G. M. Snoek.Few-example video event retrieval using tag propagation.In ICMR, 2014.
[15] A. Habibian, T. Mensink, and C. G. M. Snoek.Composite concept discovery for zero-shot video event detection.In ICMR, 2014.
[16] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In MM. ACM, 2014.
[17] L. Jiang, T. Mitamura, S.-I.Yu, and A. G. Hauptmann. Zero example event search using multimodal pseudo relevance feedback.In ICMR.ACM, 2014.
[18] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval.In Robotics: Science and Systems, 2014.
[19] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G. Hauptmann.Exploring semantic inter-class relationships (sir) for zeroshotaction recognition.In ACAI, 2015.
[20] L. Jiang, S.-I.Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann.Fast and accurate content-based semantic search in 100m internet videos.In MM. ACM, 2015.
[21] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions.In CVPR, 2015.