EXPLORING THE INFLUENCE OF FEATURE REPRESENTATION FOR DICTIONARY SELECTION BASED VIDEO SUMMARIZATION

Mingyang Ma[1], *Shaohui Mei*[1,*], *Jingyu Ji*[1], *Shuai Wan*[1], *Zhiyong Wang*[2], *and Dagan Feng*[2]

[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China.
[2] School of Information Technologies, The University of Sydney, NSW 2006, Australia.

ABSTRACT

Dictionary selection based video summarization (VS) algorithms, in which keyframes are considered as a dictionary to reconstruct all the video frames, have been demonstrated to be effective and efficient for video summarization. It has been noticed that the feature representation of video plays a great impact of the performance of VS. In this paper, the influence of feature representation of video frames on the performance of dictionary selection-based VS is for the first time investigated. In addition to the traditional hand-crafted features used in VS, such as color histogram, the deep features learned through deep neural networks are firstly used to represent video frames for dictionary selection-based VS. The impact of dimensionality reduction to the high-dimensional deep learning features on VS is further discussed. Experimental results on a benchmark video dataset demonstrate that deep learning features are able to achieve better performance than traditional hand-crafted features for dictionary selection-based VS. Moreover, the dimensionality of deep learning features can be reduced to decrease the computational cost without the degradation of VS performance.

Index Terms— Video summarization, sparse reconstruction, feature representation, deep learning

1. INTRODUCTION

The explosive development of multimedia technologies results in a huge amount of video content. For example, as reported by Youtube Statistics[1], about 400 hours of new videos are uploaded to YouTube in just 60 seconds. Consequently, traditional sequential access to video content presents significant limitations for the new emerging multimedia services such as content-based search, retrieval, navigation and video browsing. Like producing headlines for news articles, video summarization (VS) has been a desirable solution for users to gain a quick comprehension of videos [1, 2].

In the past decades, many algorithms have been proposed for VS, such as clustering based methods [3, 4], sequence reconstruction error approaches [5, 6], dictionary selection based algorithms [7, 8, 9, 10], etc. among which dictionary selection based algorithms have been demonstrated as one of the most effective ones [7, 9]. In dictionary selection based VS algorithms, the keyframes are assumed to be a dictionary that all the video frames in a video can be well reconstructed by the atoms in it. For example, Cong et al. [7] formulated video summarization as a dictionary selection problem using sparsity consistency, then a dictionary of keyframes was selected by solving the convex optimization using Nesterov method. Mei et al. [8] formulated the VS problem as an $\ell_{2,0}$ constrained sparse dictionary selection model, and keyframes were selected by the simultaneous orthogonal matching pursuit (SOMP) based method without smoothing the penalty function. Afterwards, Mei et al. [9] reformulated the video summarization task with a minimum sparse reconstruction (MSR) problem by utilizing the true sparse constraint ℓ_0 norm such that keyframes were directly selected as a sparse dictionary. Cong et al. [10] recently designed two algorithms to solve the $\ell_{2,0}$ constrained dictionary selection model for VS: a standard greedy algorithm and a gradient cues for speeding up.

In VS problems, the features of video frames are extracted to form a feature pool and summarization algorithms are conducted on such a feature pool to extract keyframes or video clips. As a result, the performance of VS is not only affected by summarization algorithms, but also the feature representation of video frames. However, previous studies have mainly focused on how to summarize the videos frames effectively by using hand-crafted features. For example, the VSUMM algorithms used a color histogram to represent the visual content of video frames [4]. In the dictionary based algorithms [7, 9], each video frame is represented with a 360-dimensional feature vector that contains two parts: a 252-dimensional feature vector extracted by CENTRIST [11] which captures the local structures of a video frame without color information and a 108-dimensional feature accounting for the color moment. The influence of feature extraction on summarization has not been explored. Therefore, in this paper, different kinds of features are extracted for video frames and the influence of these

* Corresponding author. This work is partially supported by National Natural Science Foundation of China (61671383, 61371089) and Fundamental Research Funds for the Central Universities (3102016ZB012, 3102016ZY019). Email: meish@nwpu.edu.cn.

[1] YouTube, "http://www.youtube.com/".

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

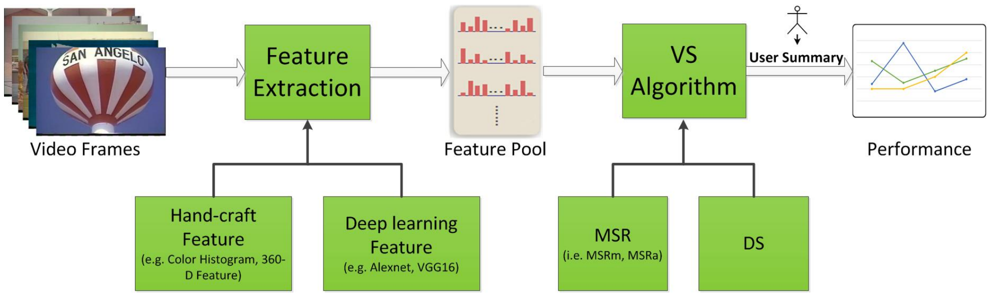


Fig. 1. Proposed scheme to exploring feature representation for VS.

features on dictionary selection based VS algorithms is investigated. Since feature learning by deep neural network has been demonstrated to be more effective than traditional handcrafted features in many applications, such as image classification and target detection, the features learned by recently proposed deep neural networks, including Alexnet [12], BN-Inception [13], Inception-v3 [14], and VGG16 [15], are also explored. Finally, experiments on a benchmark dataset have been conducted to evaluate the influence of various features on VS.

2. PROPOSED SCHEME

Keyframe extraction based VS has been proposed to select an optimal subset from the entire video frame pool through which the original video can be represented as accurate as possible. Therefore, we propose to explore feature representation for VS according to the scheme shown in Fig. 1. By applied feature extraction on video frames, a video is represented as a feature pool. In this paper, several feature extraction algorithms are used for feature representation for VS, including both traditional hand-crafted features and recently proposed deep learning features. Then VS algorithms are conducted on the feature pools to identify keyframes for VS. Three dictionary selection-based algorithms are adopted to produce video summaries. Finally, the automatic summaries (AS) generated by VS algorithms with different features are compared to the user summaries (US) to evaluate their performance.

2.1. Feature Representation of Video frames

In order to explore the influence feature representation for VS, both hand-crafted features and deep learning features are adopted to represent video frames.

2.1.1. Hand-crafted features

By investigating previous studies, two kinds of popular handcrafted features are adopted: color histogram-based features and the 360-dimensional feature used in dictionary selection-based algorithms [7, 9].

1). Color histogram-based feature

Since the HSV color space provides a more intuitive representation of color which is close to the way that humans apperceive and recognize color, the color histogram-based feature is extracted in the HSV color space. Only the histogram of hue component is used as that in the VSUMM algorithm [4]. Moreover, the quantization of the color histogram is set to 16 color bins for reducing significantly the amount of data without sacrificing important information.

2). The 360-dimensional feature

The 360-dimensional feature is a combination of CENTRIST and color features. The CENTRIST feature, which captures the local structures of a video frame without color information, is obtained by applying CENTRIST to image patches [11]. An image is processed by utilizing a spatial pyramid structure, and only the last two spatial levels are adopted and contain 5 and 1 image patches, respectively. Thus, each patch is represented by a 42-dimensional feature, where 40 dimensions are related to the eigenvectors and the other 2 dimensions are for the mean and variance of each patch. Therefore, the dimension of each CENTRIST feature is $(5 + 1) \times 42 = 252$. Since CENTRIST does not capture color information, the color moment in HSV color space is also adopted. Each frame is represented in HSV color space and partitioned into 3×4 patches. Then each image patch is described by 9 moments, i.e., 3-order color moments (i.e., Mean, Standard Deviation, and Skewness) for each of the 3 color channels (i.e., hue, saturation and value). As a result, each frame is represented with a $3 \times 4 \times 9 = 108$-dimensional color feature.

2.1.2. Deep learning features

In the last few years, deep learning has produced very good results on a variety of tasks, such as image classification [16, 17], object detection [18, 19], etc. Different with traditional feature extraction algorithms, deep learning can learn features automatically from the data itself. Generally, hundreds of thousands of parameters are involved in a deep neural network

(DNN) to explore the characteristic of big data. For example, in the convolutional neural network (CNN) proposed by Hinton et al. for ImageNet competition [12], more than 60, 000, 000 parameters are involved. All of these parameters can be optimized jointly with subsequent classifier to take full advantages of both feature extraction and classification. For example, Donahue et al. have shown that the output of last few layers of a pre-trained CNN can be used as a general visual feature descriptor for a variety of tasks [20]. Oquab et al. [21] and Simonyan et al. [15] used the outputs of the penultimate layer of a pre-trained CNN to represent full images of actions for action recognition, and achieve high performance.

Table 1. Details of the CNNs for feature extraction of video frames.

Network	Input size	Layer	Dimension
Alexnet	224 × 224 × 3	'fc6'	4096-D
BN-Inception	224 × 224 × 3	global_pool	1024-D
Inception-v3	229 × 299 × 3	global_pool	2048-D
VGG16	229 × 299 × 3	fc6	4096-D

The feature learned by CNNs has been testified to be of best performance, however it has been seldom utilized for VS. Therefore, in this paper, the deep feature representation learned by CNN models is proposed to represent the video frames for VS. The deep features of a video frame is extracted from the pre-trained networks, where an input image which is required to be converted to conform with network input is transformed into a high-level feature for various tasks. Four types of CNNs, including Alexnet [12], BN-Inception [13], Inception-v3 [14], and VGG16 [15], are selected to extract features from video frames. The details on the input size of images, the layer where the deep feature is extracted, and the dimension of a deep feature are listed in Table 1.

According to Table 1, the dimensionality of deep learning feature learned by CNNs is much greater than traditional features. As a result, it will take much more computational time when the deep learning feature is used for VS. Therefore, principle component analysis (PCA) is further used to conduct dimension reduction on the deep learning features.

2.2. VS algorithms

Dictionary selection-based VS algorithms, in which keyframes are extracted as a dictionary that all the video frames in a video can be well reconstructed by the atoms in it, have been demonstrated as one of the most effective algorithms among all the VS algorithms [7, 9]. As a result, three state-of-the-art dictionary selection-based VS algorithms are adopted for VS on feature pool of videos: including two types of minimum sparse reconstruction (MSR) algorithms (namely MSRa and MSRm) [9] and dictionary selection (DS) algorithm [10].

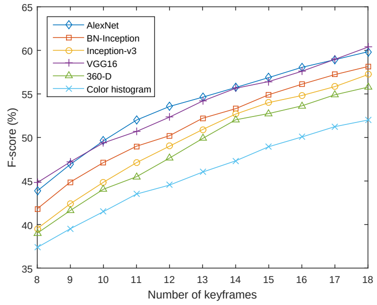


Fig. 2. Experimental results of MSRm with different features.

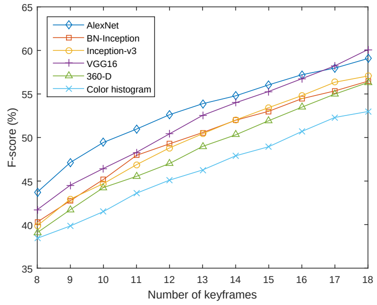


Fig. 3. Experimental results of MSRa with different features.

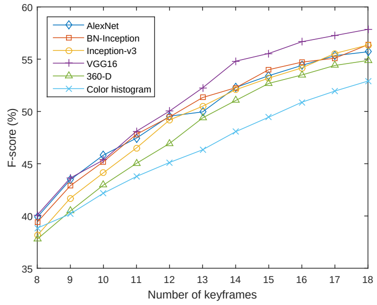


Fig. 4. Experimental results of DS with different features.

3. EXPERIMENTS AND DISCUSSION

3.1. Dataset

The dataset that contains 50 videos from the Open Video Project[2] (OVP) is used in our experiments. These videos are distributed among several genres (e.g., documentary, educational, ephemeral, historical, and, lecture) and their durations vary from 1 to 4 minutes (approximately 75 minutes in total), and the average frame rate is 30 frames per second (fps). In our experiments, each video is down-sampled to 5 fps. In all of the experiments, the number of selected keyframes in this dataset varies from 8 to 18. The metric F-score is calcu-

[2] Open Video Project, http://www.open-video.org/.

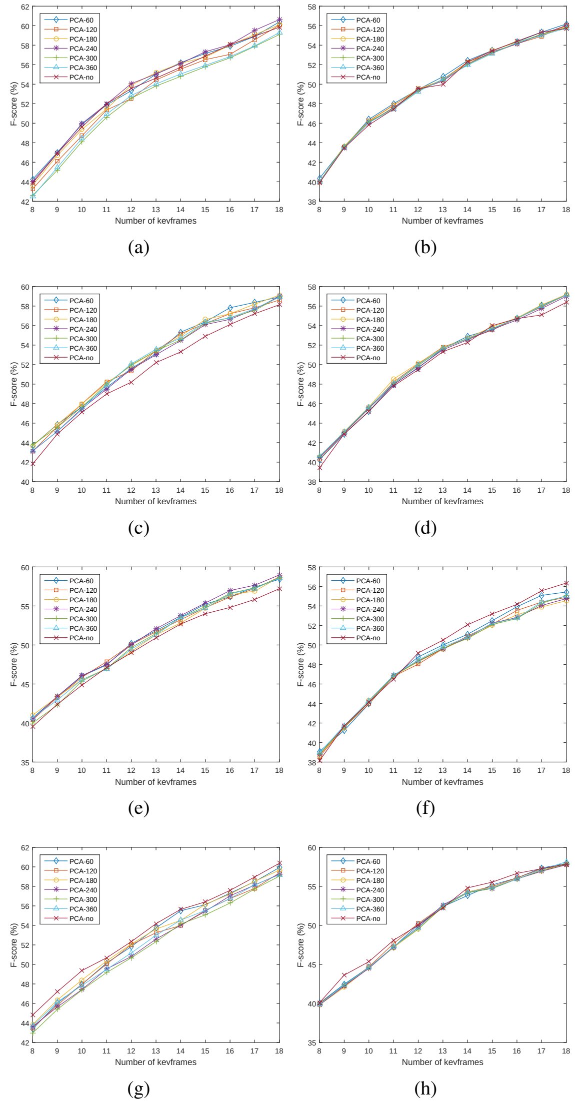
lated for quantitative evaluation by comparing the automatic summaries (AS) generated by different feature representation with five user summaries (US) which are available at VSUMM official website [4].

3.2. The VS results with different features

The experimental results of the three VS algorithms with different features are shown in Fig. 2, Fig. 3, and 4, respectively. Observed from Fig. 2, in the MSRm algorithm, the color histogram-based feature performs worst among all the considered features. The performance of 360-D feature is better than it. All the four types of deep learning features obviously outperform traditional hand-crafted features. Moreover, the features extracted by Alexnet [12] achieve best performance. Similar conclusion can be drawn for the MSRa algorithm according to the results shown in Fig. 3. When the DS algorithm is used for VS, it is also confirmed by this algorithm that the results using deep learning features are better than those using traditional hand-crafted features. The color feature only slightly outperforms Inception-v3 when the number of keyframes is very small. However, its performance become the worst when the the number of selected keyframes increases. It should also be noted that the features extracted by VGG16 achieve the best performance for the DS algorithm.

3.3. The results of dimension reduction on deep learning features

In this experiment, the dimensionality of deep learning features is reduced to 60, 120, 180, 240, 300, and 360, respectively. Since the results of MSRm and that of MSRa coincide with each other, only MSRm is adopted. Fig.5 lists the experimental results of dimension reduction on deep learning features. It is observed that, the performance of VS does not degrade or slightly degrade when the deep learning features are projected to a low-dimensional space under most cases. Under some circumstances, e.g., the MSR with BN-Inception or Inception-v3 features shown in Fig. 5 (c) and (e), the performance of VS can even be slightly improved by the PCA based projection. On the other hand, by projecting deep learning features into a low-dimensional spaces, the computational time can be significantly saved. For example, when MSR algorithms are used for VS, about 99% of computational time is saved when the Alexnet, Inception-v3, and VGG16 features are projected to 360 or smaller. Even when the DS algorithm is used, more than 85% of computational time is saved for Alexnet and VGG16 features.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5. Experimental results of dimension reduction of deep learning features: (a) MSR with ALEXNET; (b) DS with ALEXNET; (c) MSR with BN-Inception; (d) DS with BN-Inception; (e) MSR with Inception-v3; (f) DS with Inception-v3; (g) MSR with VGG16; (h) DS with VGG16;

4. CONCLUSION

In this paper, the influence of feature representation on dictionary selection-based VS is explored by using different features, including traditional hand-crafted features and recently proposed deep learning features. Moreover, the impact of dimension reduction on deep learning features for VS is discussed. Experimental results of three dictionary selection-based algorithms on a benchmark video dataset demonstrate that the deep learning features clearly outperform traditional hand-crafted features for VS and the dimensionality of these features can be further reduced to save computational time for VS without compromising summarization accuracy .

5. REFERENCES

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 37, 2007.

[2] A. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, pp. 121–143, 2008.

[3] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings of International Conference on Image Processing*. IEEE, 1998, vol. 1, pp. 866–870.

[4] S. E. F. de Avila and *et al.*, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[5] H. Lee and S. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Processing: Image Communication*, vol. 18, no. 1, pp. 1–15, 2003.

[6] T. Liu, X. Zhang, J. Feng, and K. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1451–1457, 2004.

[7] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[8] S. Mei, G. Guan, Z. Wang, M. He, X. Hua, and D. Feng, "L2,0 constrained sparse dictionary selection for video summarization," in *IEEE International Conference on Multimedia and Expo*, 2014, pp. 1–6.

[9] S. Mei, G. Guan, Z. Wang, and et al., "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522–533, 2015.

[10] Y. Cong, J. Liu, G. Sun, and et al., "Adaptive greedy dictionary selection for web media summarization," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 185–195, 2017.

[11] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Computer Science*, 2015.

[14] Szegedy C., Vanhoucke V., Ioffe S., and et al., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv: 1409.1556*, 2014.

[16] D. C. An, U. Meier, and et al., "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2011)*, 2011, pp. 1237–1242.

[17] F. Huang and Y. Lecun, "Large-scale learning with svm and convolutional for generic object categorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 284–291.

[18] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *IEEE International Conference on Computer Vision*, 2013, pp. 2056–2063.

[19] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2155–2162.

[20] J. Donahue, Y. Jia, and et al., "Decaf: A deep convolutional activation feature for generic visual recognition.," in *International Conference on Machine Learning*, 2014, vol. 32, pp. 647–655.

[21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.