



A State-of-Art Review on Automatic Video Annotation Techniques

Krunal Randive^(✉) and R. Mohan

Department of Computer Science and Engineering,
National Institute of Technology, Tiruchirappalli 620015, Tamil Nadu, India
krunalrandive@gmail.com, rmohan@nitt.edu

Abstract. Video annotation has gained attention because of the rapid development of video information and wide usage of video analysis in all directions. With the capacity of depicting video at the semantic level, video annotation has numerous applications in video analysis. Due to the shortcomings present in manual video annotation, Automatic Video Annotation was introduced. In this paper, distinctive methodologies of automatic video annotation are discussed. These models are classified into five classes namely, (1) Generative models, (2) Distance-based similarity model, (3) Discriminative model, (4) Ontology-based models, (5) Deep Learning-based models. The key theoretical contributions in the current decade in support of video annotation strategies are discussed. Additionally, the future directions concerning the research aspect of video annotation strategies are discussed.

Keywords: Automatic Video Annotation (AVA) · Deep learning · Ontology · Feature extraction · Convolutional Neural Network (CNN)

1 Introduction

With the recent development of online video data sharing platforms, for example, YouTube, the rate of growth and collection of video data have developed quickly and possess a major impact in day-to-day life. By the conventional video annotation systems marking video captions or comments at the semantic level, are not feasible with enormous data available. The primary disadvantage of manual video annotations is intuitionistic. It is difficult to annotate on a large amount of video information manually. This leads to uncertainty over video content. Hence, different people may infer different ways of the understanding video. This happens due to the differences in qualification, thinking perspective and instructive foundation.

At present, a general grouping and profound survey of Automatic Video Annotation (AVA) techniques are yet lacking. Although some reviews of Automatic Image Annotation strategies, the foci were set on Content-Based Image Retrieval (CBIR) [1–3, 20, 21, 23, 26], include extraction and semantic annotations [5–7, 12–18, 22, 24, 25, 27, 44], statistical methodologies [4, 8, 9], segmentation based [10, 11], and deep learning [28–43]. AVA has variety of applications such as AVA for archival sports videos [45], media quality assessment [46, 48], and video content analysis [47]. In this paper, various AVA methods are classified and discussed.

In recent days, AVA techniques have gained attention into a new dimension because of the shortcoming existing in the traditional annotation model. In [1], method driven by the multiple Bernoulli relevance models is proposed to carry on the annotating video by unsupervised model. These accomplishments have supported the advancement of video annotation techniques. Similarly, the annotation for the input video can be predicted from the annotated dataset through the unsupervised mode or automatically.

The main purpose of the annotation technique is to bridge the semantic gap between low-level features in the video and high-level semantic labels. Annotation can encourage video search using the content. Content-based video retrieval (CBVR) can be considered semantically significant than to perform without any query because the automated mapping between labels and video frames can be trusted [1–3]. The process of Automatic video annotation is presented in Fig. 1.

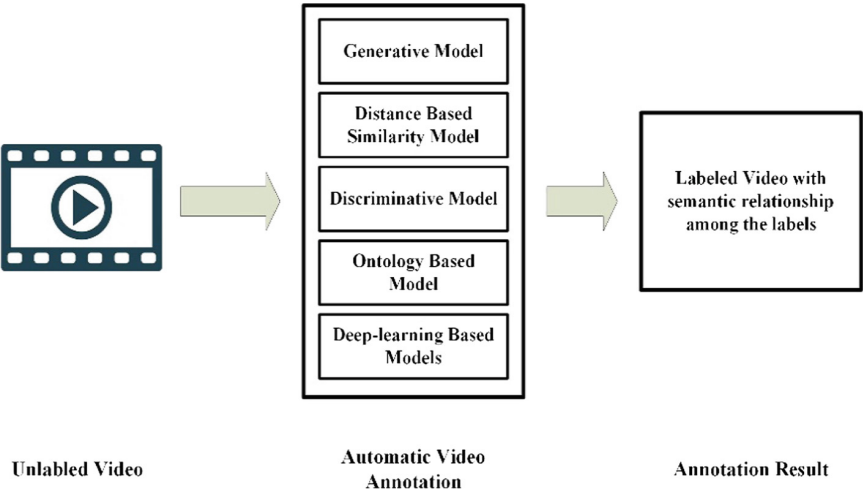


Fig. 1. The process of Automatic Video Annotation

In this paper, five AVA strategies are illustrated. The taxonomy for the model-based techniques is shown in Fig. 2. The generative models are discussed in Sect. 2. Section 3 discuss about the distance-based similarity model, discriminative models are discussed in Sect. 4. The ontology-based models are discussed in Sect. 5, and deep learning-based models are discussed in Sect. 6. Section 7 discusses about a different framework, and challenges in annotation models. The previously mentioned five classifications of AVA strategies can be additionally arranged into a few sub-classes as indicated by their fundamental ideas. Provides a taxonomy for Automatic Video Annotation techniques.

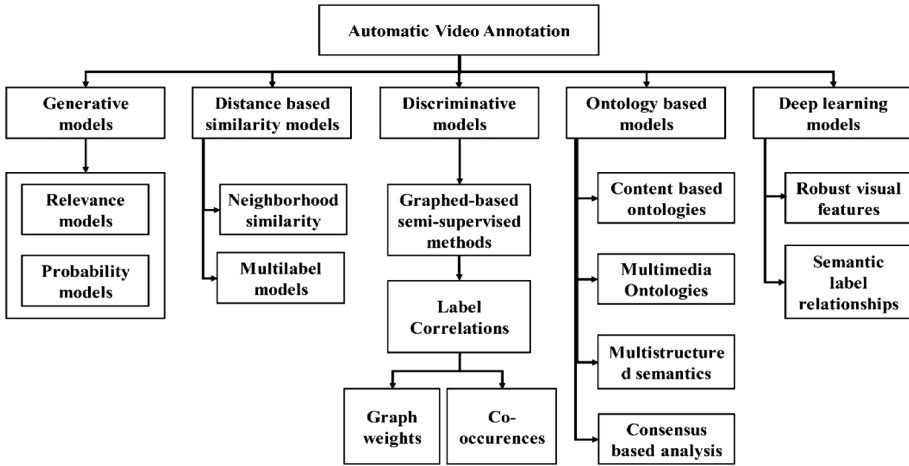


Fig. 2. Taxonomy for Automatic Video Annotation techniques.

2 Generative Model-Based Techniques

The generative model-based annotation techniques remain very prevalent, and good in an early 21st century. These techniques are committed to boost the probability of video features and tags. For an unannotated video, these techniques compute a joint probabilistic model of features extracted from video and labels to find the likelihood of a video tag using training dataset. The generative models utilized for AVA comprise of the relevance models. In probability and statistics, generative models characterize arbitrarily generating recognizable information, regularly provides some unseen parameter. It indicates a joint probability distribution over labeled sequence and observations. Generative models utilized machine learning for presenting data directly.

For the given input annotated training set, the authors in [2] have demonstrated probabilistic models. It enables us to identify the likelihood of creating a tag from the features in the video frames. In [3] the authors presented a video annotation technique, called as a double cross-media relevance model (DCMRM). It assesses the joint probability through the accumulation of tags in a pre-characterized dictionary. It includes two basic relations in video annotation. Namely, the word-to-video relationship and word-to-word relationship. In [4] the authors presented a procedure to combine the results of multiple trackers that includes evaluating the error statistics of the individual trackers by comparing their outcome to a reliable people identifier.

Outstanding contributions made by the generative models to the improvement of annotation techniques. Nevertheless, there exist challenges in the generative models-based strategies.

1. Generative models are weak to optimize the tag prediction.
2. Unable to recognize the complex semantic relationship.
3. They require high computationally complex algorithms.

3 Distance-Based Similarity Model-Based Techniques

The distance-based similarity assumes videos with similar tags for annotation. For a sample video, these techniques enable a search for a set of same videos, and the labels of that video are depending on its similarity.

In [5], the authors proposed a beyond distance measurement to manage the deficiency of the training data in video annotation by building neighborhood similarity. In [6], the authors have presented an application of event detection from video range from reducing the semantic gap. The inferred metric can be utilized in a CBVR system to extract the semantically similar video. Nearest neighbor learning takes distance metric from the given set of labeled points and marks the relation between distance in the training data.

In [7], the author not only tried to filter the semantic inferences of near- scenes for video annotation but also to retain annotations by using a hierarchical video-to-near scene annotation framework. A strategy for AVA given multi-label learning and multi-feature combination k-nearest neighbor algorithm was proposed in [8]. Amid the previous decade, several research efforts have been done on the best way to exploit the semantic correlations to enhance the image and video annotation, at the same time classifies concepts and represents them using a Correlative Multi-Label (CML) system [9].

The Distance based similarity model-based AVA methods are structure-intuitive. Due to the high flexibility of the model, tag prediction is quite successful. However, because of some intrinsic shortages, more improvements are expected.

4 Discriminative Model-Based Techniques

Video annotation is considered as the multi-label classification problem in the discriminative model-based annotation techniques. It is addressed with the use of binary classifier. To anticipate the labels for unlabeled videos, an independent binary classifier is used for every label. So, this study leads to video annotation using SVM [11], tag ranking system [13], and semi-supervised learning method [15].

In [11], the authors have trained a target classifier using the web video sets that adapt the pre-learned classifiers. This method is called as an adaptive latent structured SVM. Here the locations of the key segment are considered as latent variables. In [14], the authors presented the AVA technique, which is unsupervised leveraging video content. This method uses the search procedure followed by mining. The video content and the audio identified transcriptions are given as the queries, and the multimodal search ranks the relevant videos.

Recently, semi-supervised learning is used extensively in research. It is predominantly dedicated to the graph-based learning techniques. It visualizes entire data as the graph. In this propagation-based technique, the correlation between the labels can be effectively combined. This correlation between the labels is utilized in two ways in the graph based learning technique. In [15], authors have proposed a technique, named as optimized multigraph-based semi-supervised learning (OMG-SSL). The technique tends to tackle different deciding aspects in the video annotation simultaneously. Multiple methods and temporal consistency are included as deciding aspects. It

compares to the various relationship among video units and represented by different graphs. A semi-supervised annotation approach, named as leaning an optimized graph (OGL). The relationship between the data points can be embedded more accurately from incomplete labels and multiple features in the proposed approach [17, 18]. In [19], the authors proposed an approach for sentence generation from videos.

Despite the productivity and intuitionistic advantages of discriminative model-based techniques, there exist challenges, which are as follows:

1. The interrelationship among the visual features and tags is often excluded.
2. These techniques are often are receptive to the quality of training datasets, because of the predominant usage of label correlations.

The graph-based video techniques are transitive and can determine the labels of the unannotated video. However, this is not useful in practical applications consisting of the large video annotation tasks.

5 Ontology-Based Techniques

Concept relationships are proven to be important learning assets that can improve the effectiveness of video annotation. Ontology-based annotation techniques denote the theoretical representations of the semantic relationship between the feature present in video and labels. In object-ontology, semantics descriptors are formulated from our daily language. These semantic descriptors from a simple vocabulary give a qualitative meaning of high-level queries. Video dataset can be grouped, based on the high-level semantics (keywords) and mapping of semantic descriptors [24, 25].

An Ontology-based structured annotation technique is proposed in [20, 23] for CBVR using spatiotemporal Reasoning. These rule-based systems are efficient in describing the temporal information of events and activities in videos. In [21], the web video search engine was proposed. The combination of Boolean and temporal relations for the automatic annotation and retrieval of video content is utilized by the ontologies and the permitted queries. A framework for automatic semantic video annotation of unconstrained videos is presented in [22]. This framework combines two layers, namely annotation analysis using commonsense knowledgebase and low-level visual similarity matching. Commonsense ontology is framed by consolidating multiple-structured semantic relationships.

In [24], the authors formalized the method for analyzing the content in multimedia using the interrelationship between the low- and high-level concept descriptors. In [25], the authors demonstrated the fuzzy knowledge representations. The probability or the degree of uncertainty is determined from these representations. The fuzzy Ontology model populates the automatically generated ontology from the different video annotation datasets. The content was utilized for the improvement of video semantic interpretation.

In [26], the authors proposed a general block schematic for ontology-based video annotation. The usage of ontology-based technique helps in the semantic annotation. Either rule learning or machine learning, both together with ontology can give better annotation compared with different techniques In [27], the authors proposed a

collaborative algorithm, which consensus-based social network analysis (SNA) for semantic video annotation. In this method, the videos are sorted out similar to social networks. In this work, the content present in the video was described semantically utilizing an ontology-based approach. Simple semantics inferred from the ontology works fine for the datasets consisting of a similar type of videos. However, they may fail to annotate the dataset with different videos. So, tools that are more powerful are required to learn semantics for substantial accumulations of videos with different activities.

6 Deep-Learning Model-Based Techniques

Recently, a feature representation for the video annotation is driven by deep learning-based techniques because of the significant development in deep architectures. This allows various deep architectures to annotate large volume of video data available. The techniques are divided into two categories. The Convolution neural network (CNN) can be used for extracting features that are robust [28–30, 35]. Deep learning systems are utilized for the semantic label relationship [41, 42].

Robust visual features are the primary component for video annotation. The conventional handcrafted features are not reliable for complex concepts. The recent research works available generates robust visual features utilizing CNN because of the meritorious behavior of CNN in computer vision. In [31], a multi-semantic video annotation technique is proposed to represent the high-level semantic information using the semantic network and to map the relationships between the content. The keyframes are extracted from the video and CNN's is used to extract low-level visual features, which recognize the content in the videos. In [32], the authors have proposed a method that exhibits another way to deal with weakly labeled sequence data using the learning in a frame-based classifier by embedding a CNN in an iterative EM algorithm. An active learning method based on visualization is proposed in [33] for video annotation. In [34], the authors proposed a multimodal feature learning mechanism based on Stacked Contractive Autoencoder (SCAE) deep networks for automatic video annotation. In [36], the proposed method is utilized to generate video representation with an emphasis on temporal modeling namely Hierarchical Recurrent Neural Encoder (HRNE).

The CNN-RNN framework [40] uses Recurrent Neural Networks (RNN) with a moderate level of computational complexity to identify high-order label relationships. The framework uses convolutional and recurrent neural network mutually used for determining the correlation between the adjacent labels and video representations. The final outputs are computed based on the similarities among labels. In [37, 38], a combinational deep neural networks based two-stream structure is proposed. The structure is made out of two segments. The parallel two-stream encoding segment which 3D CNN and takes video encoding from different sources. The input encoded video representations are transferred using Long-Short-Term-Memory (LSTM)-based decoding language to the annotations. In [39], the authors have proposed a hierarchical LSTM with adjusted temporal attention (hLSTMat) approach for video annotation. In [43], a framework is presented which simultaneously explores the learning of LSTM and visual-semantic embedding (LSTM-E). In [44], a deep architecture is presented,

which combines the transferred semantic attributes gained from images and video to the CNN plus RNN framework (LSTM-TSA).

It can be inferred that the deep learning-based annotation techniques have conveyed both challenges and chances to AVA techniques. The late advancement and “leaps forward” in deep learning assures to enhance the performance of video annotation on large-scale video datasets. Yet there are two drawbacks present in the deep learning-based annotation techniques. The main drawback is the decline in the efficiency of these video annotation techniques because of the increase in the hidden layers of the deep neural network. The absence of complete theoretical guidance and unified structure to decide the network structure leads to uncontrolled training phase involved in the deep neural network.

7 Discussion and Conclusions

In this paper, five types of AVA techniques are discussed regarding its framework, merits, and challenges. The deep learning-based, the discriminative model-based, and generative model-based annotation methods fall under the category of learning-based techniques. In the discriminative model-based annotation technique, video annotation is considered as a multi-label classification problem. Hence, the binary classifier cannot determine the relation between the existing classes. Robust visual features are extracted utilizing CNN in the deep learning-based techniques. They utilize alternate frameworks, such as RNN, to infer the semantic label relationship, for automatic video annotation. Comparatively, the distance-based similarity model-based technique follows a two-step framework. The prediction is done based on similar videos. The implementation of applications with simple semantic features can be done easily using Ontology-based algorithms. However, to learn more complex semantics, machine-learning techniques are required. A decision tree is a suitable technique for video annotation and retrieval because of the ease in use for implementation. These decision rules are used for intuitive mapping from low-level features to high-level concepts. It can be inferred that the five type of AVA techniques relies on the different ways to bridge the semantic gap by analyzing the semantic context.

References

1. Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1002–1009 (2004)
2. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119–126 (2003)
3. Liu, J., Wang, B., Li, M., et al.: Dual cross-media relevance model for image annotation. In: Proceedings of the 15th International Conference on Multimedia, pp. 605–614 (2007)
4. Niño-Castañeda, J., Frías-Velázquez, A., Bo, N.B., Slembrouck, M., Guan, J., Debar, G., Vanrumste, B., Tuytelaars, T., Philips, W.: Scalable semi-automatic annotation for multi-camera person tracking. *IEEE Trans. Image Process.* **25**(5), 2259–2274 (2016)

5. Wang, M., Hua, X.S., Tang, J., Hong, R.: Beyond distance measurement: constructing neighborhood similarity for video annotation. *IEEE Trans. Multimed.* **11**(3), 465–476 (2009)
6. Wang, C., Zhang, L., Zhang, H.J.: Learning to reduce the semantic gap in web image retrieval and annotation. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 355–362 (2008)
7. Chou, C.L., Chen, H.T., Lee, S.Y.: Multimodal video-to-near-scene annotation. *IEEE Trans. Multimed.* **19**(2), 354–366 (2017)
8. Xia, S., Chen, P., Zhang, J., Li, X., Wang, B.: Utilization of rotation-invariant uniform LBP histogram distribution and statistics of connected regions in automatic image annotation based on multi-label learning. *Neurocomputing* **228**, 11–18 (2017)
9. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.: Correlative multi-label video annotation. In: *Proceedings of the 15th ACM International Conference on Multimedia*, pp. 17–26 (2007)
10. Jain, S.D., Grauman, K.: Click carving: segmenting objects in video with point clicks (2016). arXiv preprint: [arXiv:1607.01115](https://arxiv.org/abs/1607.01115)
11. Song, H., Wu, X., Liang, W., Jia, Y.: Recognizing key segments of videos for video annotation by learning from web image sets. *Multimed. Tools Appl.* **76**(5), 6111–6126 (2017)
12. Schöning, J., Faion, P., Heidemann, G., Krumnack, U.: Providing video annotations in multimedia containers for visualization and research. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 650–659 (2017)
13. Shah, R., Zimmermann, R.: Tag recommendation and ranking. In: *Multimodal Analysis of User-Generated Multimedia Content*, pp. 101–138 (2017)
14. Moxley, E., Mei, T., Hua, X.S., Ma, W.Y., Manjunath, B.S.: Automatic video annotation through search and mining. In: *2008 IEEE International Conference on Multimedia and Expo*, pp. 685–688 (2008)
15. Wang, M., Hua, X.S., Hong, R., Tang, J., Qi, G.J., Song, Y.: Unified video annotation via multigraph learning. *IEEE Trans. Circ. Syst. Video Technol.* **19**(5), 733–746 (2009)
16. Schöning, J., Faion, P., Heidemann, G.: Pixel-wise ground truth annotation in videos. In: *ICPRAM*, vol. 6, p. 11 (2016)
17. Song, J., Gao, L., Nie, F., Shen, H.T., Yan, Y., Sebe, N.: Optimized graph learning using partial tags and multiple features for image and video annotation. *IEEE Trans. Image Process.* **25**(11), 4999–5011 (2016)
18. Gao, L., Song, J., Nie, F., Yan, Y., Sebe, N., Tao Shen, H.: Optimal graph learning with partial tags and multiple features for image and video annotation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4371–4379 (2015)
19. Qian, X., Liu, X., Ma, X., Lu, D., Xu, C.: What is happening in the video?—Annotate video by sentence. *IEEE Trans. Circ. Syst. Video Technol.* **26**(9), 1746–1757 (2016)
20. Sikos, L.F.: Ontology-based structured video annotation for content-based video retrieval via spatiotemporal reasoning. In: *Bridging the Semantic Gap in Image and Video Analysis*, pp. 97–122. Springer, Cham (2018)
21. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Video annotation and retrieval using ontologies and rule learning. *IEEE Multimed.* **17**(4), 80–88 (2010)
22. Altadmri, A., Ahmed, A.: A framework for automatic semantic video annotation. *Multimed. Tools Appl.* **72**(2), 1167–1191 (2014)
23. Sikos, L.F.: RDF-powered semantic video annotation tools with concept mapping to linked data for next-generation video indexing: a comprehensive review. *Multimed. Tools Appl.* **76**(12), 14437–14460 (2017)

24. Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., Strintzis, M.G.: Semantic annotation of images and videos for multimedia analysis. In: European Semantic Web Conference, pp. 592–607 (2005)
25. Zarka, M., Ammar, A.B., Alimi, A.M.: Fuzzy reasoning framework to improve semantic video interpretation. *Multimed. Tools Appl.* **75**(10), 5719–5750 (2016)
26. Khurana, K., Chandak, M.B.: Study of various video annotation techniques. *Int. J. Adv. Res. Comput. Commun. Eng.* **2**(1), 909–914 (2013)
27. Duong, T.H., Nguyen, N.T., Truong, H.B., Nguyen, V.H.: A collaborative algorithm for semantic video annotation using a consensus-based social network analysis. *Expert Syst. Appl.* **42**(1), 246–258 (2015)
28. Wang, Y., Luo, Z., Jodoin, P.M.: Interactive deep learning method for segmenting moving objects. *Pattern Recogn. Lett.* **96**, 66–75 (2017)
29. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
30. Wu, Z., Yao, T., Fu, Y., Jiang, Y.G.: Deep learning for video classification and captioning (2016). arXiv preprint: [arXiv:1609.06782](https://arxiv.org/abs/1609.06782)
31. Yu, S., Cai, H., Liu, A.: Multi-semantic video annotation with semantic network. In: 2016 International Conference on Cyberworlds (CW), pp. 239–242, September 2016
32. Koller, O., Ney, H., Bowden, R.: Deep hand: how to train a CNN on 1 million hand images when your data is continuous and weakly labelled. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3793–3802 (2016)
33. Liao, H., Chen, L., Song, Y., Ming, H.: Visualization-based active learning for video annotation. *IEEE Trans. Multimed.* **18**(11), 2196–2205 (2016)
34. Liu, Y., Feng, X., Zhou, Z.: Multimodal video classification with stacked contractive autoencoders. *Signal Process.* **120**, 761–766 (2016)
35. Maharaj, T., Ballas, N., Rohrbach, A., Courville, A.C., Pal, C.J.: A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In: CVPR, pp. 7359–7368 (2017)
36. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1029–1038 (2016)
37. Zhang, C., Tian, Y.: Automatic video description generation via LSTM with joint two-stream encoding. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2924–2929 (2016)
38. Torabi, A., Tandon, N., Sigal, L.: Learning language-visual embedding for movie understanding with natural-language (2016). arXiv preprint: [arXiv:1609.08124](https://arxiv.org/abs/1609.08124)
39. Song, J., Guo, Z., Gao, L., Liu, W., Zhang, D., Shen, H.T.: Hierarchical LSTM with adjusted temporal attention for video captioning (2017). arXiv preprint: [arXiv:1706.01231](https://arxiv.org/abs/1706.01231)
40. Jiang, H., Lu, Y., Xue, J.: Automatic soccer video event detection based on a deep neural network combined CNN and RNN. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 490–494 (2016)
41. Karayil, T., Blandfort, P., Borth, D., Dengel, A.: Generating affective captions using concept and syntax transition networks. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 1111–1115 (2016)
42. Ashangani, K., Wickramasinghe, K.U., De Silva, D.W.N., Gamwara, V.M., Nugaliyadde, A., Mallawarachchi, Y.: Semantic video search by automatic video annotation using TensorFlow. In: Manufacturing & Industrial Engineering Symposium (MIES), pp. 1–4 (2016)

43. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4594–4602 (2016)
44. Pan, Y., Yao, T., Li, H., Mei, T.: Video captioning with transferred semantic attributes. In: *CVPR*, vol. 2, p. 3 (2017)
45. Xue, Y., Song, Y., Li, C., Chiang, A.T., Ning, X.: Automatic video annotation system for archival sports video. In: *2017 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 23–28 (2017)
46. Zhang, L., Hong, R., Nie, L., Hong, C.: A biologically inspired automatic system for media quality assessment. *IEEE Trans. Autom. Sci. Eng.* **13**(2), 894–902 (2016)
47. Loukas, C.: Video content analysis of surgical procedures. *Surg. Endosc.* **32**(2), 553–568 (2018)
48. Hudelist, M.A., Husslein, H., Münzer, B., Kletz, S., Schoeffmann, K.: A tool to support surgical quality assessment. In: *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pp. 238–239 (2017)