Video Description: A Survey of Methods, Datasets and Evaluation Metrics

Nayyer Aafaq, Syed Zulqarnain Gilani, Wei Liu, and Ajmal Mian

Abstract—Automatic video description is useful for assisting the visually impaired, human computer interaction, robotics and video indexing. The past few years have seen a surge of research interest in this area due to the unprecedented success of deep learning in computer vision and natural language processing. Numerous methods, datasets and evaluation measures have been proposed in the literature calling the need for a comprehensive survey to better focus research efforts in this flourishing direction. This paper answers exactly to this need by surveying state of the art approaches including deep learning models; comparing benchmark datasets in terms of their domain, number of classes, and repository size; and identifying the pros and cons of various evaluation metrics such as BLEU, ROUGE, METEOR, CIDEr, SPICE and WMD. Our survey shows that video description research has a long way to go before it can match human performance and that the main reasons for this shortfall are twofold. Firstly, existing datasets do not adequately represent the diversity in open domain videos and complex linguistic structures. Secondly, current measures of evaluation are not aligned with human judgement. For example, the same video can have very different, yet correct descriptions. We conclude that there is a need for improvement in evaluation measures as well as datasets in terms of size, diversity and annotation accuracy because they directly influence the development of better video description models. From an algorithmic point of view, diagnosis of the description quality is challenging because of the difficultly to assess the level of contribution from visual features compared to the bias that comes naturally from the language model adopted.

Index Terms—Deep learning, video description, video captioning, language in vision, video captioning datasets, video captioning evaluation metrics, BLEU, METEOR, ROUGE, CIDEr, SPICE, WMD.

1 Introduction

Describing a short video in words is a trivial task for most people. On the other hand, automatic generation of natural language descriptions of videos is a challenging task for machines. Automatic video description involves the understanding of many background concepts and the detection of their occurrences in a video such as *objects*, *actions*, *scenes*, *person-object relations*, and the *temporal order of events*. Moreover, it requires translation of the extracted information into a comprehensible and grammatically correct natural language narrative.

Over the past few years, the two traditionally independent fields, Computer Vision (CV) and Natural Language Processing (NLP) have joined forces to address the upsurge of research interests in understanding and describing image and video contents. Special issues of journals are published focusing on langauge in vision [1] and workshops uniting the two areas have also been held regularly at both NLP and CV conferences [2], [3], [4], [5], [6], [7].

Leveraging the recent developments in deep neural networks for NLP and CV, and the availability of large multi-modal datasets, automatically generating stories from pixels is no longer a science fiction. This growing body of work has mainly originated from the robotics community and can be labeled broadly as *language grounded meaning from vision to robotic perception* [8], [9]. Related research areas include, connecting words to pictures [10], [11], [12], narrating images in natural language sentences [13], [14], [15] and understanding natural language instructions for robotic applications [16], [17], [18]. A closely related field is information retrieval. Thanks to the release of benchmark datasets MSCOCO [19] and Flicker30k [20], research in *image captioning and retrieval* [21], [22], [23], [24], and *image question answering* [25], [26], [27], [28] has also become very active.

Automatic video description is relatively more challenging because not all objects detected in the video are relevant to the description. The detected objects that do not play a role in the observed activity should be considered as irrelevant [29]. Moreover, video description methods should additionally capture the speed, direction of relevant objects as well as causality among events, actions and objects. Finally, events in videos may span across multiple time scales and may even overlap [30]. For example while piano recitals might last for the entire duration of a long video, the applause only takes place at the end, as shown in the example of Figure 1. The example illustrates differences between three closely related areas of research, namely, image captioning, video captioning and video description with context. In this example, image captioning techniques recognize the event as mere *clapping* whereas it is actually an *applause* that resulted from a previous event - piano playing.

Figure 2 summarizes related research under the same umbrella of *Visual Content Description*. The classification is based on whether the input is still images (*Image Captioning*) or multi-frame short videos (*Video Captioning*). Note, however, that short video captioning is very different from video auto-transcribing where audio and speeches are the main focus. On the other hand, the main focus of video captioning is on the visual content as opposed to the audio signals. In

• *N. Aafaq, S. Z. Gilani, W. Liu and A. Mian are with the Department of Computer Science and Software Engineering (CSSE), University of Western Australia (UWA), WA, 6009.*
E-mail: nayyer.aafaq@research.uwa.edu.au, [zulqarnain.gilani, wei.liu, ajmal.mian]@uwa.edu.au

Fig. 1: Illustration of how video description is different from image captioning. Within video description tasks, we also highlight the differences between description with context and without context

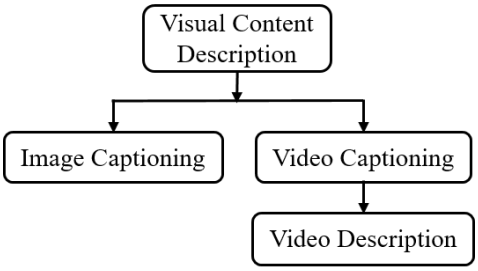


Fig. 2: Classification of visual content description.

particular, *Video Description* extends video captioning with the aim to provide a more detailed account of the visual contents in the video, such as objects, actions, scenes, and the order of events.

Below we define some terminologies used in this paper.

- *Visual Content Description*: The unifying concept of automatically generating natural language sentences to describe the visual content in still images or video clips.
- *Image Captioning*: Automatically generating a single sentence [21], [22], [31], [32], [33], [34], [35], [36] or multiple sentences [37], [38] that describe a still image.
- *Video Captioning*: Automatically generating natural language sentences to describe the contents of an input video clip that contains only one main event.
- *Video Description*: Automatically generating one or more natural language sentences that provide a narrative of a relatively longer video clip that possibly contains more than one event. The descriptions are more detailed and may be in the form of paragraphs. Video captioning is sometimes also referred to as *video summarization*.
- *Dense Video Captioning*: The detection and description of events in a video when multiple events of various timespan (that may or may not overlap) occur. It not only describes videos but also localizes events in time.

A popular approach in the current body of video description work involves the detection of Subject, Verb and Object, known as the *SVO-Triplet* approach [29], [39]. More recent works make use of *Deep Learning* models [40], [41] to build end-to-end systems resembling a typical neural machine translation system. The output of both approaches is either a single sentence [42], [43], or multiple sentences [29], [44], [45], [46], [47], [48] per video clip.

The majority of current literature on video description focuses on domain specific short video clips with limited vocabularies of objects and activities [29], [35], [39], [47], [49], [50], [51]. Description generation for open domain and relatively longer videos remains a challenging problem, as it needs large vocabularies and training data. The lack of rich models that can learn a joint feature space of a sequence of frames and the corresponding sequence of words is also a potential bottle neck in achieving the state of the art performance.

When it comes to relative performance, quantitative evaluation of video description systems is not a straightforward task. Currently, automatic evaluations are typically performed using machine translation and image captioning

metrics such as BLEU [52], ROUGE [53], METEOR [54], CIDEr [55] and the recently proposed SPICE [56] and WMD [57] metrics. Section 6.1 gives details of these measures. Here, we give a brief overview to establish motivation for our survey. BLEU is precision-based and only checks for exact matches of *n-grams* in the predicted and groundtruth references. METEOR, on the other hand, first creates an alignment between two sentences by comparing exact tokens, stemmed tokens and paraphrases. It also takes into consideration the semantically similar matches using WordNet synonyms. ROUGE, similar to BLEU, has different *n-grams* based versions and computes recall for the generated sentences and the reference sentences. CIDEr is a human-consensus-based evaluation metric which is developed specifically for evaluating image captioning but has been used in video description task as well. WMD makes use of word embeddings which are semantically meaningful vector representations of words and casts distance between two texts as an Earth Mover's Distance (EMD). This metric is less sensitive to words order and synonym changes in the sentence and like CIDEr and METEOR provides high correlation with human judgements. Lastly, SPICE is a more recent metric that correlates better with human judgement of semantic quality than previously reported metrics. It compares the semantic content of two sentences by matching their information in dependency parse trees. These metrics give very different performance measures for the same method and are not perfectly aligned with human judgements.

To the best of our knowledge, there is no survey that comprehensively covers different aspects of video description research such as existing methods, dataset characteristics, evaluation measures, benchmark results and related competitions and video Q&A challenges. We cover this gap and present a comprehensive survey of the literature. We first highlight the important practical application areas of video description in Section 2 and then classify automatic video description methods into two groups and provide a comprehensive overview of the models in each group in Section 3. In Section 4, we elaborate on the available video description datasets used for benchmarking. In Section 5, we present the details of video competitions and challenges. Furthermore, we review the evaluation metrics that are being used for quantitative analysis of the generated descriptions in Section 6. In Section 7, benchmark results achieved through the aforementioned methods are compared and discussed. Section 8 concludes our survey and discusses some insights into the findings of our survey.

2 Applications of Video Description

Video description has many practical applications, including at the very least: human-robot interaction, video indexing, automatic video subtitling, procedure generation for instructional videos, assisting the visually impaired, understanding sign language and video surveillance.

- *Video indexing*: As per Fortunelords' YouTube statistics, 300+ hours of video content is uploaded to YouTube[1] every minute. With the ever growing volume of user generated videos, automatic content based video indexing and retrieval becomes imperative. Poor tagging is the major issue that undermines the utility of videos because they are seldom found using conventional search methods [58]. Automatic video description generation has the potential to improve tagging and thus facilitate the effective and accurate retrieval of online videos.
- *Helping the visually impaired*: Video description can help the visually impaired by generating verbal descriptions of surroundings through speech synthesis, or automatically generating and reading out film descriptions. Currently, these are very costly and time-consuming manual processes.
- *Sign language translation*: Video description can facilitate sign language translation by generating descriptions from processing videos containing sign languages.
- *Service robots*: Video description can generate written procedures for human usage or service robots to take up a job after understanding actions in an instruction or demonstration video, for example, making coffee or changing a flat tyre [59].

In summary, the advancement of video description opens up enormous opportunities. It is envisaged that in the near future, we would be able to interact with robots in the same manner as with humans [35]. If video description is advanced to the stage of being able to comprehend events unfolding in the real world and render them in spoken words, *Service Robots* or *Smartphone Apps* will be able to understand human actions and other events to converse with humans in a much more meaningful and coherent manner. They could answer a user's question as to where they left their wallet or discuss what they should cook for dinner. In industry settings, they could potentially remind a worker of any actions/procedures that are missing from a routine operation.

3 Video Description Methods

3.1 Subject-Verb-Object (SVO) Tuples

The SVO tuples approach is one of the first successful methods in video description. It tackles the video description generation task in two stages. The first stage known as *content identification* focuses on visual recognition and classification of the main objects in the video clip. These typically include the performer, the action and the object of that action. The second stage involves *sentence generation* which maps the objects identified in the first stage to Subject, Verb and Object (and hence the name SVO), and filling in handcrafted templates for grammatically sound sentences. These templates are created using grammar or rule-based systems, which are only effective in very constrained environments, i.e. short clips or videos with limited number of objects and actions.

A variety of approaches have been proposed for detecting objects, humans, actions and events in videos. Below we summarize the recognition techniques used in the Stage I of the SVO tuples based approaches in the literature.

1. https://fortunelords.com/youtube-statistics/ accessed on on 15-04-2018

- *Object Recognition:* Object recognition has been conventionally performed with Model-based shape matching by edge detection or color matching [39], matching HAAR features [60], context based object recognition [61], Scale Invariant Feature Transform (SIFT) [62] discriminatively trained part-based models [63], [64] and Deformable Parts Model (DPM) [65], [66].
- *Human and Activity Detection:* Human detection has been conventionally performed with Histograms of Oriented Gradient (HOG) [67]. For activity detection, Bayesian Networks (BN) [68], Dynamic Bayesian Networks (DBNs) [69], Hidden Markov Models (HMM) [70], state machines [71], and PNF Networks [72] have been used.
- *Integrated Approaches:* Instead of detecting the description-relevant entities separately, Stochastic Attribute Image Grammar (SAIG) [73] and Stochastic Context Free Grammars (SCFG) [74], allows for compositional representation of visual entities and objects present in a video, an image or a scene based on their spatial and functional relations. Using the visual grammar, image content extraction is formulated as a parse graph to find a specific configuration produced by the grammar that best describes the image. In other words, not all entities present in an image are of equal relevance, which is a distinct feature of this method compared to the aforementioned non-integrated approaches.

For sentence generation, a number of methods have been proposed including HALogen representation [75], Head-driven Phrase Structure Grammar (HPSG) [76], planner and surface realizer [77]. The primary common task of these methods is to define templates. A template is a pre-defined language structure with slots for user specified parameters. Each template requires three parts for its proper functioning, namely, lexicons, template rules and grammar. *Lexicon* is a vocabulary that describes high level features extracted from a video stream. *Template rules* are user-defined rules guiding the selection of appropriate lexicons for sentence generation. *Grammar* is core to the study of computational linguistics, which defines the body of linguistic rules that describe the structure of expressions in a language and assures syntactical correctness of the sentence. Known for their expressive power, Grammar enables the generation of a very large set of configurations from a small vocabulary using production rules.

In a template based approach, a sentence is generated by fitting the most important entities to each of the categories required by the template, e.g. subject, verb, object, and place. Entities and actions recognized in the content identification stage are used as lexicons. Correctness of the generated sentence is ensured by Grammar. Figure 3 presents an example list of templates used for sentence generation in a template based approach.

Below is a detailed account of the key literature and the approaches taken in each stage of the SVO.

The pioneer work of Kojima et al. [39] in 2002 focuses primarily on describing videos of a single person performing a single action. To detect humans in a scene, they calculated

Subject + Verb
Woman is walking.
A man is standing.

Subject + Verb + Object
Man is smoking a cigarette.
A man is drinking coffee.

Subject + Verb + Object + Place
A woman is cooking in the kitchen.
A boy is playing on the beach.

Subject + Verb + Complement
Man looks tired.
Woman is old

Fig. 3: An example list of templates for sentence generation. To fill in the template, subject, verb and object are used. Verb is obtained from action/activity detection methods using spatio-temporal features whereas subject and object are obtained from object detection methods using spatial features.

the probability that a pixel belongs to the background or the skin region from the chromaticity values of the pixel and their distributions. Once the human head and hands are detected, they estimate human posture by considering three kinds of geometric information i.e. position of head, direction of head and position of hands. For example, to obtain the head direction, the detected head image is compared against a list of pre-collected head models and a threshold is used to decide on the matching head direction. For object detection, they applied two-way matching, i.e. shape based matching and pixel based color matching to a list of predefined known objects. Actions detected are all related to object handling and the difference image is used to detect actions such as putting an object down or lifting an object. To generate the description in sentences, predefined case frames and verb patterns as proposed by Nishida et al. [78], [79] are used. Case frame is a type of frame expression used for representing the relationship between cases, which are classified into 8 categories. The frequently used ones are *agent*, *object*, and *locus*. For example, "a person walks from the table to the door", is represented as:

```
[PRED:walk, AG:person, GO-LOC:by(door),
SO-LOC:front(table)]
```

where PRED is the predicate for action, AG is the agent or actor, GO-LOC is the goal location and SO-LOC is the source location. A list of semantic primitives are defined about movements, which are organised using body action state transitions. For example, if `moving` is detected and the speed is `fast`, then the activity state is transitioned from `moving` to `running`. They also distinguish durative actions such as walk from instantaneous actions such as stand up. The major drawback of this approach is that it cannot be easily extended to more complex scenarios. Its heavy reliance on the correctness of manually created activity concept hierarchy and state transition model also

prevents it from being used in practical situations.

Lee et al. [80] proposed a framework for semantic annotation of visual events in three steps: image parsing, event inference and language generation. The first component is carried out by an image parsing engine using stochastic attribute image grammar (SAIG) [73]. The output of the image parsing engine is a visual vocabulary i.e. a list of visual elements and objects present in the frame and how they are related. This output is then fed into an event inference engine, which extracts semantic and contextual information of visual events, along with their relationships. They used Video Event Markup Language (VEML) [81] for semantic representation. Finally, in the text generation stage, head-driven phrase structure grammar (HPSG) [76] is used to convert the semantic representation into text description. Compared to Kojima et al. [39], the grammar based approaches can infer and annotate a broader range of scenes and events. Ten sequences of urban traffic and maritime scenes spanned over the duration of 120 minutes, containing more than 400 moving objects are used for evaluation. Some detected events include entering and exiting the scene, moving, turning, stopping, approaching traffic intersection; watercraft approaching maritime markers or land areas and one object following another object. Recall and Precision rates are used to compare the detected events with manually labeled ground truth. Due to poor estimation of the motion direction from low number of perspective views, their method does not perform well on "turning" events.

Khan et al. [48] introduced an approach to generate natural language descriptions for human related contents such as actions (limited to 5 actions only) and emotions in videos. They implemented a suite of conventional image processing techniques, including face detection [82], emotion detection [83], action detection [70], non-human object detection [60] and scene classification [84], to extract the high level entities of interests from video frames. These include humans, objects, actions, gender, position and emotion. In their approach, a human is rendered as *Subject* who is performing some action and objects affected by human actions or activities are rendered as *Object*. A template based approach is adopted to generate natural language sentences based on the detected entities. They evaluated the method on a dataset of 50 snippets where each snippet spanned over 5 to 20 seconds duration. Out of 50, 20 snippets were human close-ups and 30 showed human activities such as stand, walk, sit, run and wave. The primary focus of their research was on activities inolving a human interacting with some objects. Hence, their method does not generate any description until a human is detected in the video. The method also fails at identifying actions with little movement such as smoking and drinking and does not capture interactions among humans.

Hanckmann et al. [85] proposed a method to automatically describe events involving multiple actions (7 on average), performed by one or more individuals. Unlike Khan et al. [48], human-human interactions are taken into account in addition to human-object interactions. Bag-of-features (48 in total) are collected as action detectors [86] for detecting and classifying actions in a video. The description generator subsequently describes the verbs with a rule-based method that relates the actions to the entities in the scene. It finds the appropriate actors among objects or persons and connects them to the appropriate verbs. In contrast to Khan et al. [48] who assume that the subject is always a person, Hanckmann et al. [85] generalizes subjects to people and vehicles. Furthermore, the number of human actions is much richer, compared to the 5 verbs in Khan et al. [48]), they have 48, capturing a diverse range of actions such as `approach`, `arrive`, `bounce`, `carry`, `catch` and etc.

Barbu et al. [29] generated sentence descriptions for short videos of highly constrained domains consisting of 70 object classes, 48 action classes and a vocabulary of 118 words. They rendered the detected object as a noun, the observed action as a verb, the identified object properties as adjectives and the spatial relationships between the objects using prepositions. Their approach consists of three steps. First, object detection [87] is performed on each frame by limiting 12 detections per frame to avoid over detections. Second, object tracking [88], [89] is performed to increase the precision. Third, using dynamic programming the optimal set of detections is chosen that is temporally coherent with the optical flow, yielding a set of object tracks for each video. Hidden Markov Models (HMMs) are then employed as time series classifiers to produce verb labels for actions in each video. After getting the verb, all tracks are merged to generate template based sentences that comply to grammar rules.

Despite the reasonably accurate lingual descriptions generated for videos in constrained environments, the aforementioned methods have trouble scaling to accommodate increased number of objects and actions in open domain and large video corpora. The challenge of enumerating all relevant world concepts requires a customized detector for each entity. Furthermore, the text generated by existing methods of the time has largely been in the form of putting together lists of keywords using grammars and templates without any semantic verification.

To address the issue of lacking semantic verification, Das et. al [47] proposed a hybrid method that shows more relevant content generation over simple keyword annotation methods of videos. They borrowed ideas from the bottom-up and top-down approaches for image description. The hybrid model comprised of low level initial keyword annotation, a middle level concept detection and a high level final lingual description. First, in a bottom up approach, keywords are predicted using low level video features. In this approach they first find a proposal distribution over some training vocabulary using *multimodal latent topic models*. Then by using grammar rules and parts of speech (POS) tagging, most probable subjects, objects and verbs are selected. Second, in a top down approach, a set of concepts is detected and stitched together which subsequently is converted to lingual descriptions using a tripartite graph template. Third, for high level semantic verification, they produced a ranked set of natural language sentences by comparing the predicted caption keywords with the detected concepts. Quantitative evaluation of this hybrid method shows that it was able to generate more relevant content compared to its predecessors [29], [50].

While most of the prior mentioned works are restricted

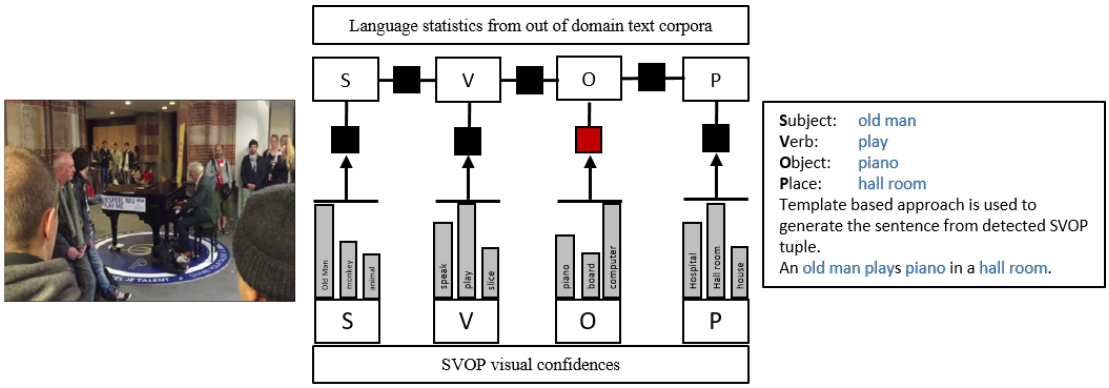


Fig. 4: Example of the Subject-Verb-Object-Place (SVOP) [90] approach where confidences are obtained by integrating probabilities from visual recognition system with statistics from out of domain English text copora to determine the most likely SVOP tuple.

to constrained domains, Krishnamoorthy et al. [91] lead the early works of describing open domain video data. They used selected open domain YouTube videos, however, the subjects and objects were limited to the 20 entities that were available in the classifier training set. Their main contribution is the introduction of text-mining using web-scale text copora to aid the selection of the best SVO tuple to improve sentence coherence.

In addition to focusing on open domain videos and utilising web scaled text corpora, Guadarrama et al. [92] and Thomason et al. [90] also dealt with relatively larger vocabulary. Compared to Krishnamoorthy et al. [91], instead of using only the 20 objects in the PASCAL dataset [93], all videos of the YouTube corpora are used for the detection of 241 objects, 45 subjects, and 218 verbs. To describe short YouTube videos, Guadarrama et al. [92] proposed a novel language driven approach. They introduced "zero-shot" verb recognition for selecting unseen verbs in the training set. For example, given the subject "person", object "car" and the model-predicted verb "move", the most likely verb would be "drive". Thomason et al. [90], utilized similar visual recognition techniques as Guadarrama et al. [92] on YouTube videos for probabilistic estimations of subjects, verbs, and objects. The object and action classifiers were trained on ImageNet [94]. In addition to detecting subjects, verbs and objects, places (12 scenes) where actions are performed, e.g. kitchen or play ground are also identified. To further improve the accuracy of assigning visually detected entities to the right category, probabilities using language statistics obtained from four "out of domain" English text corpora: English Gigaword, British National Corpus (BNC), ukWac and WaCkypedia EN are used to enhance the confidence of word-category alignment for sentence generation. A small "in domain" corpus comprising sentences from human-generated descriptions of the video description dataset is also constructed and incorporated in the sentence generation stage. Co-occurring bi-gram (SV, VO, and OP) statistics from the candidate SVOP tuples are calculated using both the "out of domain" and the "in domain" corpus, which are used in a Factor Graph Model (FGM) to predict the most probable SVO and place combination. Finally, the detected SVOP tuple is used to generate an English sentence through a template based approach. An illustration of the approach is shown in Figure 4.

In order to overcome the tedious efforts of rule based engineering methods when the problem scales, some later methods [35] train statistical models for lexical entries. Rohrbach et. al [35] introduced to learn the conversion from visual content to natural language descriptions from a parallel corpus of videos and textual descriptions. Their method follows a two-step approach, first learning an intermediate representation of semantic labels from the video, and then translating it to natural language adopting techniques from Statistical Machine Translation (SMT) [95]. In the first step, they generated the rich semantic representation of the visual content using the maximum a posterior estimate (MAP) of a CRF taking in video features as unaries. They used dense trajectories [96] and SIFT features [62] as well as temporal context reasoning modeled in a CRF. This representation is then translated to a natural language sentence using phrase-based statistical machine translation. As a second step, they generated natural language description as a machine translation problem using the semantic representation as source language and the generated sentences as target language.

Naïve SVO tuple rule-based engineering approaches and statistical methods soon become inadequate in dealing with large open domain videos, such as YouTubeClips [97], TACoS-MultiLevel [44], MPII-MD [98], M-VAD [99], MSR-VTT [42] and ActivityNet Captions [30]. These datasets contain thousands of lexical entries and dozens of hours of videos. The meaning of "large" is three folds here. Firstly, open domain videos contain unforeseeable diverse set of subjects, objects, activities and places. Secondly, due to the sophisticated nature of human languages, such datasets are often annotated with multiple viable meaningful descriptions. Thirdly, it also means that the videos to be described are often longer, potentially stretching through many hours. Descriptions in multiple sentences or even paragraphs become more desirable.

In summary, for object and activity recognition stage, the research moved from earlier threshold-based detection [39] to manual feature engineering and traditional classifiers [47], [90], [91], [92]. For the sentence generation stage, we observe an uptake of machine learning methods in recent years for addressing the issue of large vocabulary, evidenced by recent methods using trained models for

lexical entries as, either in a fully [91], [92], [100], [101] or weakly [35], [44], [102], [103] supervised fashion. However, the separation of the two stages makes this camp of methods incapable of capturing the interplay of visual features and linguistic patterns, let alone learning a transferable state space between visual artifacts and linguistic representations. In the next section, we review the deep learning methods in addressing the scalability, language complexity and domain transferability issues facing large open domain video descriptions.

3.2 Deep Learning Models

The whirlwind success of deep learning in almost all subfields of computer vision, has also revolutionized how we approach video description. In particular, Convolutional Neural Networks (CNNs) [104] are now the state of the art for visual data modeling and excel at tasks such as object recognition [104], [105], [106]. Long Short-Term Memory (LSTMs) [107] and the more general deep Recurrent Neural Networks (RNNs), on the other hand, are now dominating the area of sequence modeling, setting new benchmarks in machine translation [108], [109], speech recognition [110] and the closely related task of image captioning [21], [36]. While conventional methods struggle to cope with large-scale, more complex and diverse datasets for video description, researchers have combined these deep nets in various configurations with promising performances.

As shown in Figure 5, the deep learning approaches to video description can also be divided into two sequential stages, namely, visual recognition and sequence generation. However, in contrast to the SVO Tuple Methods in Section 3.1, where lexical word tokens are generated as a result of the first stage through visual content recognition, visual features represented by fixed or dynamic real-valued vectors are produced instead. For an end-to-end system, and borrowing terms from machine translation, this is often referred to as the *video encoding stage*. CNN, RNN or Long Short-Term Memory (LSTM) are used in this encoding stage to learn these visual features, which are then fed through the second stage for text generation, also known as the *decoding stage*. For decoding, different flavours of RNNs are used, such as deep RNN, Bi-directional RNN, LSTM or Gated Recurrent Units (GRU). The resulting description can be a single sentence or multiple sentences. Figure 6 illustrates a typical end-to-end video description system, in this case CNNs are used for object and action encoding, and a two-layered LSTM for sentence generation. Hereby, we group the literature based on the different combinations for the encoding and decoding stage, namely:

- CNN - RNN Video Description, where convolution architectures are used for visual encoding and recurrent structures are used for decoding;
- RNN - RNN Video Description, where recurrent networks are used for both stages; and the relatively new research area,
- Deep reinforcement networks for video description.

3.2.1 CNN-RNN Video Description

Given its success in computer vision, CNN is still by far the most popular network structure used for visual encoding. The encoding process can be broadly categorized into fixed-size and variable-size video encoding.

Donahue et al. [21] presented the first application of deep network models to the video description task. They proposed three architectures for video description with an assumption to have predictions of objects, subjects, and verbs present in the video from a CRF, based on the complete video input. This allows the architecture to observe the video as a whole at each time step instead of incrementally frame by frame. The first architecture, LSTM encoder-decoder with CRF max, is motivated by the statistical machine translation (SMT) based video description approach by Rohrbach et al. [35] mentioned earlier in Section 3.1. Recognizing the state-of-the-art performance of LSTM in machine translation tasks, the SMT module in [35] is replaced with a two-layer LSTM for encoding and decoding. Similar to [108], the first LSTM layer encodes the one-hot vector of the input sentence allowing for variable-length inputs. The final hidden representation from the first encoder stage is then fed into the decoder stage which decodes the input into a sentence, one word at each time step. Another variant of the architecture, LSTM decoder with CRF max, incorporates max predictions. This architecture encodes the semantic representation as a single fixed length vector. The entire visual input representation is provided at each time step to the LSTM, similar to image description. An advantage of LSTM is that it can naturally incorporate probability vectors during training and test time which allows the LSTM to learn uncertainties in visual generation. This virtue of LSTM is exploited in the third variant of the architecture, LSTM decoder with CRF probabilities. Instead of using max predication like in second variant (LSTM decoder with CRF max), this architecture incorporates probability distributions. Although the LSTM outperformed the SMT based approach of [35], it is not yet an end-to-end trainable deep network model.

In contrast to the work by Donahue et al. [21], where an intermediate role representation is adopted, Venugopalan et al. [111] presented the first end-to-end deep model for video-to-text generation. Their model is able to simultaneously learn a latent "meaning" state and fluent grammatical model of the associated language. Moreover, Donahue et al. [21] showed results on the narrow domain cooking videos with a small set of pre-defined objects and actors, whereas Venugopalan et al. [111] reported results on YouTube Clips [112]. An LSTM is used to model sequence dynamics. To avoid supervised intermediate representations, they connected the LSTM directly to a deep CNN to process input video frames. First, to generate a fixed-length visual input that effectively summarizes a short video, a CNN [113], a minor variant of AlexNet [104] was adopted. The CNN was pre-trained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [94]. It provides a robust and efficient way without manual feature selection for initialization for recognizing objects in the videos. They sampled one in every ten frames in the video, extracted *fc7* layer feature vectors and performed a mean pooling over all the extracted frames to generate a single 4,096 dimensional vector to represent the whole video. These feature vectors are then fed into a two-layered LSTM [114]. The feature vectors from CNN form the input to

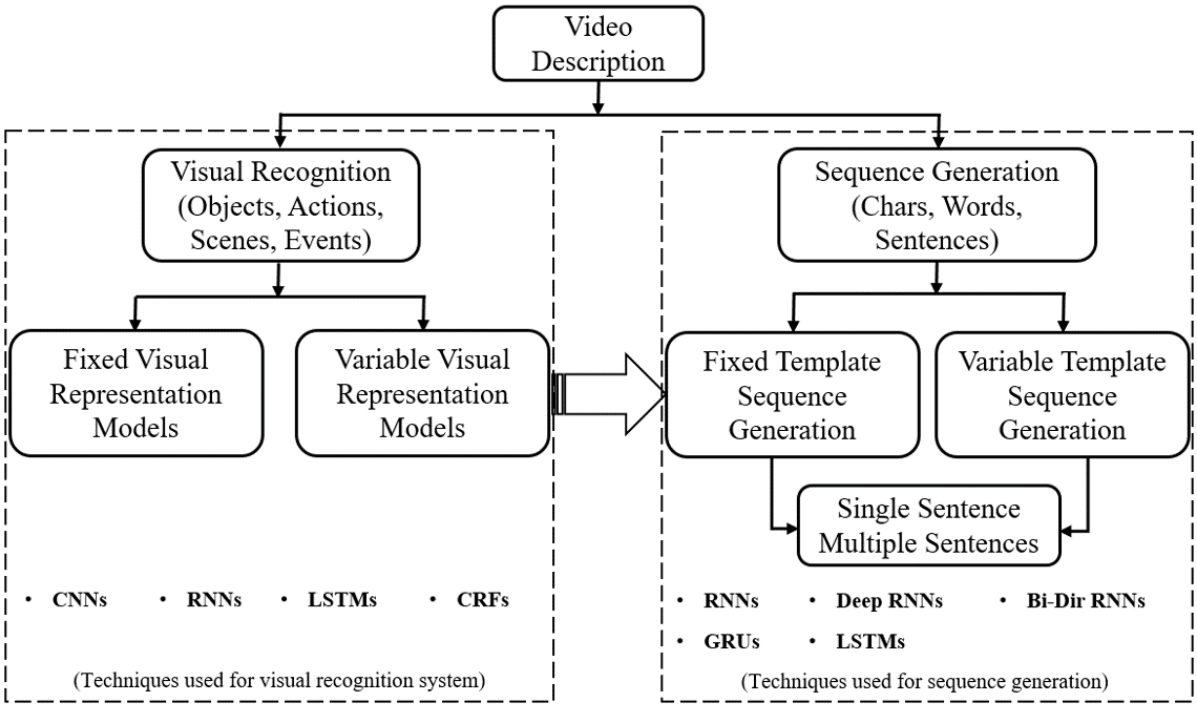


Fig. 5: Breakdown of deep learning based video description techniques used in the literature.

the first LSTM layer. A second LSTM layer is stacked on top of first LSTM layer, where the hidden state of the first LSTM layer becomes the input to the second layer LSTM unit for caption generation. In essence, the transforming of multiple frame-based feature vectors into a single aggregated video-based vector, reduces the video description problem into an image captioning one. This end-to-end model was better than the previous video description works at the time and effectively models the sequence generation task without requiring the use of fixed template based sentences. However, as a result of simple averaging, valuable temporal information of the video, such as the order of appearances of any two objects, are lost. Therefore, this approach is only suitable of generating captions for short clips with a single major action in the clip.

Open domain videos are rich in complex interactions among actors and objects. Attempting to represent such videos using a temporally collapsed single feature vector is therefore prone to produce clutter. Consequently, the descriptions produced are bound to be inadequate because valuable temporal ordering information of events are not captured in the representation. To incorporate temporal structure into the description generation, Li et al. [115], proposed a novel spatio-temporal 3D CNN to model input video clips. The 3D CNN is based on GoogLeNet [106] and pre-trained on an activity recognition dataset, to capture local fine-grained motion information from consecutive frames. The local motion information is then subesequently summarized and preserved through higher-level representations, which is achieved by dividing the input video clip into 3D spatio-temporal cuboids. Each cuboid is then represented by concatenating the histograms of oriented gradients, oriented flow and motion boundary (HoG, HoF,

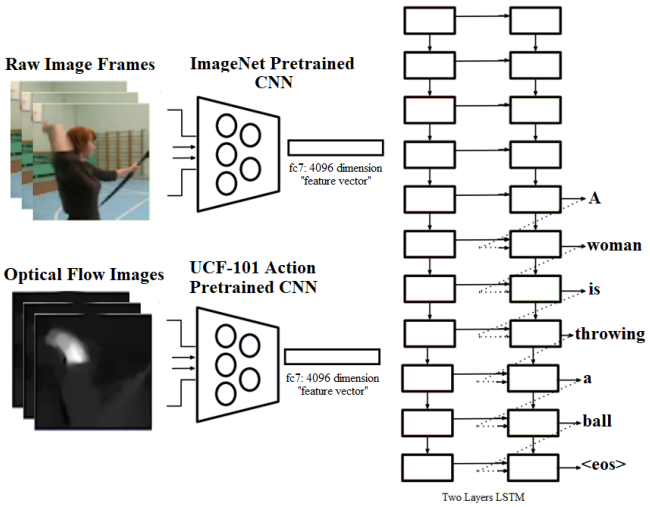


Fig. 6: S2VT Model for video description. The visual input to the model comprises CNN outputs of RGB raw frames and/or Optical Flow images. Figure adapted from [40]

MbH) [116], [117]. These transformations not only help capture local motion features but also reduce the computation of the subsequent 3D CNN. For global temporal structure, a temporal attention mechanism is proposed and adapted from soft attention [118] that allows the decoder to weight each temporal feature vector in a text-generating RNN. Using 3-D CNN and attention mechanism in RNN they were able to improve results.

Unlike the *fixed video representation models*, as discussed above, *variable visual representation models* are able to directly map the variable length inputs (video frames) to variable

length outputs (words or sentences) and are successful in modeling various complex temporal dynamics. Venugopalan et al. [40] proposed an architecture to address the variable representation problem for both the video encoding and the sentence decoding stage. They used a stacked two-layered LSTM. The first LSTM for reading the sequence of video frames and the second LSTM conditioned on the last hidden state of the first to generate the corresponding caption. The LSTM parameters are shared in both stages to improve learning. That was the first time when sequence to sequence approach, that was successfully applied to machine translation [108] for the video captioning problem. Later works have followed a similar kind of approach, either by incorporating attention mechanisms [115] in the sentence decoder, by building a common visual-semantic-embedding [43] or by adding external knowledge with language models [119] or visual classifiers [120].

Although deep learning based methods have achieved much better results compared to previously used classifier based approaches, most of the methods focused on producing single sentence from a short video clip containing only one main event. In real-world applications on the other hand, most videos contain more than a single event. Description of such multi-events and semantically rich videos by only one sentence end up in overally simplified, and hence, uninformative sentences. For example, instead of saying "someone sliced the potatoes with knife, chopped the onions into pieces and put the onions and potatoes into the pot", a single sentence generation method would probably say "someone is cooking". Yu et al. [46] proposed a hierarchical recurrent neural network (h-RNN) that exploits both the temporal and spatial attention mechanisms. They focused on the sentence decoder and proposed a hierarchical model containing a sentence and a paragraph generator. First, short sentences are produced by a Gated Recurrent Unit (GRU) layer [121] that takes video features as input and generates a single short sentence. The other recurrent layer is responsible for generating paragraphs by combining sentence vectors from the sentence generator. The paragraph generator thus captures the inter-sentence dependencies and generates a sequence of relevant and consecutive sentences.

Recently, Krishna et al. [30] introduced the task of dense-captioning events in a video and employed action detection techniques to predict the temporal intervals. They proposed a model to identify multiple events in a single pass of a video, attempting to describe the detected events simultaneously. This is the first work of its kind detecting and describing multiple and overlapping events in a video. However the model has not shown significant improvement in captioning benchmark, but set the baseline for challenging videos.

3.2.2 RNN - RNN Video Description

Srivastava et al. [122] performed visual recognition by using an LSTM to encode the video frames and then feed the feature vector into another LSTM for decoding. They also introduced some variants of their models and predicted the future sequences from the previous frames. The authors adopted the language translation model [108] for visual recognition but could not achieve significant improvement in classification accuracy.

Yu et al. [46] also proposed a similar approach and used two RNN structures for video description task. Their configuration builds a hierarchical decoder using multiple Gated Recurrent Units (GRU) for sentence generation. The output of this decoder is then fed to a paragraph generator which models the time dependencies between the sentences while focusing on linguistic aspects. The authors improved the state-of-the-art results for video description, but not significant on object classification accuracy. Their method is inefficient for videos involving fine-grained activities and small interactive objects.

3.2.3 Deep Reinforcement Learning Models

Deep Reinforcement Learning (DRL), a sub-category of Reinforcement Learning (RL), where artificial intelligent agents learn from the environment by trial-and-error, adjust learning policies purely from environmental rewards or punishments, has out-performed human counterparts in many real-word games. DRL approaches are pioneered by Google Deep Mind [123], [124] since 2013. Due to the absence of a straight forward cost function, learning mechanisms in this approach are considerably harder to devise as compared to traditional supervised techniques. Two distinct challenges are evident in reinforcement learning when compared with conventional supervised approaches: (1) The model does not have full access to the function being optimized. It has to query the function through interaction. (2) The interaction with the environment is state based where the present input depends on previous actions. The choice of reinforcement learning alogorithms then depends on the scope of the problem at hand. For example, variants of Hierarchical Reinforcement Learning (HRL) framework have been applied to Atari games [125], [126]. Similarly, different variants of DRL have been used to address the problems of image captioning [127] and video description [128], [129], [130], [131], [132].

Xwang et al. [129] proposed a fully-differentiable deep neural network using reinforcement learning for the video description task. Their method follows a general encoder-decoder framework. The encoding stage captures the video frame features using ResNet-152 [133] (a pre-trained CNN model on ImageNet). These frame level features are then passed through a low-level Bi-LSTM [134] encoder and a high-level LSTM [107] encoder successively to obtain encoder stage output. In the decoding stage, HRL agent plays the role of a decoder and outputs a natural language descriptions word by word. The HRL agent comprises of three components, a low level worker that selects primitive actions at every time step by following the goals set by Manager, a high-level manager that sets goals at a lower temporal resolution, and an internal critic that determines whether a goal is accomplished or not. The manager sets the goal for the worker to generate a semantic segment, and the worker generates the corresponding words in order to fulfill the job. The internal critic determines if the worker has accomplished the goal and informs the manager once the goal is accomplished to help manager update the goals. The whole process repeats until the end of sentence token is reached. This method is demonstrated to be able to capture

more details of the video content and generate more fine-grained descriptions. However, this method has shown very little improvement over existing baseline methods.

Recently, Chen et al. [132] proposed a RL based model for the selection of *key informative frames* to represent the complete video, in an attempt to minimize noise and unnecessary computations. The reward for each frame selection is calculated based on maximizing visual diversity and minimizing the textual discrepancy. A complete video can then be represented by a compact subset of 6-8 frames on average. Evaluated against several popular benchmarks, it was demonstrated that video captions can be produced without performance degradation at a significantly reduced computational cost. The method did not use motion features for encoding, a design trade-off between speed and accuracy.

4 DATASETS

The countless applications and availability of labelled datasets for video description have been the main driving forces behind the fast advancement of this research area. In this survey, we describe and summarize the characteristics of these datasets and give an overview in Table 1.

4.1 Cooking

4.1.1 MP-II Cooking

Max Plank Institute for Informatics (MP-II) Cooking dataset [135] comprises 65 fine grained cooking activities performed by 12 participants preparing 14 dishes such as *fruit salad* and *cake* etc. The data are recorded in the same kitchen with camera attached to the ceiling which records a person working at the kitchen counter from the front. The 65 cooking activities include "wash hands", "put in bowl", "cut apart", "take out from drawer" etc. When the person is not in the scene for 30 frames (one second) or is performing an unusual activity which is not annotated, a "background activity" is generated. These fine grained activities, for example "cut slices", "pour", or "spice" are distinguished by body motions with low inter-class and high intra-class variability. In total, the dataset comprises 44 videos (888,775 frames) with an average length per clip of approximately 600 seconds. The dataset spans a total of 8 hours play length for all videos, and a total of 5,609 annotations.

4.1.2 YouCook

The YouCook dataset [47] consists of 88 YouTube cooking videos of different people cooking various recipes. The background (kitchen/scene) is different in most of the videos and includes camera movements. This dataset represents a more challenging visual problem than the MP-II Cooking [135] dataset that is recorded with a fixed camera view point in the same kitchen with the same background. The dataset is uniformly split into six different cooking styles, such as *baking* and *grilling*. For machine learning, the training set contains 49 videos while the rest 39 videos are used for testing. The training data also comes with frame-by-frame object and action annotations. The object categories for the dataset include "utensils", "bowls" and "foods" etc. The Amazon Mechanical Turk (AMT) was employed for human generated multiple natural language descriptions of each video. Each AMT worker provided at least three sentences per video as a description, and on average 8 descriptions were collected per video. See Figure 7(b) for example clips and descriptions.

4.1.3 TACoS

Textually Annotated Cooking Scenes (TACoS) is a subset of MP-II Composites [139]. TACoS was further processed to provide coherent textual descriptions for high quality videos. Note that MP-II Composites contain more videos but less activities than the MP-II Cooking [135]. It contains 212 high resolution videos with 41 cooking activities. Videos in the MP-II Composites dataset span over different lengths ranging from 1-23 minutes with an average length of 4.5 minutes. The TACoS dataset was constructed by filtering through MP-II Composites while restricting to only those activities that involve manipulation of cooking ingredients and have at least 4 videos for the same activity. As a result, TACoS contains 26 fine grained cooking activities in 127 videos. AMT workers were employed for generating high-quality alignment of sentences and video segments such as "preparing carrots", "cutting a cucumber" or "separating eggs" etc. For each video, 20 different textual descriptions were collected. It contains 17,334 action descriptions realized in 11,796 different sentences. A total of 146,771 words are used, out of which 75,210 are content words i.e. nouns, verbs and adjectives. The vocabulary contains 28,292 verb tokens. The dataset also provides the alignment of sentences describing activities by obtaining approximate timestamps for start and end of each activity. Figure 7(d) shows some example clips and descriptions.

4.1.4 TACoS-MultiLevel

TACoS Multilevel [44] corpus annotations were also collected via AMT workers on the TACoS corpus [136]. For each video in the TACoS corpus, three levels of descriptions were collected that include: (1) detailed description of video with no more than 15 sentences per video; (2) a short description that comprises 3-5 sentences per video; and finally (3) a single sentence description of the video. This corpus is annotated with the tuples like activity, object, tool, source and target with a person always being the subject. See Figure 7(e) for example clips and descriptions.

4.2 Movies

4.2.1 MPII-MD

MPII-Movie Description Corpus [98] contains transcribed audio descriptions extracted from 94 Hollywood movies. These movies are subdivided into 68,337 clips with an average length of 3.9 seconds paired with 68,375 sentences amounting to almost one sentence per clip. Each clip is accompanied with a single sentence that is sourced from the script of the movie and the audio description data. The Audio Descriptions (ADs) were collected first by retrieving the audio streams from the movie using online services MakeMkV [2] and Subtitle Edit [3]. The audio segments

2. https://www.makemkv.com/
3. http://www.nikse.dk/SubtitleEdit/

TABLE 1: Standard datasets for benchmarking video description methods.

Dataset	Domain	No of classes	No of Videos	Avg Length	No of Clips	No of Sentences	No of Words	Vocab	Length (Hrs)
MSVD [112]	Open (YTube)	218	1970	10 sec	1,970	70,028	607,339	13,010	5.30
MPII Cooking [135]	Cooking	65	44	600 sec	-	5,609	-	-	8
YouCook [47]	Cooking (YTube)	6	88	-	Nil	2,688	42,457	2,711	2.30
TACoS [136]	Cooking	26	127	360 sec	7,206	18,227	146,771	28,292	15.9
TACos-MLevel [44]	Cooking	1	185	360 sec	14,105	52,593	2,000	-	27.1
MPII-MD [98]	Movie	-	94	3.9 sec	68,337	68,375	653,467	24,549	73.60
M-VAD [99]	Movie	-	92	6.2 sec	48,986	55,904	519,933	17,609	84.60
MSR-VTT [42]	Open	20	7,180	20 sec	10,000	200,000	1,856,523	29,316	41.20
Charades [137]	Human	157	9,848	30 sec	-	27,847	-	-	82.01
VTW [138]	Open (YTube)	-	18,100	90 sec	-	44,613	-	-	213.20
ActyNet Cap [30]	Open	-	20,000	180 sec	-	100,000	1,348,000	-	849.00

were then transcribed using crowd sourced transcription service [140]. The transcribed texts were also aligned to match the time-stamps for each spoken sentence. In order to remove any potential misalignment between the time of speech and the corresponding visual content, they manually aligned each sentence to the movie clip. During the manual alignment of the scripts irrelevant sentences describing the content not present in the video were filtered out. The audio descriptions track is an additional audio track that is added to the movies to describe visual content for the visually impaired. The movie snippets were manually aligned to the sentences with minor misalignments present among them. The total time span of the dataset videos is almost 73.6 hours and the vocabulary size is 653,467. Example clips and descriptions are shown in Figure 7(f).

4.2.2 M-VAD

Montreal Video Annotation Dataset (M-VAD) [99] is based on the Descriptive Video Service (DVS) and contains 48,986 video clips from 92 different movies. Each clip has an average length of 6.2 seconds and the total time for the complete dataset is 84.6 hours. The total number of sentences is 55,904, where some clips are paired with more than one sentence. The vocabulary of the dataset spans about 17,609 words (Nouns-9,512: Verbs-2,571: Adjectives-3,560: Adverbs-857). The dataset split consists of 38,949, 4,888 and 5,149 video clips for training, validation and testing respectively. See Figure 7(g) for example clips and descriptions.

4.3 General

4.3.1 MSVD

Microsoft Video Description (MSVD) dataset [112] is a collection of 1,970 YouTube snippets with human annotated sentences. This dataset was also annotated by AMT workers. The audio is muted in all clips to avoid bias from lexical choices in the descriptions. Furthermore, videos containing subtitles or overlaid text were removed during the quality control process of the dataset formulation. Finally, manual filtering was carried out over the submitted videos to ensure that each video met the prescribed criteria and was free of inappropriate and ambiguous content. The duration of each video in this dataset is typically between 10 to 25 seconds, mainly depicting a single activity. The dataset comprises multilingual (such as Chinese, English, German etc) human generated descriptions. On average, there are 41 single sentence descriptions per clip. This dataset has been frequently used by the research community as detailed in the Results Section 7. Almost all research groups have split this dataset into training, validation and testing partitions of 1200, 100 and 670 videos respectively. Figure 7(a) shows example clips and descriptions from MSVD dataset.

4.3.2 MSR-VTT

MSR-Video to Text (MSR-VTT) [42] is a dataset for general video captioning derived from a wide variety of videos and comprises 7,180 videos of 20 general categories and 10,000 clips. An example is shown in Figure 7(c). Among the 10,000 clips, 6,513 are for training, 497 for validation and the remaining 2,990 are for testing. Each video contains 20 human annotated reference captions created by AMT workers. The dataset forms one of the largest collection of clip-sentence pairs where each video is annotated with multiple sentences. In addition to video content, this dataset also contains audio information that can potentially be used for multimodal research.

4.3.3 Charades

This dataset [137] contains 9,848 videos of daily indoor household activities. These videos are recorded by 267 AMT workers from three different continents. They were given scripts with actions and objects and were required to follow the scripts to perform actions with the specified objects. The objects and actions used in the scripts are from a fixed vocabulary. The dataset contains 66,500 temporal annotations for 157 action classes, 41,104 labels for 46 object classes and 27,847 textual descriptions of the videos. The dataset contains videos of daily activities that are on average 30 seconds long. The dataset is split into training and test subset comprising 7,985 and 1,863 videos respectively. Videos are collected in 15 different types of indoor scenes, involving 46 object classes and comprise a vocabulary of 30 verbs leading to 157 action classes.

C1: The monkey pushed the other monkey.
C2: A gorilla is pushing another gorilla.
C3: The gorilla snuck up behind another one and pushed him down.
C4: Two apes are playing.

(a) MSVD

C1The woman has all the ingredients ready for making muffins . . .
C2: In this video, a woman pours ingredients into a large metal bowl . . .
C3: A well organized kitchen with a microwave and cooking range in the background . . .

(b) YouCook

C1: A woman giving speech on news channel
C2: Hillary Clinton gives a speech.
C3: Hillary Clinton is making a speech at the conference of mayors.
C4: A woman is giving a speech on stage.
C5: A lady speak some news on TV.

(c) MSR-VTT

C1: The person rinses the carrot.
C2: The person cuts off the ends of the carrot.
C3: He starts chopping the carrot in small pieces.
C4: He finishes chopping the carrot in small pieces.

(d) TACoS

C1: The person entered the kitchen
C2: The person took out a drawer.
C3: The person took an egg from the refrigerator.
C4: The person put the egg in the bowl.

(e) TACoS MLevel

C1: A police officer takes a piece of paper from the typewriter.
C2: An officer blows his whistle several times.
C3: There is pandemonium as native guards begin to round up people.

(f) MPII-MD

C1: He plants a tender kiss on her shoulder.

(g) M-VAD

C1: A small group of men are seen running around a basketball . . .
C2: One player moves all around the net holding the ball . . .
C3: He bounces the ball around a bit and more shots. . .

(h) ActivityNet Captions

Fig. 7: Example video frames (3 non-consecutive frames per clip) and captions from the various benchmark video description datasets.

4.3.4 VTW

Video Titles in the Wild (VTW) [138] consists of 18,100 video clips with an average of 1.5 minutes duration per clip. Each video is associated with a single sentence. The vocabulary is very diverse and each word appears only in two sentences on average across the whole dataset. Besides the single sentence per video, the dataset also provides accompanying descriptions (known as augmented sentences) that describe non visual information of the video. The dataset is proposed for video title generation as opposed to video content description but can also be used for language-level understanding tasks including video question answering.

4.3.5 ActivityNet Captions

ActivityNet Captions dataset [30] contains 20,000 videos that correspond to 849 hours with 100k descriptions. Each sentence has an average length of 13.48 words, each sentence describes 36 seconds of video and 31% of their respective videos. The descriptions for each video are detailed and cover 94.6% of the entire video content i.e. each annotation of corresponding video covers almost all major actions within the video. Furthermore, the dataset incorporates 10% overlap in the temporal descriptions which makes the dataset very interesting and challenging for addressing the scenario where multiple events occur at the same time. An example of this dataset is given in Figure 7(h).

5 Video Description Competitions

Another major driving force of the fast paced development in video description research comes from the many competitions and challenges organised by companies and conferences in recent years. Some of the major competitions are listed below.

5.1 LSMDC

The Large Scale Movie Description Challenge (LSMDC) [141] started in 2015 in conjunction with ICCV 2015 and was also presented as an ECCV 2016 workshop. The challenge comprises of public and blind test sets and an evaluation server for automatic evaluation [142] of the results. Primarily it consists of three tasks that includes *Movie Description*, *Movie Annotation and Retrieval* and *Movie Fill-in-the-Blank*. Since 2017, the *MovieQA* challenge has also been included in LSMDC in addition to the previous three tasks.

The dataset for this challenge was first introduced in ICCV 2015 workshop [141]. The LSMDC dataset is a combination of two benchmark datasets MPII-MD [98] and M-VAD [99] which were initially collected independently as mentioned in our dataset section (Section 4.2). The two datasets were merged for this challenge, with overlaps removed to avoid repetition of the same movie in the training and test set. Further, script-based movie alignments from the validation and test sets of MPII-MD have also been removed. The dataset was then augmented by clips only (without aligned annotations) from 20 additional movies to form the blind test of the challenge. These clips are only used for evaluation.

The final LSMDC dataset has 118,081 video clips extracted from 202 unique movies. It has on average approximately one sentence per clip. Names of characters in the reference captions are replaced with the token word "SOMEONE". The dataset is further split into a training set of 91,908 clips, a validations set of 6,542 clips, a public test set of 10,053 clips and a blind test set of 9,578 clips. The average clip length is approximately 4.8 seconds. The training set captions consists of 22,829 unique words. A summary of the LSMDC dataset can be found in Table 2.

A survey of the benchmark results on video description (Section-7) shows that LSMDC has emerged as the most challenging dataset, evident by the poor performances of several models. As mentioned in the dataset section (Section 4.2), the natural language descriptions of movie clips are typically sourced from movie scripts and audio descriptions, so misalignments between captions and videos often occur when text refer to objects that appeared just before or after the cutting point of a clip. Misalignment is certainly a key contributing factor causing poor performance on this dataset. Another factor could be the indiscriminate replacement of various names with "SOMEONE".

Submission protocol of the challenge is similar to the MS COCO Image Captioning Challenge [143] and uses the same automatic evaluation protocol. The winner is selected based on human evaluation. The latest results of automatic evaluation on LSMDC are publicly available [144].

5.2 MSR-VTT

In 2016, to further motivate and challenge the academic and industrial research community, Microsoft started the Microsoft Research - Video to Text (MSR-VTT) [145] competition aiming at bridging together video and language research. The dataset used for this competition is MSR-VTT [146], see our dataset section (Section 4.3) for more details. The participants of the competition are asked to develop a video to language model using MSR-VTT dataset. External datasets, either public or private can be used to help for better objects, actions, scenes and events detection, so long as the external data used are explicitly cited and explained in the submission file.

Unlike LSMDC, MSR-VTT challenge focuses only on video to language task. In this challenge, given an input video clip, the task is to automatically generate at least one complete sentence that should encapsulate the most informative dynamics of the video. Accuracy is benchmarked against human generated captions during the evaluation stage. The evaluation is based on a automatically computed score using multiple common metrics such as BLEU@4, METEOR, ROUGE-L, and CIDEr-D. Details of these metrics are given in Section- 6. Like LSMDC, human evaluations are also used to rank the generated sentences.

5.3 TRECVID

Text Retrieval Conference (TREC) is a series of workshops focusing on various subareas of Information Retrieval (IR) research. In particular, the TREC Video Retrieval Evaluation (TRECVID) [147] workshops, started in 2001, are dedicated to research on content-based exploitation of digital videos. The primary areas of interests include but are not limited to

TABLE 2: LSMDC Dataset Statistics.

Dataset Split	No of Movies	No of Clips	Words	Sentences	Avg Length (sec)	Total Length (hrs)
LSMDC Training	153	91,908	913,841	91,941	4.9	124.90
LSMDC Validation	12	6,542	63,789	6,542	5.2	9.50
LSMDC Public Test	17	10,053	87,147	10,053	4.2	11.60
LSMDC Blind Test	20	9,578	83,766	9,578	4.5	12.00
LSMDC (Total)	202	118,081	1,148,543	118,081	4.8	158.00

(a) (b)

Fig. 8: Example video frames from TRECVID-VTT dataset. (a) Frames from the Easy-Video category and (b) frames from the Hard-Video category.

semantic indexing, video summarization, video copy detection, multimedia event detection and ad-hoc video search. Since TREC-2016, a new Video to Text Description (VTT) [148] task using natural language has also been included in the challenge tasks.

TRECVID-2017 VTT task used a dataset of over 50K automatically collected Twitter Vine videos where each clip is approximately 6 seconds long. In this task, a subset of 1,880 videos were randomly selected and annotated manually. The dataset is further divided into four groups based on the number of descriptions (2 to 5) per videos. These groups are referred to as G2, G3, G4 and G5. Each group has multiple sets of descriptions containing all the descriptions for the entire videos subset. Furthermore, the videos are also tagged as easy or hard according to how difficult it is to describe a video. A set of example frames from the VTT are show in Figure 8.

In addition to the metrics BLEU, METEOR and CIDEr (see the evaluation metrics Section- 6), TRECVID also introduced the Semantic Text Similarity (STS) [149] metric. This metric measures the semantic similarity between the candidate and the ground truth description. Human evaluations are also employed to measure the quality of the automatically generated captions. The recently proposed Direct Assessment (DA) [150] method, has shown to produce highly reliable human evaluation results for machine translation. It has now become the official ranking method in machine translation benchmark evaluations [151]. As per DA, for video description evaluation, human assessors are given a video and a sentence pair. After watching the video, assessors are asked to rate how well the caption describes the events in the video on a scale of $0-100$ [152].

5.4 ActivityNet Challenge

ActivityNet Dense-Captioning Events in Videos [153] was first introduced in 2017 as a task of the ActivityNet Large Scale Activity Recognition Challenge [154], [155], running as a CVPR Workshop since 2016. This task studies the detection and description of multiple events in a video. In the ActivityNet Captions Dataset, each video clip is associated with a set of temporally annotated sentence descriptions, where each description covers a unique portion of the clip. Together, multiple events in that clip can be covered and narrated using the set of sentences. The events may be of variable durations (long or short) or even overlap. Details of this dataset have been described in Section 4.3.5 and Table 1.

Server based evaluations [156] are performed for this challenge. The precision of captions generated are measured using traditional evaluation metrics: BLEU, METEOR and CIDEr. The latest results for the challenge are also publicly available and can be found online [157].

6 Evaluation Metrics

6.1 Automatic Sentence Generation Evaluation

Evaluation of video descriptions is a highly challenging task as there is no specific ground truth or "right answer" that may be taken as a reference for benchmarking accuracy. A video can be correctly described in a wide variety of sentences that may differ not only syntactically but in terms of semantic content as well. Take an sample from MSVD dataset as shown in Figure 9 for instance, several ground truth captions are available for the same video clip. Note that each caption describes the clip in an equally valid but different way with varied attentions and levels of details in the clip, ranging from "jet", "commercial airplane" to "South African jet" and from "flying", "soaring" to "banking" and lastly from "air", "blue sky" to "clear sky".

For automatic evaluation, when comparing the generated sentences with ground truth descriptions, three evaluation metrics are borrowed from machine translation, namely, Bilingual Evaluation Understudy (BLEU) [52], Recall Oriented Understudy of Gisting Evaluation (ROUGE) [53] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [54]. Consensus based Image Description Evaluation (CIDEr) [55] and Semantic Propositional Image

C1: A jet is flying.
C2: A commercial plane flying.
C3: A South African jet banked itself in the air.
C4: A South African Airways plane is flying in a blue sky.
C5: An airplane is flying in the clear sky.
C6: The plane is soaring through the air.

Fig. 9: An example from MSVD dataset with the associated ground truth captions. Note how the same video clip has been described very differently. Each caption describes the activity wholly or partially in a different way.

Captioning Evaluation (SPICE) [56] are two other recently introduced metrics specifically designed for image captioning tasks, that are also being used for automatic evaluation of video descriptions. A summary of the metrics described in this section is given in Table 3.

In addition to these automatic evaluation metrics, human evaluations are also employed to determine the performance an automated video description algorithm.

6.1.1 Bilingual Evaluation Understudy (BLEU, 2002)

BLEU is a popular algorithm used to quantify the quality of machine generated text. The quality measures the correspondence between a machine's output and that of a human. BLEU scores are computed based on the overlap between predicted *uni–gram*s (single word) or higher order *n–gram* (a contiguous sequence of n words) and a set of one or more candidate reference sentences. According to BLEU, a high-scoring description should match the reference sentence in length, in word choice and in word order. The BLEU evaluation will score 1 for an exact match. Note that the more reference sentences in the ground truth per video there are, the higher the BLEU score. It is primarily designed to evaluate text at a corpus level and, therefore, its use as an evaluation metric over individual sentences may not be fair. BLEU is calculated as,

$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n,$$

where c is the length of the candidate translation and r is the effective reference corpus length, w_n are positive weights, and p_n is the geometric average of the modified n-gram precisions. The first term is referred to as the brevity penalty that penalizes descriptions that are shorter than the reference description.

6.1.2 Recall Oriented Understudy for Gisting Evaluation (ROUGE, 2004)

ROUGE [53] metric was proposed in 2004 for text summaries evaluation. It computes *n–gram* based recall score for candidate sentences with respect to the reference sentences. Similar to BLEU, ROUGE is also computed by varying the *n–gram* count. However, BLEU is a precision-based measure (that measures how well a candidate sentence matches a set of reference sentences by computing the percentage of n-grams in the candidate sentence overlapping with the reference sentences) where as ROUGE is a recall based metric. Moreover, other than *n–gram* variants ROUGE$_n$, it has other versions known as ROUGE$_S$, ROUGE$_W$, ROUGE$_{SU}$ (extension of ROUGE$_S$) and ROUGE$_L$. The version used in image and video captioning evaluation is ROUGE$_L$ which computes recall and precision scores of the longest common subsequences (LCS) between candidate and each reference sentence. The metric compares common subsequences of words in candidate and reference sentences. The intuition behind is that the longer the LCS of candidate and reference sentences, the higher the similarity between the two summaries. The words need not be consecutive but should be in sequence. ROUGE-N is computed as

$$\text{ROUGE-N} = \frac{\sum_{S \in R_{Sum}} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in R_{Sum}} \sum_{g_n \in S} C(g_n)},$$

where n is the length of the n-gram, g_n, and $C_m(g_n)$ is the maximum number of n-grams co-occurring in candidate as well as reference summaries and R$_{Sum}$ stands for reference summaries.

To estimate the similarity between two summaries, X of length m and Y of length n, LCS-based F-measure score is computed. Where X is a reference summary sentence and Y is a candidate summary sentence, using the following equations:

$$R_{lcs} = \frac{\text{LCS}(X,Y)}{m},$$

$$P_{lcs} = \frac{\text{LCS}(X,Y)}{n},$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}},$$

Where LCS(X,Y) is the length of longest common subsequence between X and Y, $\beta = P_{lcs}/R_{lcs}$. The LCS-based F-measure score computed by equation F_{lcs} is known as ROUGE$_L$ score. ROUGE$_L$ is 1 when $X = Y$, and it is zero when there is nothing common between X and Y i.e. LCS$(X,Y) = 0$.

One of the advantages of ROUGE$_L$ is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order as n-grams. Moreover, it automatically includes longest in-sequence common n-grams, therefore, predefining the n-gram length is not required.

TABLE 3: Summary of metrics used for video description evaluation.

Metric Name	Designed For	Methodology
BLEU [52]	Machine translation	*n-gram* precision
ROUGE [53]	Document summarization	*n-gram* recall
METEOR [54]	Machine translation	*n-gram* with synonym matching
CIDEr [55]	Image captioning	*tf-idf* weighted *n-gram* similarity
SPICE [56]	Image captioning	Scene-graph synonym matching
WMD [57]	Document similarity	Earth mover distance on word2vec

6.1.3 Metric for Evaluation of Translation with Explicit Ordering (METEOR, 2005)

METEOR [54] was designed to explicitly fix the weaknesses in BLEU [52] of exact lexical match by introducing semantic matching. METEOR takes WordNet [158], a lexical database of the English language to account for various match levels, including exact token matches, stemmed tokens matches, synonymy matching as well as the paraphrase matching.

METEOR scores are computed based on the alignment between a candidate sentence and a set of reference sentences. Each sentence is taken as a set of unigrams and alignment is done by mapping unigrams of candidate and reference sentences. During mapping, every unigram in each sentence either maps to zero or one unigram in the other sentence. In case there are more than one alignments with the same number of mappings, the alignment with the fewest crosses is chosen i.e. with fewer intersections of two mappings. After the final alignment, the score is computed as follows:

First of all, unigram precision P is calculated as: $P = m/w_t$ where m is the number of unigrams in the candidate translation that are also found in the reference translation, and w_t is the number of unigrams in the candidate translation. Unigram recall is calculated as $R = m/w_r$ where m is same as for precision and w_r is the number of unigrams in the reference translation. Precision and recall are combined using the harmonic mean with recall weighted 9 times more than precision.

$$F_{mean} = \frac{10PR}{R + 9P},$$

The precision, recall and F-score measures account for congruity with respect to single words only but do not cater larger segments that appear in both the reference and the candidate sentence. Longer *n–gram* matches are used to compute a penalty p for the alignment. The penalty is calculated for the mappings that are not adjacent in the reference and the candidate sentence. To compute this penalty, unigrams are grouped into the fewest possible chunks, where a chunk is defined as a set of unigrams that are adjacent in the hypothesis and in the reference. A translation that is identical to the reference will give just one chunk. The penalty is then computed as

$$p = 0.5(\frac{c}{U_m})^2,$$

where c is the number of chunks and U_m is the number of unigrams that have been mapped. The final score for a segment is calculated as

$$M = F_{mean}(1 - p),$$

To calculate a score over a whole corpus or collection of segments, the aggregate values for P, R and p are then combined using the same formula.

In case of multiple reference sentences, the maximum METEOR score between the candidate and reference sentence is considered. To date, METEOR has shown better correlation with human judgements compared to BLEU. Moreover, a recent study [159] has shown METEOR to be a better evaluation metric compared to contemporary metrics.

6.1.4 Consensus based Image Description Evaluation (CIDEr, 2015)

CIDEr [55] is a recent metric proposed for evaluating image captions. Given an image I_i, CIDEr evaluates how well a candidate sentence c_i matches the consensus of a set of image descriptions $S_i = s_{i1}, \ldots, s_{im}$. It performs stemming and converts all the words from candidate as well as reference sentences into their root forms e.g. "stems", "stemmer", "stemming", and "stemmed" to their root word "stem". CIDEr treats each sentence as a set of n-grams containing 1 to 4 words. To encode the consensus between candidate sentence and reference sentence, it measures the coexistence frequency of n-grams in both the sentences. Similarly, n-grams not present in the reference sentence are not expected to be in the candidate sentence. Finally, n-grams that are very common among the reference sentences of all the images are given lower weightage as they are likely to be less informative about the image content and more biased towards lexical structure of the sentences. The weightage for each n-gram is encoded using Term Frequency Inverse Document Frequency (TF-IDF) [160]. The term "TF" puts higher weightage on frequently occurring n-grams in the reference sentence of the image, whereas "IDF" puts lower weightage on those n-grams that commonly appear across the whole dataset.

Finally, CIDEr_n score for n-grams of length n is calculated using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m}\sum_j \frac{g^n(c_i).g^n(s_{ij})}{\|g^n(c_i)\|.\|g^n(s_{ij})\|},$$

where $g^n(c_i)$ is a vector corresponding to all n-grams of length n and $\|g^n(c_i)\|$ is the magnitude of the vector $g^n(c_i)$. Same is true for $g^n(s_{ij})$. Further, CIDEr uses higher order n-grams (higher the order, longer the sequence of words) to capture the grammatical properties and richer semantics of

the text. For that matter, it combines the scores of different n-grams using the following equation:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i),$$

The most popular version of CIDEr in image and video description evaluation is CIDEr-D that incorporates a few modifications in the originally proposed CIDEr to prevent higher scores for the captions that badly fail in human judgements.

6.1.5 Word Mover's Distance (WMD, 2015)

The WMD [57] makes use of word embeddings which are semantically meaningful vector representations of words learnt from text corpora. WMD distance measures the dissimilarity between two text documents. Two captions with different words may still have the same semantic meanings. On the contrary, two captions may contain the same objects, attributes or relations and yet their meanings could be completely different. WMD was proposed to address this problem. This is because word embeddings are good at capturing semantic meanings and are easier to compute than WordNet thanks to the distributed vector representations of words. The distance between two texts is casted as an Earth Mover's Distance (EMD) [161] typically used in transportation to calculate the travel cost using word2vec embeddings [162]. In this metric, captions or text documents are first transformed into normalized bag-of-words (nBOW) vectors, accounting for all words except special words like start and stop words. Semantic similarities between individual word pairs are incorporated into the document similarity metric by using Euclidean distance in the word2vec embedding space. The distance between two documents or captions is then defined as the cost required to move all words between captions. Figure 10 illustrates an example WMD calculation process. The cost is minimized as a linear optimization problem, which can be solved as a special case of EMD [161]. Compared to BLUE, ROUGE and CIDEr, WMD is less sensitive to words order or synonym swapping. Further, similar to CIDEr and METEOR, it gives high correlation against human judgements.

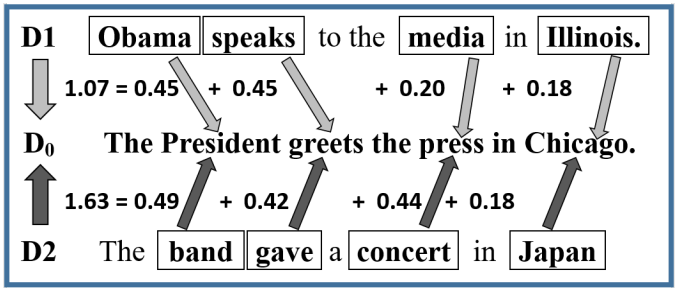


Fig. 10: Components of the WMD metric between a query D_0 and two sentences D_1 and D_2 with the same BOW distance. The arrows show flow between two words and are labeled with their distance contribution. Figure adapted from [57].

6.1.6 Semantic Propositional Image Captioning Evaluation (SPICE, 2016)

SPICE [56] is the latest proposed evaluation metric for image and video descriptions. It is based on scene graph tuples of the candidate descriptions and all the reference sentences. Scene graph is used to parse the candidate sentence to semantic tokens such as object classes C, relation types R and attribute types A. SPICE parses candidate caption c into a scene graph tuple using relationship

$$G(c) = [O(c), E(c), K(c)]$$

where $G(c)$ denotes the scene graph of caption c, $O(c)$ is the set of objects, $E(c)$ represents relationships between objects and $K(c)$ represents objects attributes. After successful parsing, a set of tuples is created using the elements of G with all possible combinations. SPICE is computed based on F1-score between the candidate and reference caption tuples. Like METEOR, SPICE also uses WordNet for synonym matching. Although, in the current literature, the SPICE score has not been reported much but one obvious limiting factor on its performance is the quality of the parsing.

6.2 Human Evaluations

Given the lack of reference captions and low correlation with human judgements of automated evaluation metrics, human evaluations are also often used to judge the quality of machine generated captions. Human evaluations may either be crowed-sourced such as AMT workers or specialist judges as in some competitions. Such human evaluations can be further structured using measurements such as *Relevance* or *Grammar Correctness*. In relevance based evaluation, video content relevance is given subjective scores, with highest score given to the *"Most Relevant"* and minimum score to the *"Least Relevant"*. No two sentences should be given the same score unless they are identical. In the approaches where grammar correctness is measured, the sentences are graded based on grammatical correctness without showing the video content to the evaluators in which case, more than one sentence may have the same score.

6.3 Drawbacks of Evaluation Metrics

In a recent study by Kilickaya et.al. [163], several experiments were performed to evaluate the behavior of metric scores. First, the original caption was evaluated with itself and maximum score was achieved by all the metrics. Then small modifications were introduced to candidate sentences in order to measure how the evaluation metrics score changes due to these changes. It was reported that all the scores from all metrics were reduced even when the original words were replaced with their synonyms. Out of all six metrics, the change was more prominent in SPICE and CIDEr due to incorrect *parsing*, failure of synonym matching and unbalanced *tf-idf* weighting. Further study reveals that most metrics are not affected much with the addition of few redundant words in the sentences. However, if the order of words in a sentence is changed, BLEU, ROUGE and CIDEr scores are affected significantly due to their n-gram methodology. SPICE and WMD are not affected much by word order changes in a sentence.

6.4 Reliability of Evaluation Metrics

A good method to evaluate the video descriptions is to compare the machine generated captions with the reference captions annotated by humans. However, as shown in Figure 9, the reference captions can vary within itself and can only represent few samples out of all valid samples for the same video clip. Having more reference sample captions create a better solution space and hence lead to more reliable evaluation.

Another aspect of the evaluation problem is the syntactic variations in candidate sentences. This problem also exists in the well studied field of machine translation. In this case, a sentence in a source language can be translated into various sentences in a target language. Syntactically different sentences may still have the same semantic content.

In a nutshell, evaluation metrics evaluate the suitability of a caption to the visual input by comparing how well the candidate caption matches with that of reference caption(s). The performance of the metrics improved with increased number of reference captions as found in [55]. Numerous studies [40], [46], [55], [55], [164] also found that CIDEr and METEOR have higher correlations to human judgements and are regarded superior amongst the contemporary reported metrics. WMD and SPICE are very recent automatic caption evaluation metrics and have not been reported much in the literature at the time of this survey.

7 Benchmark Results

We summarize the benchmark results of various techniques applied to the video description problem on different datasets in Table 4. We group the methods based on the dataset they reported results on and then list them in chronological order. Moreover, for multiple variants of the same model, only their best reported results are selected. For a detailed analysis of each method and its variants, the original paper should be consulted. In addition, where multiple n-gram scores are reported for the BLEU metric, we have chosen only the BLEU@4 result as this is the closest to human evaluations. From Table 4, we can see that most methods have reported results on the MSVD dataset, followed by MPII-MD, M-VAD, MSR-VTT and ActivityNet Captions. The popularity of MSVD may be attributed to the diverse nature (i.e. open domain) of videos and the large number of reference captioning. MPII-MD, M-VAD, MSR-VTT and ActivityNet Captions are popular because of their size and their usage in competitions (see Section 5).

Another key observation is that earlier works are mainly reporting results in terms of subject, verb, object (SVO) and in some cases place (scene) detection accuracies in the video, whereas more recent works started to report sentence level matches using the automatic evaluation metrics. Considering the diverse nature of the datasets and the limitations of automatic evaluation metrics, we have analyzed the results of different methods in Table 4 using four popular metrics namely BLEU, METEOR, CIDEr and ROUGE.

For the MSVD dataset, LSTM-TSA [165] has achieved the best BLEU and METEOR score. However, it stands at third position on CIDEr score whereas SCN-LSTM [166] takes the first position for best CIDEr score and joins LSTM-TSA in the best METEOR score. However, SCN-LSTM stands at number two in the BLEU metric score. TDDF [167] stands at second position on METEOR score followed by HRNE [164] at third place. On TACoS Multilevel dataset, h-RNN [46] has produced best results on all reported metrics i.e. BLEU, METEOR and CIDEr.

On the challenging M-VAD dataset, the reported results are overall very poor, however, within the presented results we see that so far only Temporal-Attention [115] and HRNE [164] reported results using the BLEU metric reporting a BLEU score of 0.7. All papers reported METEOR results for this dataset and so far BAE [168] has produced the best METEOR score followed by LSTM-TSA [165]. HRNE [164] and Glove+Deep Fusion Ensemble [119] share the third place for METEOR score.

MPII-MD is another very challenging dataset and still has very low benchmark results similiar to the M-VAD dataset. Only BAE [168] has reported BLEU score for this dataset. LSTM-TSA [165] has achieved the best METEOR score followed by LSTM-E [43] and S2VT [40] at second and third place respectively.

Results on another very popular dataset MSR-VTT are overall better than M-VAD and MPII-II. As shown in Table 4, CST-GT-None [131] has reported the highest score on all four metrics. The second highest BLEU score is reported by DenseVidCap [169] followed by HRL [129] respectively. However, HRL [129] reported the second best METEOR, CIDEr and ROUGE scores. We report four results for ActivityNet datasets and two for LSMDC and one for Charades at the end of Table 4.

8 Discussion and Conclusion

We presented the first comprehensive survey of video description research covering algorithms and models reported in the literature, public datasets that are available for training and testing the models, various metrics that are used for evaluation and benchmark results on these datasets. We categorized the existing literature into Subject-Verb-Object (SVO-Triplet) and Deep Learning based models and presented their detailed analysis. Despite extensive research efforts, our survey shows that the results are still far below an acceptable level for most real world applications.

One of the main bottlenecks hindering progress along this line of research is the lack of effective and purpose designed video description evaluation metrics. Current metrics have been adopted from machine translation or image captioning and fall short in measuring the quality of machine generated captions and their agreement with human judgements. One way to improve these metrics is to increase the number of refererence sentences. However, larger number of reference sentences is expensive to collect.

From an algorithm design perspective, although LSTMs have shown competitive caption generation performance, the interpretablity and intelligibility of the underlying model are low. Specifically, it is hard to differentiate how much visual features have contributed to the generation of a specific word compared to the bias that comes naturally from the language model adopted. This problem is exacerbated when the aim is to diagnose the generation of erroneous captions. For example, when we see a caption "red fire hydrant" generated by a video description model

TABLE 4: Performance of video description methods on varius benchmark datasets.

Techniques / Models / Methods	Yr	Dataset	Results			
			BLEU	**METEOR**	**CIDEr**	**ROUGE**
Sentence-Tracker [29]	2012	Mind's Eye	Human judgements			
RBS+RBS & RF-TP+RBS [85]	2012	MSVD	SVO Accuracy			
SVO-LM (VE) [91]	2013	MSVD	0.45+_0.05	0.36+_0.27		
FGM [90]	2014	MSVD	SVOP Accuracy			
LSTM-YT [111]	2015	MSVD	33.3	29.1	-	-
Temporal-Attention (TA) [115]	2015	MSVD	41.9	29.6	51.67	-
S2VT [40]	2015	MSVD	-	29.8	-	-
h-RNN [46]	2016	MSVD	49.9	32.6	65.8	-
MM-VDN [170]	2016	MSVD	37.6	29.0	-	-
Glove + Deep Fusion Ensemble [119]	2016	MSVD	42.1	31.4	-	-
S2FT [171]	2016	MSVD	-	29.9	-	-
HRNE [164]	2016	MSVD	43.8	33.1	-	-
GRU-RCN [172]	2016	MSVD	43.3	31.6	68.0	-
LSTM-E [43]	2016	MSVD	45.3	31.0	-	-
SCN-LSTM [166]	2017	MSVD	51.1	**33.5**	**77.7**	-
LSTM-TSA [165]	2017	MSVD	**52.8**	**33.5**	74.0	-
TDDF [167]	2017	MSVD	45.8	33.3	73.0	69.7
BAE [168]	2017	MSVD	42.5	32.4	63.5	-
PickNet [132]	2018	MSVD	46.1	33.1	76.0	69.2
MRF [35]	2013	TACoS	5.6	-	-	-
SMT(SR) + Prob I/P [44]	2014	TACoS MLevel	28.5	-	-	-
CRF + LSTM-Decoder [21]	2015	TACoS MLevel	28.8	-	-	-
h-RNN [46]	2016	TACoS MLevel	**30.5**	**28.7**	**160.2**	-
JEDDi-Net [173]	2018	TACoS MLevel	18.1	23.85	103.98	50.85
Temporal-Attention (TA) [115]	2015	M-VAD	**0.7**	5.7	6.1	-
S2VT [40]	2015	M-VAD	-	6.7	-	-
Visual-Labels [120]	2015	M-VAD	-	6.4	-	-
HRNE [164]	2016	M-VAD	**0.7**	6.8	-	-
Glove + Deep Fusion Ensemble [119]	2016	M-VAD	-	6.8	-	-
LSTM-E [43]	2016	M-VAD	-	6.7	-	-
LSTM-TSA [165]	2017	M-VAD	-	7.2	-	-
BAE [168]	2017	M-VAD	-	**7.3**	-	-
S2VT [40]	2015	MPII-MD	-	7.1	-	-
Visual-Labels [120]	2015	MPII-MD	-	7.0	-	-
SMT [98]	2015	MPII-MD	-	5.6	-	-
Glove + Deep Fusion Ensemble [119]	2016	MPII-MD	-	6.8	-	-
LSTM-E [43]	2016	MPII-MD	-	7.3	-	-
LSTM-TSA [165]	2017	MPII-MD	-	**8.0**	-	-
BAE [168]	2017	MPII-MD	**0.8**	7.0	**10.8**	**16.7**
Alto [174]	2016	MSR-VTT	39.8	26.9	45.7	59.8
VideoLab [175]	2016	MSR-VTT	39.1	27.7	44.4	60.6
RUC-UVA [176]	2016	MSR-VTT	38.7	26.9	45.9	58.7
v2t-navigator [177]	2016	MSR-VTT	40.8	28.2	44.8	61.1
TDDF [167]	2017	MSR-VTT	37.3	27.8	43.8	59.2
DenseVidCap [169]	2017	MSR-VTT	41.4	28.3	48.9	61.1
CST-GT-None [131]	2017	MSR-VTT	**44.1**	**29.1**	**49.7**	**62.4**
PickNet [132]	2018	MSR-VTT	38.9	27.2	42.1	59.5
HRL [129]	2018	MSR-VTT	41.3	28.7	48.0	61.7
Dense-Cap Model [30]	2017	ActivityNet	**3.98**	9.5	**24.6**	-
LSTM-A+PG+R [178]	2017	ActivityNet	-	**12.84**	-	-
TAC [179]	2017	ActivityNet	-	9.61	-	-
JEDDi-Net [173]	2018	ActivityNet	1.63	8.58	19.88	**19.63**
CT-SAN [180]	2016	LSMDC	**0.8**	7.1	**10.0**	**15.9**
GEAN [181]	2017	LSMDC	-	**7.2**	9.3	15.6
HRL [129]	2018	Charades	18.8	19.5	23.6	41.4

from a frame containing a white fire hydrant, it is difficult to diagnose whether the color feature is incorrectly encoded by the visual feature extractor or is due to the bias in the used language model towards red fire hydrants.

Some challenges come from the diverse nature of the videos themselves. For instance, multiple activities in a video, where captions represent only some activities, could lead to low video description performance of a model. Similarly, longer duration videos pose further challenges since most action features can only encode short term actions such as trajectory features and C3D features [182] that are dependent on video segment lengths. Finally, most feature extractors are suitable only for static or smoothly changing images and hence struggle to handle abrupt scene changes.

We reviewed popular benchmark datasets that are commonly used for training and testing video description models. We also discussed four competitions/challenges that are regularly held to promote the video analysis research in general and the video description research more specifically. We discussed in detail the available automatic evaluation metrics that are used for video description highlighting their attributes and drawbacks. Lastly, we presented a comprehensive summary of results obtained by recent methods on the benchmark datasets. These results not only show the relative performance of existing methods but also highlight the relative difficulty of the datasets and strictness of the evaluation metrics. In a nut shell, the merge of vision and language is a rapidly growing research area with numerous promising real world applications. This survey give readers a clear picture of what has been achieved in this field so far and where the gaps exist so that future research efforts can be better focused.

Acknowledgement

The research was supported by ARC Discovery Grant DP160101458.

References

[1] "Language in Vision," https://www.sciencedirect.com/journal/computer-vision-and-image-understanding/vol/163, 2017.

[2] B. Andrei, E. Georgios, H. Daniel, M. Krystian, N. Siddharth, X. Caiming, and Z. Yibiao, "A Workshop on Language and Vision at CVPR 2015," http://languageandvision.com/2015.html, 2015.

[3] B. Andrei, M. Tao, N. Siddharth, D. Puneet, Z. Quanshi, S. Nishant, L. Jiebo, and S. Rahul, "A Workshop on Language and Vision at CVPR 2017," http://languageandvision.com/2017.html, 2017.

[4] B. Andrei, M. Tao, N. Siddharth, Z. Quanshi, S. Nishant, L. Jiebo, and S. Rahul, "A Workshop on Language and Vision at CVPR 2018," http://languageandvision.com/, 2018.

[5] R. Anna, T. Atousa, R. Marcus, P. Christopher, L. Hugo, C. Aaron, and S. Bernt, "The Joint Video and Language Understanding Workshop at ICCV 2015," https://sites.google.com/site/describingmovies/workshop-at-iccv-15, 2015.

[6] R. Anna, T. Makarand, T. Atousa, M. Tegan, R. Marcus, F. Sanja, P. Christopher, and S. Bernt, "The Joint Video and Language Understanding Workshop at ICCV 2017," https://sites.google.com/site/describingmovies/workshop-at-iccv-17, 2017.

[7] M. Margaret, M. Ishan, H. Ting-Hao, and F. Frank, "Story Telling Workshop and Visual Story Telling Challenge at NAACL 2018," http://www.visionandlanguage.net/workshop2018/, 2018.

[8] D. Roy and E. Reiter, "Connecting language to the world," *Artificial Intelligence*, vol. 167, no. 1–2, 2005. [Online]. Available: /ref/roy/ConnectingLanguage.pdf

[9] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1–2, 2005. [Online]. Available: /ref/roy/SemioticSchemas.pdf

[10] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1107–1135, 2003.

[11] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth, "Names and faces in the news," in *CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[12] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei, "Construction and analysis of a large scale image ontology," *Vision Sciences Society*, vol. 186, p. 2, 2009.

[13] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*. Springer, 2010, pp. 15–29.

[14] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating image descriptions," in *Proceedings of the 24th CVPR*. Citeseer, 2011.

[15] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *CNLL*. Association for Computational Linguistics, 2011, pp. 220–228.

[16] S. Guadarrama, L. Riano, D. Golland, D. Go, Y. Jia, D. Klein, P. Abbeel, T. Darrell *et al.*, "Grounding spatial relations for human-robot interaction," in *IROS, 2013*. IEEE, 2013, pp. 1640–1647.

[17] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 2010, pp. 251–258.

[18] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation." in *AAAI*, 2011.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *ACL*, vol. 2, pp. 67–78, 2014.

[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term rcnn for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[23] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *CVPR*, 2015, pp. 1473–1482.

[24] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *ICCV*, 2015, pp. 2533–2541.

[25] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems*, 2014, pp. 1682–1690.

[26] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[27] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Advances in neural information processing systems*, 2015, pp. 2953–2961.

[28] L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual madlibs: Fill in the blank description generation and question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2461–2469.

[29] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi *et al.*, "Video in sentences out," *arXiv preprint arXiv:1204.2742*, 2012.

[30] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," *arXiv:1705.00754*, 2017.

[31] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422–2431.

[32] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[33] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions." *TACL*, vol. 2, no. 10, pp. 351–362, 2014.

[34] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal rnn," *arXiv preprint arXiv:1412.6632*, 2014.

[35] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 433–440.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[37] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[38] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2641–2649.

[39] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.

[40] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4534–4542.

[41] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

[42] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5288–5296.

[43] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4594–4602.

[44] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, "Coherent multi-sentence video description with variable level of detail," in *German conference on pattern recognition*. Springer, 2014, pp. 184–195.

[45] A. Shin, K. Ohnishi, and T. Harada, "Beyond caption to narrative: Video captioning with multiple sentences," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3364–3368.

[46] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4584–4593.

[47] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *CVPR*, 2013, pp. 2634–2641.

[48] M. U. G. Khan, L. Zhang, and Y. Gotoh, "Human focused video description," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1480–1487.

[49] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann, "Beyond audio and video retrieval: towards multimedia summarization," in *2nd ACM ICMR*, 2012, p. 2.

[50] M. U. G. Khan and Y. Gotoh, "Describing video contents in natural language," in *Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*. Association for Computational Linguistics, 2012, pp. 27–35.

[51] H. Yu and J. M. Siskind, "Grounded language learning from video sentences." in *ACL(1)*, 2013, pp. 53–63.

[52] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on ACL*. Association for Computational Linguistics, 2002, pp. 311–318.

[53] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[54] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, 2005, pp. 65–72.

[55] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[56] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[57] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[58] N. Morsillo, G. S. Mann, and C. J. Pal, "Youtube scale, large vocabulary video annotation." *Video Search and Mining*, vol. 287, pp. 357–386, 2010.

[59] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *CVPR*, 2016, pp. 4575–4583.

[60] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[61] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *null*. IEEE, 2003, p. 273.

[62] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[63] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[64] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[65] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[66] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 2241–2248.

[67] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[68] S. Hongeng, F. Brémond, and R. Nevatia, "Bayesian framework for video surveillance application," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 1. IEEE, 2000, pp. 164–170.

[69] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 742–749.

[70] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 12, pp. 1325–1337, 1997.

[71] D. Koller, N. Heinze, and H.-H. Nagel, "Algorithmic characterization of vehicle trajectories from image sequences by motion verbs," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*. IEEE, 1991, pp. 90–95.

[72] C. S. Pinhanez and A. F. Bobick, "Human action detection using pnf propagation of temporal constraints," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, 1998, pp. 898–904.

[73] S.-C. Zhu, D. Mumford *et al.*, "A stochastic grammar of images," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2007.

[74] D. Moore and I. Essa, "Recognizing multitasked activities from video using stochastic context-free grammar," in *AAAI/IAAI*, 2002, pp. 770–776.

[75] I. Langkilde-Geary and K. Knight, "Halogen input representation."

[76] C. Pollard and I. A. Sag, *Head-driven phrase structure grammar*. University of Chicago Press, 1994.

[77] E. Reiter and R. Dale, *Building natural language generation systems*. Cambridge university press, 2000.

[78] F. Nishida, S. Takamatsu, T. Tani, and T. Doi, "Feedback of correcting information in postediting to a machine translation system," in *Proceedings of the 12th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1988, pp. 476–481.

[79] F. Nishida and S. Takamatsu, "Japanese-english translation through internal expressions," in *Proceedings of the 9th conference on Computational linguistics-Volume 1*. Academia Praha, 1982, pp. 271–276.

[80] M. W. Lee, A. Hakeem, N. Haering, and S.-C. Zhu, "Save: A framework for semantic annotation of visual events," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[81] R. Nevatia, J. Hobbs, and B. Bolles, "An ontology for video event representation," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*. IEEE, 2004, pp. 119–119.

[82] P. Kuchi, P. Gabbur, P. S. Bhat, and S. S. David, "Human face detection and tracking using skin color modeling and connected component operators," *IETE journal of research*, vol. 48, no. 3-4, pp. 289–293, 2002.

[83] I. Maglogiannis, D. Vouyioukas, and C. Aggelopoulos, "Face detection and recognition of natural human emotion using markov random fields," *Personal and Ubiquitous Computing*, vol. 13, no. 1, pp. 95–101, 2009.

[84] W. Kim, J. Park, and C. Kim, "A novel method for efficient indoor–outdoor image classification," *Journal of Signal Processing Systems*, vol. 61, no. 3, pp. 251–258, 2010.

[85] P. Hanckmann, K. Schutte, and G. J. Burghouts, "Automated textual descriptions for a wide range of video events with 48 human actions," in *European Conference on Computer Vision*. Springer, 2012, pp. 372–380.

[86] G. Burghouts, H. Bouma, R. de Hollander, S. Van den Broek, and K. Schutte, "Recognition of 48 human behaviors from video," in *Int. Symp. Optronics in Defense and Security, OPTRO*, 2012.

[87] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 2241–2248.

[88] C. Tomasi and T. Kanade, "Detection and tracking of point features," 1991.

[89] J. Shi *et al.*, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[90] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild." in *Coling*, vol. 2, no. 5, 2014, p. 9.

[91] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge." in *AAAI*, vol. 1, 2013, p. 2.

[92] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "Recognizing and describing activities using semantic hierarchies and zero-shot recognition," in *ICCV*, 2013, pp. 2712–2719.

[93] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[94] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[95] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 2007, pp. 177–180.

[96] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[97] D. L. Chen, W. B. Dolan, S. Raghavan, T. Huynh, and R. Mooney, "Collecting highly parallel data for paraphrase evaluation," in *JAIR: - Volume 37*. Association for Computational Linguistics, 2010, pp. 397–435.

[98] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[99] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," *arXiv preprint arXiv:1503.01070*, 2015.

[100] J. Corso, "Gbs: Guidance by semantics-using high-level visual inference to improve vision-based mobile robot localization," STATE UNIV OF NEW YORK AT BUFFALO AMHERST, Tech. Rep., 2015.

[101] C. Sun and R. Nevatia, "Semantic aware video transcription using random forest classifiers," in *European Conference on Computer Vision*. Springer, 2014, pp. 772–786.

[102] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework." in *AAAI*, vol. 5, 2015, p. 6.

[103] H. Yu and J. M. Siskind, "Learning to describe video with weak supervision by exploiting negative sentential information." in *AAAI*, 2015, pp. 3855–3863.

[104] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

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

[106] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions." Cvpr, 2015.

[107] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[108] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[109] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[110] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1764–1772.

[111] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[112] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *ACL: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 190–200.

[113] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[114] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[115] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *ICCV*, 2015, pp. 4507–4515.

[116] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*. Springer, 2006, pp. 428–441.

[117] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, pp. 124–1.

[118] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[119] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving lstm-based video description with linguistic knowledge mined from text," *arXiv preprint arXiv:1604.01729*, 2016.

[120] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," in *German Conference on Pattern Recognition*. Springer, 2015, pp. 209–221.

[121] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[122] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning*, 2015, pp. 843–852.

[123] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[124] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[125] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," in *Advances in neural information processing systems*, 2016, pp. 3675–3683.

[126] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu, "Feudal networks for hierarchical reinforcement learning," *arXiv preprint arXiv:1703.01161*, 2017.

[127] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," *arXiv preprint arXiv:1704.03899*, 2017.

[128] R. Pasunuru and M. Bansal, "Reinforced video captioning with entailment rewards," *arXiv preprint arXiv:1708.02300*, 2017.

[129] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," *arXiv preprint arXiv:1711.11135*, 2017.

[130] L. Li and B. Gong, "End-to-end video captioning with multitask reinforcement learning," *arXiv preprint arXiv:1803.07950*, 2018.

[131] S. Phan, G. E. Henter, Y. Miyao, and S. Satoh, "Consensus-based sequence training for video captioning," *arXiv preprint arXiv:1712.09532*, 2017.

[132] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," *arXiv preprint arXiv:1803.01457*, 2018.

[133] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[134] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[135] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1194–1201.

[136] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[137] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

[138] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *European conference on computer vision*. Springer, 2016, pp. 609–625.

[139] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *European Conference on Computer Vision*. Springer, 2012, pp. 144–157.

[140] "Castingwords transcription service." http://castingwords.com/, 2014.

[141] "Describing and understanding video and the Large Scale Movie Description Challenge (LSMDC)," https://sites.google.com/site/describingmovies/, 2015.

[142] "The Large Scale Movie Description Challenge (LSMDC) online evaluations," https://competitions.codalab.org/competitions/6121#learn_the_details-evaluation, 2017.

[143] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[144] "The Large Scale Movie Description Challenge (LSMDC) online results," https://competitions.codalab.org/competitions/6121#results, 2017.

[145] "Microsoft Research - Video to Text (MSR-VTT) challenge," http://ms-multimedia-challenge.com/2016/challenge, 2016.

[146] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[147] "TREC Video Retrieval Evaluation (TRECVID) challenge," https://trecvid.nist.gov/.

[148] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Aly, G. J. Jones *et al.*, "Trecvid 2016. evaluating video search, video event detection, localization and hyperlinking," 2016.

[149] L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese, "Umbc_ebiquity-core: semantic textual similarity systems," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, vol. 1, 2013, pp. 44–52.

[150] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, "Can machine translation systems be evaluated by the crowd alone," *Natural Language Engineering*, vol. 23, no. 1, pp. 3–30, 2017.

[151] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva *et al.*, "Findings of the 2017 conference on machine translation (wmt17)," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 169–214.

[152] Y. Graham, G. Awad, and A. Smeaton, "Evaluation of automatic video captioning using direct assessment," *arXiv preprint arXiv:1710.10586*, 2017.

[153] "Activity Net Captions Challenge, Task5: Dense-Captioning Events in Videos," http://activity-net.org/challenges/2017/index.html#, 2017.

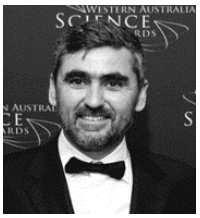
[154] "Activity Net challenge," http://activity-net.org/challenges/2017/index.html#, 2017.

[155] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, R. Khrisna, V. Escorcia, K. Hata, and S. Buch, "Activitynet challenge 2017 summary," *arXiv preprint arXiv:1710.08011*, 2017.
[156] "Activity Net Captions Challenge, Evaluations," https://github.com/ranjaykrishna/densevid_eval, 2017.
[157] "Activity Net Captions Challenge, Results," http://activity-net.org/challenges/2017/evaluation.html, 2017.
[158] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.
[159] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, vol. 452, no. 457, 2014, p. 457.
[160] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.
[161] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.
[162] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
[163] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," *arXiv preprint arXiv:1612.07600*, 2016.
[164] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1029–1038.
[165] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[166] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017.
[167] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Task-driven dynamic fusion: Reducing ambiguity in video description," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[168] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[169] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[170] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, and K. Saenko, "A multi-scale multiple instance video description network," *arXiv preprint arXiv:1505.05914*, 2015.
[171] Y. Liu and Z. Shi, "Boosting video description generation by explicitly translating from frame-level captions," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 631–634.
[172] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *arXiv preprint arXiv:1511.06432*, 2015.
[173] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko, "Joint event detection and description in continuous video streams," *arXiv preprint arXiv:1802.10250*, 2018.
[174] R. Shetty and J. Laaksonen, "Frame-and segment-level features and candidate pool evaluation for video caption generation," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1073–1076.
[175] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko, "Multimodal video description," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1092–1096.
[176] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. Snoek, "Early embedding and late reranking for video captioning," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1082–1086.
[177] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann, "Describing videos using multi-modal fusion," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1087–1091.
[178] T. Yao, Y. Li, Z. Qiu, F. Long, Y. Pan, D. Li, and T. Mei, "Msr asia msm at activitynet challenge 2017: Trimmed action recognition, temporal action proposals and dense-captioning events in videos."
[179] J. Qin, C. Shizhe, C. Jia, Chen, and H. Alexander, "Ruc-cmu: System descriptions for the dense video captioning task," *arXiv preprint arXiv:1710.08011*, 2017.
[180] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," *arXiv preprint arXiv:1610.02947*, 2016.
[181] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim, "Supervising neural attention models for video captioning by human gaze data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[182] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR*, vol. abs/1412.0767, 2014. [Online]. Available: http://arxiv.org/abs/1412.0767

Nayyer Aafaq received BE degree with distinction in Avionics from the College of Aeronautical Engineering (CAE), National University of Sciences and Technology (NUST), Pakistan, in 2007 and MS degree with high distinction in Systems Engineering from Queensland University of Technology (QUT), Australia, in 2012. He is currently working towards the Ph.D. degree at The University of Western Australia (UWA). He is a recipient of SIRF scholarship at UWA. He has served as a Research Assistant at STG Research Institute, Pakistan, from 2007 to 2011 and as a lecturer at College of Aeronautical Engineering (CAE), NUST, Pakistan from 2013 till 2017. His current research interests includes Deep Learning, Video Analysis and intersection of Natural Language Processing (NLP), Computer Vision (CV) and Machine Learning.

Syed Zulqarnain Gilani received his PhD from the University of Western Australia where he is now working as a Research Fellow. He did his MS in EE from the National University of Sciences and Technology (NUST), Pakistan in 2009 and secured the Presidents Gold Medal. His research interests include 3D facial morphometrics with applications to syndrome delineation and machine learning.

Wei Liu received her PhD from the University of Newcastle, Australia in 2003. She is now working at the Department of Computer Science and Software Engineering at the University of Western Australia, and co-lead the faculty's Big Data research group. Her research impact in the field of knowledge discovery from natural language text data is evident by a series of highly cited papers, and the reputable top data mining and knowledge management journals and conferences that she has been published in. These include for example, ACM Computer Surveys, Journal of Data Mining and Knowledge Discovery, Knowledge and Information Systems, International Conference on Data Engineering (ICDE), ACM International Conference on Information and Knowledge Management (CIKM). She has won three Australian Research Council Grants and several industry grants. Her current research focuses on deep learning methods for knowledge graph construction from natural language text, sequential data mining and text mining.



Ajmal Mian is an Associate Professor of Computer Science at The University of Western Australia. He completed his PhD from the same institution in 2006 with distinction and received the Australasian Distinguished Doctoral Dissertation Award from Computing Research and Education Association of Australasia. He received the prestigious Australian Postdoctoral and Australian Research Fellowships in 2008 and 2011 respectively. He received the UWA Outstanding Young Investigator Award 2011, the West Australian Early Career Scientist of the Year 2012 award, the Vice-Chancellors Mid-Career Research Award 2014 and the Aspire Professional Development Award 2016. He has published over 150 scientific papers in reputable journals and conferences. He has secured seven Australian Research Council grants, a National Health and Medical Research Council grant and a DAAD German Australian research cooperation grant. He has served as a guest editor of Pattern Recognition, Computer Vision and Image Understanding and Image and Vision Computing journals. His research interests include computer vision, machine learning, 3D shape analysis, hyperspectral image analysis, pattern recognition, and multimodal biometrics.