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

NAYYER AAFAQ, AJMAL MIAN, and WEI LIU, The University of Western Australia, Australia
SYED ZULQARNAIN GILANI, The University of Western Australia and Edith Cowan University
MUBARAK SHAH, University of Central Florida, USA

Video description is the automatic generation of natural language sentences that describe the contents of a given video. It has applications in human-robot interaction, helping the visually impaired and video subtitling. The past few years have seen a surge of research in this area due to the unprecedented success of deep learning in computer vision and natural language processing. Numerous methods, datasets, and evaluation metrics have been proposed in the literature, calling the need for a comprehensive survey to focus research efforts in this flourishing new direction. This article fills the gap by surveying the state-of-the-art approaches with a focus on deep learning models; comparing benchmark datasets in terms of their domains, number of classes, and repository size; and identifying the pros and cons of various evaluation metrics, such as SPICE, CIDEr, ROUGE, BLEU, METEOR, and WMD. Classical video description approaches combined subject, object, and verb detection with template-based language models to generate sentences. However, the release of large datasets revealed that these methods cannot cope with the diversity in unconstrained open domain videos. Classical approaches were followed by a very short era of statistical methods that were soon replaced with deep learning, the current state-of-the-art in video description. Our survey shows that despite the fast-paced developments, video description research is still in its infancy due to the following reasons: Analysis of video description models is challenging, because it is difficult to ascertain the contributions towards accuracy or errors of the visual features and the adopted language model in the final description. Existing datasets neither contain adequate visual diversity nor complexity of linguistic structures. Finally, current evaluation metrics fall short of measuring the agreement between machine-generated descriptions with that of humans. We conclude our survey by listing promising future research directions.

CCS Concepts: • **Computing methodologies → Computer vision**; *Natural language processing*;

Additional Key Words and Phrases: Video description, video captioning, video to text, language in vision

ACM Reference format:
Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video Description: A Survey of Methods, Datasets, and Evaluation Metrics. *ACM Comput. Surv.* 52, 6, Article 115 (October 2019), 37 pages.
https://doi.org/10.1145/3355390

This research was supported by ARC Discovery Grants DP160101458 and DP190102443. Research was also sponsored in part by the Army Research Office and was accomplished under Grant no. W911NF-19-1-0356.
Authors' Addresses: N. Aafaq (corresponding author), A. Mian, W. Liu, and S. Z. Gilani, The University of Western Australia, 35 Stirling Hwy, WA, 6009; emails: nayyer.aafaq@research.uwa.edu.au; {ajmal.mian, wei.liu, zulqarnain.gilani}@uwa.edu.au; M. Shah, University of Central Florida, 4000 Central Florida Blvd, Orlando, Florida, 32816; email: shah@crcv.ucf.edu.

0360-0300/2019/10-ART115 $15.00
https://doi.org/10.1145/3355390

1 INTRODUCTION

Describing a short video in natural language is a trivial task for most people, but a very challenging one for machines. Automatic video description involves understanding of many entities and the detection of their occurrences in a video employing computer vision techniques. These entities include *background scene, humans, objects, human actions, human-object interactions, human-human interactions, other events*, and the *order* in which events occur. All this information must then be articulated using a comprehensible and grammatically correct text employing Natural Language Processing (NLP) techniques. Over the past few years, these 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 images and videos. Special issues of journals are published focusing on language in vision [2] and workshops uniting the two areas have also been held regularly at both NLP and CV conferences [6–8, 89].

Automatic video description has many applications in human-robot interaction, automatic video subtitling, and video surveillance. It can be used to help the visually impaired by generating verbal descriptions of surroundings through speech synthesis or automatically generate and read out film descriptions. Currently, these are achieved through very costly and time-consuming manual processes. Another application is the description of sign-language videos in natural language. Video description can also generate written procedures for human or service robots by automatically converting actions in a demonstration video into simple instructions; for example, assembling furniture, installing a CD-ROM, making coffee, or changing a flat tire [4, 20].

The advancement of video description opens up enormous opportunities in many application domains. It is envisaged that in the near future, we will be able to interact with robots in the same manner as with humans [119]. 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, then *Service Robots* or *Smart phone Apps* will be able to understand human actions and other events to converse with humans in a much more meaningful and coherent manner. For example, 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. The recent release of a dialogue dataset, *Talk the Walk* [149], has introduced yet another interesting application where a natural language dialogue between a *guide* and a *tourist* helps the tourist to reach a previously unseen location on a map using perception, action, and interaction modeling.

Leveraging the recent developments in deep neural networks for NLP and CV, and the increased availability of large multi-modal datasets, automatically generating stories from pixels is no longer 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* [121]. Related research areas include connecting words to pictures [15, 16, 31], narrating images in natural language sentences [38, 74, 80], and understanding natural language instructions for robotic applications [50, 90, 136]. Another closely related field is *Visual Information Retrieval (VIR)*, which takes visual (image, drawing, or sketch), text (tags, keywords, or complete sentences), or mixed visual and text queries to perform content-based search. Thanks to the release of benchmark datasets MS COCO [83] and Flicker30k [164], research in *image captioning and retrieval* [33, 37, 66, 88], and *image question answering* [9, 87, 111, 168] has also become very active.

Automatically generating natural language sentences describing the video content has two components: *understanding* the visual content and *describing* it in grammatically correct natural language sentences. Figure 1 shows a simple deep learning–based video captioning framework. The task of video description is relatively more challenging, compared to image captioning, because not all objects in the video are relevant to the description—such as the detected objects that do not

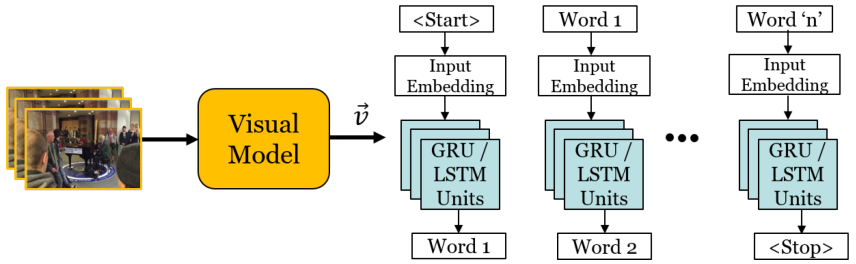

Fig. 1. A basic framework for deep learning–based video captioning. A visual model encodes the video frames into a vector space. The language model takes input of visual vector and word embeddings to generate the sentence that describes the input visual content.

play any role in the observed activity [14]. Moreover, video description methods must additionally capture the speed, direction of relevant objects, as well as causality among events, actions, and objects. Finally, events in videos can be of varying lengths and may even result in a possible overlap of events [70]. See Figure 2, for example: The event of piano recitals is spanned over almost the entire duration of the video, however, the applause is a very short event that only takes place at the end. The example illustrates differences between three related areas of research—namely, image captioning, video captioning, and dense video captioning. In this example, image captioning techniques recognize the event as mere *clapping*, whereas it is actually *applause* that resulted from a previous event—piano playing.

Figure 3 summarizes related research under the umbrella of *Visual 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 autotranscription where audio and speeches are the main focus. Video captioning concerns mainly the visual content as opposed to the audio signals. In particular, *Video Description* extends video captioning with the aim to provide a more detailed account of the visual contents in the video.

Below, we define some terminology used in this article.

- *Visual Description*: The unifying concept encompassing (see Figure 3) the automatic generation of single or multiple natural language sentences that convey the information in still images or video clips.
- *Video Captioning*: Conveying the information of a video clip as a whole through a single automatically generated natural language sentence based on the premise that short video clips usually contain one main event [11, 33, 43, 101, 144, 162].
- *Video Description*: Automatically generating multiple natural language sentences that provide a narrative of a relatively longer video clip. The descriptions are more detailed and may be in the form of paragraphs. Video description is sometimes also referred to as *story -telling* or *paragraph generation* [114, 167].
- *Dense Video Captioning*: Detection and conveying information of all, possibly overlapping, events of different lengths in a video using a natural language sentence per event. As illustrated in Figure 2, dense video captioning localizes events in time [70, 107, 158, 163] and generates sentences that are not necessarily coherent. However, video description gives a more detailed account of one or more events in a video clip using multiple coherent sentences without having to localize individual events.

Video captioning research started with the classical template-based approaches in which Subject (S), Verb (V), and Object (O) are detected separately and then joined using a sentence

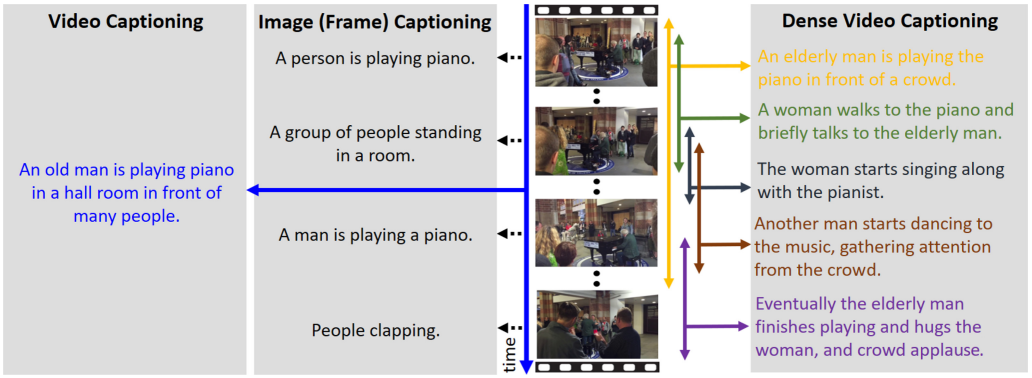

Fig. 2. Illustration of differences between image captioning, video captioning, and dense video captioning. Image (video frame) captioning describes each frame with a single sentence. Video captioning describes the complete video with one sentence. In dense video captioning, each event is temporally detected and described by a single sentence eventually resulting in multiple sentences localized in time but not necessarily coherent.

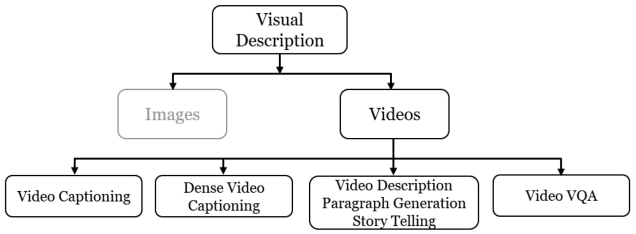

Fig. 3. Classification of visual content description. This survey focuses on video only and not images.

template. These approaches are referred to as *SVO-Triplets* [14, 68]. However, the advent of deep learning and the tremendous advancements in CV and NLP have equally affected the area of video captioning. Hence, latest approaches follow deep learning–based architectures [117, 144] that encode the visual features with 2D/3D-CNN and use LSTM/GRU to learn the sequence. The output of both approaches is either a single sentence [100, 160] or multiple sentences [14, 29, 62, 114, 129, 167] per video clip. Early research on video description mostly focused on domain-specific short video clips with limited vocabularies of objects and activities [14, 29, 61, 68, 119, 165]. Description of open domain and relatively longer videos remains a challenge, as it needs large vocabularies and training data. Methods that follow CNN-LSTM/GRU framework mainly differ from each other in the different types of CNNs and language models (vanilla RNN, LSTM, and GRUs) they employ as well as how they pass the extracted visual features to the language model (at the first time-step only or all time-steps). Later methods progressed by introducing additional transformations on top of the standard encoder-decoder framework. These transformations include attention mechanism [162], where the model learns which part of the video to focus on; sequence learning [144], which models a sequence of video frames with the sequence of words in the corresponding sentence; semantic attributes [43, 101], which exploit the visual semantics in addition to CNN features, and joint modeling of visual content with compositional text [100]. More recently, video-based visual description problem has evolved towards dense video captioning and video story-telling. New datasets have also been introduced to progress along these lines.

When it comes to performance comparison, quantitative evaluation of video description systems is not straightforward. Currently, automatic evaluations are typically performed using machine

translation and image captioning metrics, including Bilingual Evaluation Understudy (BLEU) [102], Recall Oriented Understudy for Gisting Evaluation (ROUGE) [82], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [12], Consensus-based Image Description Evaluation (CIDEr) [142], and the recently proposed Semantic Propositional Image Captioning Evaluation (SPICE) [5] and Word Mover's Distance (WMD) [76] metrics. Section 4.1 presents these measures. Here, we give a brief overview to establish motivation for our survey. BLEU is a precision-based metric, which accounts for precise matching of *n-grams* in the generated and ground truth references. METEOR, however, first creates an alignment between the 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 was developed specifically for evaluating image captioning methods but has also been used in video description tasks. WMD makes use of word embeddings (semantically meaningful vector representations of words) and compares two texts using the Earth Mover's Distance (EMD). This metric is relatively less sensitive to word order and synonym changes in a sentence and, like CIDEr and METEOR, it provides high correlation with human judgments. Last, SPICE is a more recent metric that correlates more with human judgment of semantic quality as compared to previously reported metrics. It compares the semantic information of two sentences by matching their content in dependency parse trees. These metrics capture very different performance measures for the same method and are not perfectly aligned with human judgments. Also, due to the hand-engineered nature of these metrics, their scores are unstable when the candidate sentence is perturbed with synonyms, word order, length, and redundancy. Hence, there is a need for an evaluation metric that is *learned* from training data to score in harmony with human judgments in describing videos with diverse content.

The current literature lacks a comprehensive and systematic survey that covers different aspects of video description research, including methods, dataset characteristics, evaluation measures, benchmark results and related competitions, and video Q&A challenges. We fill this gap and present a comprehensive survey of the literature. We first highlight the important applications and major trends of video description in Section 1 and then classify automatic video description methods into three groups, giving an overview of the models from each group in Section 2. In Section 3, we elaborate on the available video description datasets used for benchmarking. Furthermore, we review the evaluation metrics that are used for quantitative analysis of the generated descriptions in Section 4. In Section 5, benchmark results achieved through the aforementioned methods are compared and discussed. In Section 6, we discuss the possible future directions. Section 7 concludes our survey and discusses some insights into the findings.

2 VIDEO DESCRIPTION METHODS

Video description literature can be divided into three main phases: The classical methods phase, where pioneering visual description research employed classical CV and NLP methods to first detect entities (objects, actions, scenes) in videos and then fit them to standard sentence templates. The statistical methods phase, which employed statistical methods to deal with relatively larger datasets. This phase lasted for a relatively short time. Finally, the deep learning phase, which is the current state-of-the-art and is believed to have the potential to solve the open domain automatic video description problem. Below, we give a detailed survey of the methods in each category.

2.1 Classical Methods

The SVO (Subject, Object, Verb) tuples-based methods are among the first successful methods used specifically for video description. However, research efforts were made long before to describe

visual content into natural language, albeit not explicitly for captioning or description. The first-ever attempt goes back to Koller et al. [69] in 1991, who developed a system that was able to characterize motion of vehicles in real traffic scenes using natural language verbs. Later, in 1997, Brand et al. [20] dubbed this as "Inverse Hollywood Problem" (since in Hollywood script (description) is converted into video; here, the problem is opposite), and described a series of actions into semantic tag summaries to develop a storyboard from instructional videos. They also developed a system, "video gister," which was able to heuristically parse the videos into a series of key actions and generate a script that describes actions detected in the video. They also generated key frames depicting the detected causal events and defined the series of events into semantics representation, e.g., `Add by enter`, `motion`, `detach and remove by attach, move`, `leave.` Video gister was limited to only one human arm (actor) interacting with non-liquid objects and was able to understand only five actions (touch, put, get, add, remove).

Getting back to SVO tuple-based methods, which tackle 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 or *actor*, 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 fills 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. Numerous methods 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.

- *Object Recognition:* Object recognition in SVO approaches was performed typically using conventional methods, including model-based shape matching through edge detection or color matching [68], HAAR features matching [148], context-based object recognition [140], Scale Invariant Feature Transform (SIFT) [85], discriminatively trained part-based models [42], and Deformable Parts Model (DPM) [40, 41].
- *Human and Activity Detection:* Human detection methods employed features such as Histograms of Oriented Gradient (HOG) [27] followed by SVM. For activity detection, features like Spatiotemporal Interest Points such as Histogram of Oriented Optical Flow (HOOF) [21], Bayesian Networks [56], Dynamic Bayesian Networks [46], Hidden Markov Models (HMM) [17], state machines [69], and PNF Networks [105] have been used by SVO approaches.
- *Integrated Approaches:* Instead of detecting the description-relevant entities separately, Stochastic Attribute Image Grammar (SAIG) [176] and Stochastic Context Free Grammars (SCFG) [94], allow for compositional representation of visual entities present in a video, an image, or a scene based on their spatial and functional relations. Using the visual grammar, the content of an image is first extracted as a parse graph. A parsing algorithm is then used to find the best scoring entities that describe the video. In other words, not all entities present in a video are of equal relevance, which is a distinct feature of this class of methods compared to the aforementioned approaches.

For Stage II, sentence generation, a variety of methods have been proposed, including HALogen representation [77], Head-driven Phrase Structure Grammar (HPSG) [106], planner and surface realizer [110]. The primary common task of these methods is to define templates. A template is a user-defined language structure containing placeholders. To function properly, a template is composed of three parts named lexicons, grammar, and template rules. *Lexicon* represents vocabulary that describes high-level video features. *Template rules* are user-defined rules guiding the selection

Subject + Verb	**Subject + Verb + Object**
Woman is walking.	Man is smoking a cigarette.
A man is standing.	A man is drinking coffee.
Subject + Verb + Object + Place	**Subject + Verb + Complement**
A woman is cooking in the kitchen.	Man looks tired.
A boy is playing on the beach.	Woman is old

Fig. 4. An example of various templates used for sentence generation from videos. Subject, verb, and object are used to fill in these templates. 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.

of appropriate lexicons for sentence generation. *Grammar* defines linguistic rules to describe the structure of expressions in a language, ensuring that a generated sentence is syntactically correct. Using production rules, Grammar can generate a large number of various configurations from a relatively small vocabulary.

In template-based approaches, 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 4 presents examples of some popular templates used for sentence generation in template-based approaches. Figure 5 gives a timeline of how the classical methods evolved over time; whereas below, we provide a survey of SVO methods by grouping them into three categories—namely, subject (human) focused, action and object focused, and methods that use the SVO approach on open domain videos. Note that the division boundaries are frequently blurred between these categories.

(1) Subject (Human) Focused: In 2002, Kojima et al. [68] proposed one of the earliest methods designed specifically for video captioning. This method focuses primarily on describing videos of one person performing one action only. To detect humans in a scene, they calculated the probability of a pixel coming from the background or the skin region using the values and distributions of pixel chromaticity. Once a human's head and hands are detected, the human posture is estimated by considering three kinds of geometric information, i.e., position of the head and hands and direction of the head. 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 up. To generate the description in sentences, pre-defined *case frames* and verb patterns as proposed by Nishida et al. [96, 97] are used. Case frame is a type of frame expression used for representing the relationship between cases, which are classified into eight 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 organized 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 (e.g., walk) from instantaneous actions (e.g., stand up). The major drawback of their approach is that it cannot be easily extended to more complex scenarios such as multiple actors, incorporating temporal information, and capturing causal relationship between

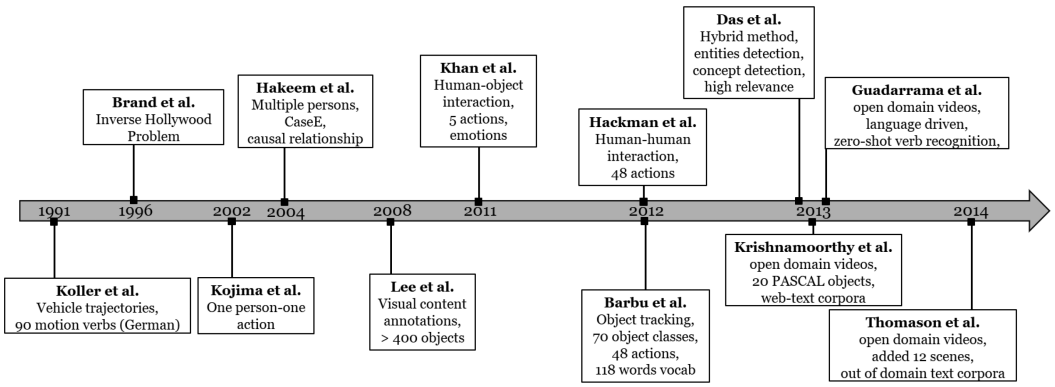

Fig. 5. Evolution of classical methods over time. In general the focus of these methods moved from subjects (humans) to actions and objects and then to open domain videos containing all three SVO categories.

events. The heavy reliance on the correctness of manually created activity concept hierarchy and state transition model also prevents it from being used in practical situations.

Hakeem et al. [51] addressed the shortcomings of Kojima et al's [68] work and proposed an extended case framework (CASE^E) using hierarchical CASE representations. They incorporated multiple agent events, temporal information, and causal relationship between the events to describe the events in natural language. They introduced case-list to incorporate multiple agents in AG, `[PRED:move, AG:{person1, person2},...]`. Moreover, they incorporated temporal information into CASE using temporal logic to encode the relationship between sub-events. As some events are conditional on other events, they also captured causal relationship between events. For example, in the sentence "a man played piano and the crowd applauded," the applause occurred because the piano was played. `[CAUSE: [PRED:play, D:crowd, FAC:applaud]]`.

Khan et al. [62] introduced a framework to describe human-related contents such as actions (limited to five only) and emotions in videos using natural language sentences. They implemented a suite of conventional image processing techniques, including face detection [73], emotion detection [86], action detection [17], non-human object detection [148], and scene classification [65], to extract the high-level entities of interest from video frames. These include humans, objects, actions, gender, position, and emotion. Since their approach encapsulates human-related actions, human is rendered as *Subject* and the objects upon which action is performed 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, each spanning 5s to 20s 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 involving a human interacting with some objects. Hence, their method does not generate any description until a human is detected in the video. The method cannot identify actions with subtle movements (such as smoking and drinking) and interactions among humans.

(2) Action and Object Focused: Lee et al. [78] proposed a method for semantically annotating visual content in three sequential stages: namely, image parsing, event inference, and language generation. An "image parsing engine" using stochastic attribute image grammar (SAIG) [176] is employed to produce a visual vocabulary, i.e., a list of visual entities present in the frame along with their relationships. This output is then fed into an "event inference engine," which extracts semantic and contextual information of visual events along with their relationships. Video Event Markup Language (VEML) [95] is used to represent semantic information. In the final stage, head-driven phrase structure grammar (HPSG) [106] is used to generate text description from the semantic

representation. Compared to Kojima et al. [68], grammar-based methods can infer and annotate a wider range of scenes and events. Ten streams of urban traffic and maritime scenes over a period of 120mins, containing more than 400 moving objects, are used for evaluation. Some detected events include "entering the scene, moving, stopping, turning, approaching traffic intersection, watercraft approaching maritime markers and land areas and scenarios where one object follows the other" [78]. Recall and Precision rates are employed to evaluate the accuracy of the events that are detected with respect to manually labeled ground truth. Due to poor estimation of the motion direction from a few perspective views, their method does not perform well on "turning" events.

Hanckmann et al. [52] proposed a method to automatically describe events involving multiple actions (seven on average) performed by one or more individuals. Unlike Khan et al. [62], human-human interactions are taken into account in addition to human-object interactions. Bag-of-features (48 in total) are collected as action detectors [18] for detecting and classifying actions in a video. The description generator subsequently describes the verbs relating the actions to the scene entities. It finds the appropriate actors among objects or persons and connects them to the appropriate verbs. In contrast to Khan et al. [62], who assume that the subject is always a person, Hanckmann et al. [52] generalizes subjects to include vehicles as well. Furthermore, the number of human actions is much richer. Compared to the five verbs in Khan et al. [62], they have 48 verbs capturing a diverse range of actions, such as `approach`, `arrive`, `bounce`, `carry`, `catch`, and so on.

Barbu et al. [14] 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 a detected object and action as noun and verb, respectively. Adjectives are used for the object properties, and prepositions are used for their spatial relationships. Their approach is composed of three steps. In the first step, object detection [41] is carried out on each frame by limiting 12 detections per frame. Second, object tracking [128, 138] is performed to increase the precision. Third, using dynamic programming, the optimal set of detections is chosen. Verb labels corresponding to actions in the videos are then produced using Hidden Markov Models. 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. To incorporate all the relevant concepts, these methods require customized detectors for each entity. Furthermore, the texts generated by existing methods of the time have mostly 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. [29] proposed a hybrid method that produces content of high relevance compared to simple keyword annotation methods. They borrowed ideas from image captioning techniques. This hybrid model is composed of three steps in a hierarchical manner. First, in a bottom-up approach, keywords are predicted using low-level video features. In this approach, they first find a proposal distribution over the training set of 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. A tripartite graph template is then used for converting the stitched concepts to a natural language description. Finally, for semantic verification, they produced a ranked set of natural language sentences by comparing the predicted 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 [14, 61].

(3) SVO Methods for Open Domain Videos: While most of the prior mentioned works are restricted to constrained domains, Krishnamoorthy et al. [71] led the early works of describing

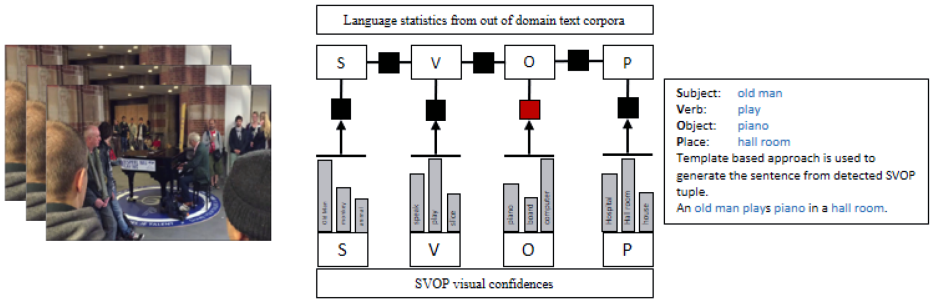

Fig. 6. Example of the Subject-Verb-Object-Place (SVOP) [137] approach where confidences are obtained by integrating probabilities from visual recognition system, with statistics from out-of-domain English text corpora to determine the most likely SVOP tuple. The red block shows low probability given to a correct object by the visual system that is rectified by the high probability from the linguistic model.

open domain videos. 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 corpora to aid the selection of the best SVO tuple to improve sentence coherence.

In addition to focusing on open domain videos and utilizing web-scaled text corpora, Guadarrama et al. [49] and Thomason et al. [137] started dealing with relatively larger vocabularies. Compared to Krishnamoorthy et al. [71], instead of using only 20 objects in the PASCAL dataset [36], 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. [49] proposed a novel language-driven approach. They introduced "zero-shot" verb recognition for selecting unseen verbs in the training set. For example, if subject is "person," object refers to "car" and the model-predicted verb is "move," then the most suitable verb would be "drive." Thomason et al. [137] used visual recognition techniques on YouTube videos for probabilistic estimations of subjects, verbs, and objects. Their approach is illustrated in Figure 6. The object and action classifiers were trained on ImageNet [124]. In addition to detecting subjects, verbs, and objects, places (12 scenes) where actions are performed, e.g., kitchen or playground, 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 composed of human-annotated sentences for the video description dataset is also constructed and incorporated into 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" corpora, 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.

Classical methods focused mainly on the detection of pre-defined entities and events separately. These methods then tried to describe the detected entities and events using template-based sentences. However, to describe open domain videos or those with more events and entities, classical methods must employ object and action detection techniques for each entity, which is unrealistic due to the computational complexity. Moreover, template-based descriptions are insufficient to describe all possible events in videos given the linguistic complexity and diversity. Consequently, these methods failed to describe semantically rich videos.

2.2 Statistical Methods

Naïve SVO tuple rule-based engineering approaches are indeed inadequate to describe open domain videos and large datasets, such as YouTubeClips [23], TACoS-MultiLevel [114], MPII-MD [116], and M-VAD [139]. These datasets contain very large vocabularies as well as tens of hours of videos. There are three important differences between these open domain and previous datasets. First, open domain videos contain unforeseeable diverse sets of subjects, objects, activities, and places. Second, due to the sophisticated nature of human languages, such datasets are often annotated with multiple viable meaningful descriptions. Third, the videos to be described are often long, potentially stretching through many hours. Descriptions of such videos with multiple sentences or even paragraphs become more desirable.

To avoid the tedious efforts required in rule-based engineering methods, Rohrbach et al. [119] proposed a machine learning method to convert visual content into natural language. They used parallel corpora of videos and associated annotations. Their method follows a two-step approach. First, it learns to represent the video as intermediate semantic labels using maximum posterior estimate (MAP). Then, it translates the semantic labels into natural language sentences by using techniques borrowed from Statistical Machine Translation (SMT) [67]. In this machine translation approach, the intermediate semantic label representation is the source, while the expected annotations are regarded as the target language.

For the object and activity recognition stages, the research moved from earlier threshold-based detection [68] to manual feature engineering and traditional classifiers [29, 49, 71, 137]. For the sentence generation stage, an uptake of machine learning methods can be observed in recent years to address the issue of large vocabulary. This is also evidenced by the trend in recent methods that use models for lexical entries that are learned in a weakly supervised [114, 119, 161, 166] or fully supervised [26, 49, 71, 133] 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 and discuss how they address the scalability, language complexity, and domain transferability issues faced by open domain video description.

2.3 Deep Learning Models

The whirlwind success of deep learning in almost all sub-fields of computer vision has also revolutionized video description approaches. In particular, Convolutional Neural Networks (CNNs) [72] are the state-of-the-art for modeling visual data and excel at tasks such as object recognition [72, 131, 135]. Long Short-Term Memory (LSTMs) [55] and the more general deep Recurrent Neural Networks (RNNs), however, are now dominating the area of sequence modeling, setting new benchmarks in machine translation [25, 134], speech recognition [47], and the closely related task of image captioning [33, 147]. 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 7, deep learning approaches to video description can also be divided into two sequential stages: visual content extraction and text generation. However, in contrast to the SVO tuple methods (Section 2.1), where lexical word tokens are generated as a result of the first stage through visual content extraction, visual features represented by fixed or dynamic real-valued vectors are produced instead. 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 visual features that are then used in the second stage for text generation, also known as the *decoding stage*. For decoding, different flavors of RNNs are used, such as deep RNN, Bi-directional RNN, LSTM, or Gated

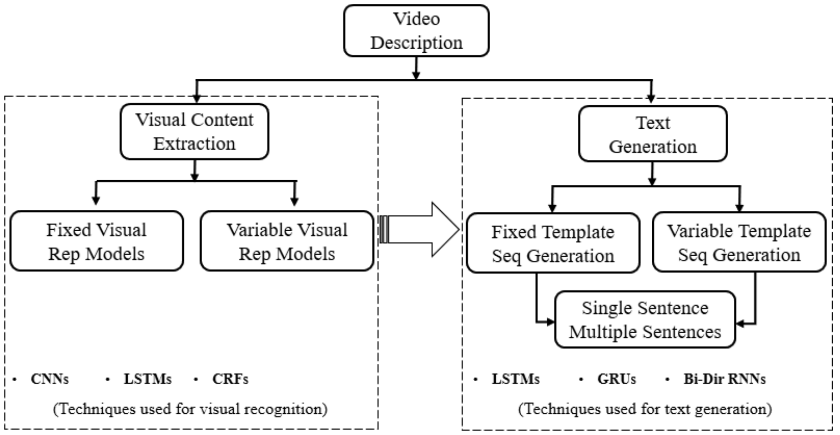

Fig. 7. Deep learning–based video description techniques in the literature are composed of two main stages. The first stage involves visual content extraction and is represented either by a fixed length vector or by dynamic vectors. The second stage takes input of visual representation vectors from the first stage for text generation and generates single/multiple sentence(s).

Recurrent Units (GRU). The resulting description can be a single sentence or multiple sentences. Figure 8 illustrates a typical end-to-end video description system with encoder-decoder stages. The encoding part is followed by transformations such as mean pooling, temporal encoding, or attention mechanisms to represent the visual content. Some methods apply sequence-to-sequence learning and/or semantic attributes learning. The aforementioned mechanisms have been used in different combinations by contemporary methods. We group the literature based on the different combinations of deep learning architectures for encoding and decoding stages, namely:

- CNN-RNN Video Description, where convolution architectures are used for visual encoding and recurrent structures are used for decoding. This is the most common architecture employed in deep learning–based video description methods;
- RNN-RNN Video Description, where recurrent networks are used for both stages; and
- Deep reinforcement networks, the relatively new research area for video description.

2.3.1 CNN-RNN Video Description. Given its success in computer vision and simplicity, 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. [33] were the first to use deep neural networks to solve the video captioning problem. They proposed three architectures for video description. Their model is based on the assumption to have CRF-based predictions of subjects, objects, and verbs after full pass of complete video. This allows the architecture to observe the complete video at each time-step. 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. [119] mentioned earlier in Section 2.2. Recognizing the state-of-the-art machine translation performance of LSTMs, the SMT module in Reference [119] is replaced with a stacked LSTM composed of two layers for encoding and decoding. Similar to Reference [134], 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 to generate a sentence by producing one word per time-step. Another variant of the architecture, LSTM decoder with CRF max, incorporates max predictions. This architecture encodes the semantic representation into a fixed length vector. Similar

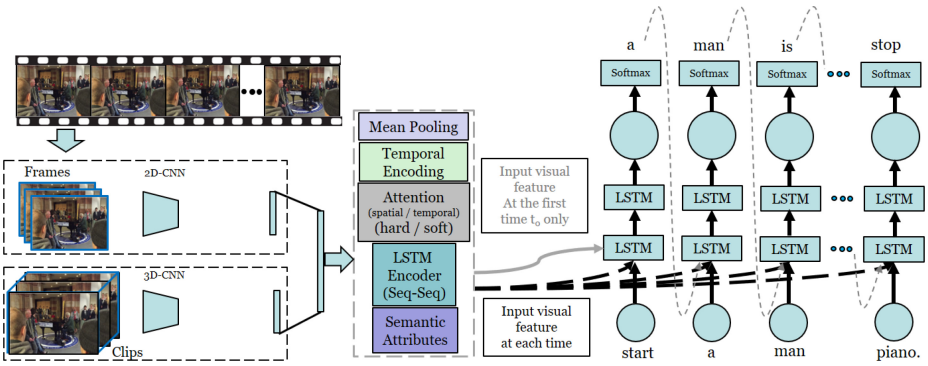

Fig. 8. Summary of deep learning–based video description methods. Most methods employ mean pooling of frame representations to represent a video. More advanced methods use attention mechanisms, semantic attribute learning, and/or employ a sequence-to-sequence approach. These methods differ in whether the visual features are fed only at first time-step or all time-steps of the language model.

to image description, LSTM is able to see the whole visual content at every time-step. An advantage of LSTM is that it is able to incorporate probability vectors during training as well as testing. 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 the second variant (LSTM decoder with CRF max), this architecture incorporates probability distributions. Although the LSTM outperformed the SMT-based approach of Reference [119], it was still not trainable in an end-to-end fashion.

In contrast to the work by Donahue et al. [33], where an intermediate role representation was adopted, Venugopalan et al. [145] presented the first end-to-end trainable network architecture for generating natural language description of videos. Their model is able to simultaneously learn the semantic as well as grammatical structure of the associated language. Moreover, Donahue et al. [33] presented results on *domain -specific* cooking videos composed of pre-defined objects and actors. However, Venugopalan et al. [145] reported results on `open domain` YouTube Clips [22]. To avoid supervised intermediate representations, they connected an LSTM directly to the output of the CNN. The CNN extracts visual features, whereas the LSTM models the sequence dynamics. They transformed a short video into a fixed length visual input using a CNN model [58] that is slightly different from AlexNet [72]. The CNN model [58] was learned using the ILSVRC-2012 object classification dataset (composed of 1.2M images), which is a subset of ImageNet [124]. It provides a robust and efficient way without manual feature selection for initialization object recognition in the videos. They sampled every tenth frame in the video and extracted features for all sample frames from the *fc7* layer of the CNN. Furthermore, they represented a complete video by averaging all the extracted frame-wise feature vectors into a single vector. These feature vectors are then fed into a two-layered LSTM [48]. The feature vectors from CNN form the input to the first layer of the LSTM. A second LSTM layer is stacked on top of the first LSTM layer, where the hidden state of the first LSTM layer is the input to the second 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 performed better than the previous systems at the time and was able to effectively generate the sequence without any templates. However, as a result of simple averaging, valuable temporal information of the video, such as the order of appearances of any two objects, is lost. Therefore, this approach is only suitable for generating captions for short clips with a single major action per clip.

Open domain videos are rich in complex interactions among actors and objects. Representation of such videos using a temporally averaged 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 is not captured in the representation. With the success of C3D [141] in capturing spatio-temporal action dynamics in videos, Li et al. [162] proposed a novel 3D-CNN to model the spatio-temporal information in videos. Their 3D-CNN is based on GoogLeNet [135] and pre-trained on an activity recognition dataset. It captures local fine motion information between consecutive frames. This local motion information is then subsequently summarized and preserved through higher-level representations by modeling a video as a 3D spatio-temporal cuboid. It is further represented by concatenation of HoG, HoF, MbH [28, 151]. 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 [10]. Using 3D CNN and attention mechanisms in RNN, they were able to improve results. Recently, GRU-EVE [3] was proposed as an effective and computationally efficient technique for video captioning. GRU-EVE uses a standard GRU for language modeling but with Enriched Visual Encoding as follows: It applies the Short Fourier Transform on 2D/3D-CNN features in a hierarchical manner to encapsulate the spatio-temporal video dynamics. The visual features are further enriched with high-level semantics of the detected objects and actions in the video. Interestingly, the enriched features obtained by applying Short Fourier Transform on 2D-CNN features alone [3] outperform C3D [141] features.

Unlike the *fixed video representation models* discussed above, *variable visual representation models* are able to directly map input videos composed of different number of frames to variable-length words or sentences (outputs), and are successful in modeling various complex temporal dynamics. Venugopalan et al. [144] proposed an architecture to address the variable representation problem for both the input (video frames) and the output (sentence) stage. For that purpose, they used a two-layered LSTM framework, where the sequence of video frames is input to the first layer of the LSTM. The hidden state of the first LSTM layer forms the input to the second layer of the LSTM. The output of the second LSTM layer is the associated caption. The LSTM parameters are shared in both stages. Although sequence-to-sequence learning had previously been used in machine translation [134], this is the first method [144] to use a sequence-to-sequence approach in video captioning. Later methods have adopted a similar framework, with minor variations including attention mechanisms [162], making a common visual-semantic-embedding [100], or using out-of-domain knowledge either with language models [143] or visual classifiers [115].

While deep learning has achieved much better results compared to previously used classifier-based approaches, most methods have aimed at producing one sentence from a video clip containing only one major event. In real-world applications, videos generally contain more than a single event. Description of such multi-events and semantically rich videos by only one sentence ends up to be overly simplified, and hence, uninformative. For example, instead of saying "someone sliced the potatoes with a 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. [167] proposed a hierarchical recurrent neural network (h-RNN) that applies the attention mechanisms on both the temporal and spatial aspects. They focused on the sentence decoder and introduced a hierarchical framework composed of a sentence generator and on top of that a paragraph generator. First, a Gated Recurrent Unit (GRU) layer takes video features as input and generates a single short sentence. The other recurrent layer generates paragraphs using context and the sentence vectors obtained from the sentence generator. The paragraph generator thus captures the dependencies between sentences and generates a paragraph of sentences that are related. Recently, Krishna et al. [70] introduced the concept of dense-captioning of events in a video and

employed action-detection techniques to predict the temporal intervals. They proposed a model to extract multiple events with one single pass of a video, attempting to describe the detected events simultaneously. This is the first work that detects and describes multiple overlapping events in a video. However, the model did not achieve significant improvement on the captioning benchmark.

2.3.2 RNN-RNN Video Description. Although not as popular as the CNN-RNN framework, another approach is to also encode the visual information using RNNs. Srivastava et al. [132] use one LSTM to extract features from video frames (i.e., encoding) and then pass the feature vector through another LSTM for decoding. They also introduced some variants of their models and predicted the future sequences from the previous frames. The authors adopted a machine translation model [134] for visual recognition but could not achieve significant improvement in classification accuracy.

Yu et al. [167] proposed a similar approach and used two RNN structures for the video description task. Their configuration is a hierarchical decoder with multiple Gated Recurrent Units (GRU) for sentence generation. The output of this decoder is then fed to a paragraph generator that models the time dependencies between the sentences while focusing on linguistic aspects. The authors improved the state-of-the-art results for video description; however, their method is inefficient for videos involving fine-grained activities and small interactive objects.

2.3.3 Deep Reinforcement Learning Models. Deep Reinforcement Learning (DRL) has outperformed humans in many real-word games. In DRL, artificial intelligence agents learn from the environment through trial-and-error and adjust learning policies purely from environmental rewards or punishments. DRL approaches have been popularized by Google Deep Mind [92, 93] since 2013. Due to the absence of a straightforward 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 algorithms 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 [75, 146]. Similarly, different variants of DRL have been used to meet the challenging requirements of image captioning [112] as well as video description [24, 79, 103, 104, 155].

Xwang et al. [155] proposed a fully differentiable neural network architecture using reinforcement learning for video description. Their method follows a general encoder-decoder framework. The encoding stage captures the video frame features using ResNet-152 [54]. The frame-level features are processed through two-stage encoder, i.e., low-level LSTM [125] followed by a high-level LSTM [55]. For decoding, they employed HRL to generate the word-by-word natural language descriptions. The HRL agent is composed of three components, a low-level worker that accomplishes tasks as set by manager, a high-level manager that sets goals, and internal critic to ascertain whether the task has been accomplished or not and informs the manager accordingly to help the manager update the goals. The process iterates till reaching the end of sentence token. This method is demonstrated to be capable of capturing more details of the video content, thus generating more fine-grained descriptions. However, this method has shown very little improvement over existing baseline methods.

In 2018, Chen et al. [24] proposed an RL-based model selecting *key informative frames* to represent a complete video in an attempt to minimize noise and unnecessary computations. Key frames are selected such that they maximize visual diversity and minimize the textual discrepancy. Hence, a compact subset of 6–8 frames on average can represent a full video. Evaluated against several

Table 1. Standard Datasets for Benchmarking Video Description Methods

Dataset	Domain	# classes	# videos	avg len	# clips	# sent	# words	vocab	len (hrs)
MSVD [22]	open	218	1970	10s	1,970	70,028	607,339	13,010	5.3
MPII Cooking [118]	cooking	65	44	600s	-	5,609	-	-	8.0
YouCook [29]	cooking	6	88	-	-	2,688	42,457	2,711	2.3
TACoS [109]	cooking	26	127	360s	7,206	18,227	146,771	28,292	15.9
TACos-MLevel [114]	cooking	1	185	360s	14,105	52,593	2K	-	27.1
MPII-MD [116]	movie	-	94	3.9s	68,337	68,375	653,467	24,549	73.6
M-VAD [139]	movie	-	92	6.2s	48,986	55,904	519,933	17,609	84.6
MSR-VTT [160]	open	20	7,180	20s	10K	200K	1,856,523	29,316	41.2
Charades [130]	human	157	9,848	30s	-	27,847	-	-	82.01
VTW [171]	open	-	18,100	90s	-	44,613	-	-	213.2
YouCook II [174]	cooking	89	2K	316s	15.4K	15.4K	-	2,600	176.0
ActyNet Cap [70]	open	-	20K	180s	-	100K	1,348,000	-	849.0
ANet-Entities [173]	social media	-	14,281	180s	52K	-	-	-	-
VideoStory [45]	social media	-	20K	-	123K	123K	-	-	396.0

popular benchmarks, it was demonstrated that video captions can be produced without performance degradation but at a significantly reduced computational cost. The method did not use motion features for encoding, a design trade-off between speed and accuracy. DRL-based methods are gaining popularity and have shown comparable results in video description. Due to their unconventional learning methodology, DRL methods are unlikely to suffer from paucity of labeled training data, hardware constraints, and overfitting problems. Therefore, these methods are expected to flourish.

3 DATASETS

The availability of labeled datasets for video description has been the main driving force behind the fast advancement of this research area. In this survey, we summarize the characteristics of these datasets and give an overview in Table 1. The datasets are categorized into four main classes, namely: *Cooking*, *Movies*, *Videos in the Wild*, and *Social Media*. In most of the datasets, a single caption per video is assigned except for a few datasets that contain multiple sentences or even paragraphs per video snippet.

3.1 Cooking

3.1.1 MP-II Cooking. Max Plank Institute for Informatics (MP-II) Cooking dataset [118] is composed of 65 fine-grained cooking activities, performed by 12 participants preparing 14 dishes such as *fruit salad, cake*, and so on. The data are recorded in the same kitchen with camera installed on the ceiling. The 65 cooking activities include "wash hands," "put in bowl," "cut apart," "take out from drawer," and so on. When the person is not in the scene for 30 frames (1s) or is performing an activity that is not annotated, a "background activity" is generated. These fine-grained activities—for example "cut slices," "pour," or "spice"—are differentiated by movements with low inter-class and high intra-class variability. In total, the dataset is composed of 44 videos (888,775 frames), with an average length per clip of approximately 600s. The dataset spans a total of 8h play length for all videos and 5,609 annotations.

3.1.2 YouCook. The YouCook dataset [29] consists of 88 YouTube cooking videos of different people cooking various recipes. The background (kitchen/scene) is different in most of the videos.

This dataset represents a more challenging visual problem than the MP-II Cooking [118] dataset that is recorded with a fixed camera viewpoint in the same kitchen and with the same background. The dataset is divided into six different cooking styles; for example, *grilling*, *baking*, and so on. For machine learning, the training set contains 49 videos and the test set contains 39 videos. Frame-wise annotations of objects and actions are also provided for the training videos. The object categories for the dataset include "utensils," "bowls," "food," and so on. 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 eight descriptions were collected per video.

3.1.3 TACoS. Textually Annotated Cooking Scenes (TACoS) is a subset of MP-II Composites [120]. 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 [118]. It contains 212 high-resolution videos with 41 cooking activities. Videos in the MP-II Composites dataset span over different lengths ranging from 1–23mins with an average length of 4.5mins. 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 to align the sentences and associated videos; for example, "preparing carrots," "cutting a cucumber," "separating eggs," and so on. For each video, 20 different textual descriptions were collected. The dataset is composed of 11,796 sentences containing 17,334 actions descriptions. A total of 146,771 words are used in the dataset. Almost 50% of the words, i.e., 75,210, describe the content for example nouns, verbs, adjectives, and so on. These words includes a vocabulary size of 28,292 verb tokens. The dataset also provides the alignment of sentences describing activities by obtaining approximate time stamps where each activity starts and ends.

3.1.4 TACoS-MultiLevel. TACoS Multilevel [114] corpus annotations were also collected via AMT workers on the TACoS corpus [109]. For each video in the TACoS corpus, three levels of descriptions were collected, which include: (1) detailed description of video with no more than 15 sentences per video; (2) a short description composed of 3–5 sentences per video; and finally (3) a single sentence description of the video. Annotation of the data is provided in the form of tuples such as object, activity, tool, source, and target with a person always being the subject.

3.1.5 YouCook II. YouCook-II Dataset [174] consists of 2K videos uniformly distributed over 89 recipes. The cooking videos are sourced from YouTube and offer all the challenges of open domain videos, such as variations in camera position, camera motion, and changing backgrounds. The complete dataset spans a total play time of 175.6h and has a vocabulary of 2600 words. The videos are further divided into 3–16 segments per video with an average of 7.7 segments per video elaborating procedural steps. Individual segment length varies from 1s to 264s. All segments are temporally localized and annotated. The average length of each video is 316s, reaching up to a maximum of 600s. The dataset is randomly split into train, validation, and test sets with the ratio of 66%:23%:10%, respectively.

3.2 Movies

3.2.1 MPII-MD. MPII-Movie Description Corpus [116] contains transcribed audio descriptions extracted from 94 Hollywood movies. These movies are subdivided into 68,337 clips with an average length of 3.9s paired with 68,375 sentences amounting to almost one sentence per clip. Every clip is paired with one sentence that is extracted 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[1] and Subtitle Edit.[2] These audio streams are further transcribed using crowd-sourced transcription service [1]. Then the transcribed texts were aligned with associated spoken sentences using their time stamps. To remove the misalignments of audio content with the visual content itself, each sentence was also manually aligned with the corresponding video clip. During the manual alignment process, sentences describing the content not present in the video clip were also filtered out. The audio descriptions track is an added feature in the dataset trying to describe the visual content to help visually impaired persons. The total time span of the dataset videos is almost 73.6h, and the vocabulary size is 653,467.

3.2.2 M-VAD. Montreal Video Annotation Dataset (M-VAD) [139] is based on the Descriptive Video Service (DVS) and contains 48,986 video clips from 92 different movies. Each clip is spanned over 6.2s on average and the entire time for the complete dataset is 84.6h. The total number of sentences is 55,904, with few clips associated 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.

3.3 Social Media

3.3.1 VideoStory. VideoStory [45] is a multi-sentence description dataset composed of 20K social media videos. This dataset is aimed to address the story narration or description generation of long videos that may not be sufficiently illustrated with a single sentence. Each video is paired with at least one paragraph. The average number of temporally localized sentences per paragraph is 4.67. There are a total of 26,245 paragraphs in the dataset composed of 123K sentences with an average of 13.32 words per sentence. On average, each paragraph covers 96.7% of video content. The dataset contains about 22% temporal overlap between co-occurring events. The dataset has training, validation, and test split of 17,908, 999, and 1,011 videos, respectively and also proposes a blind test set composed of 1,039 videos. Each training video is accompanied with one paragraph, however, videos in the validation and test sets have three paragraphs each for evaluation. Annotations for the blind test are not released and are only available on server for benchmarking different methods.

3.3.2 ActivityNet Entities. ActivityNet Entities dataset (or ANet-Entities) [173] is the first video dataset with entities grounding and annotations. This dataset is built on the training and validation splits of the ActivityNet Captions dataset [70], but with different captions. In this dataset, noun phrases (NPs) of video descriptions have been grounded to bounding boxes in the video frames. The dataset is composed of 14,281 annotated videos, 52K video segments with at least one noun phrase annotated per segment, and 158K bounding boxes with annotations. The dataset employs a training set (10K) similar to ActivityNet Captions. However, validation set of ActivityNet Captions is randomly and evenly split into ANet-Entities validation (2.5K) and testing (2.5K) sets.

3.4 Videos in the Wild

3.4.1 MSVD. Microsoft Video Description (MSVD) dataset [22] is composed of 1,970 YouTube clips 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

[1] https://www.makemkv.com/.
[2] http://www.nikse.dk/SubtitleEdit/.

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 10s and 25s mainly showing one activity. The dataset is composed of 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 Section 5. Almost all research groups have split this dataset into training, validation, and testing partitions of 1,200, 100, and 670 videos, respectively.

3.4.2 MSR-VTT. MSR-Video to Text (MSR-VTT) [160] contains a wide variety of open domain videos for video captioning task. It is composed of 7,180 videos subdivided into 10K clips. The clips are grouped into 20 different categories. The dataset is divided into 6,513 training, 497 validation, and 2,990 test videos. Each video is composed of 20 reference captions annotated by AMT workers. In terms of the number of clips with multiple associated sentences, this is one of the largest video captioning datasets. In addition to video content, this dataset also contains audio information that can potentially be used for multimodal research.

3.4.3 Charades. This dataset [130] 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 describing 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. Videos are recorded in 15 different indoor scenes and restricted to use 46 objects and 157 action classes only. The dataset is composed of 66,500 annotations describing 157 actions. It also provides 41,104 labels to its 46 object classes. Moreover, it contains 27,847 descriptions covering all the videos. The videos in the dataset depict daily life activities with an average duration of 30s. The dataset is split into 7,985 and 1,863 videos for training and test purposes, respectively.

3.4.4 VTW. Video Titles in the Wild (VTW) [171] contains 18,100 video clips with an average of 1.5mins duration per clip. Each clip is described with one sentence only. However, it incorporates a diverse vocabulary, where on average one word appears in not more than two sentences across the whole dataset. Besides the single sentence per video, the dataset also provides accompanying descriptions (known as augmented sentences) that describe information not present in the visual content of the clip. 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.

3.4.5 ActivityNet Captions. ActivityNet Captions dataset [70] contains 100K dense natural language descriptions of about 20K videos from ActivityNet [176] that correspond to approximately 849h. On average, each description is composed of 13.48 words and covers about 36s of video. There are multiple descriptions for every video and when combined together, these descriptions cover 94.6% of content present in the entire video. In addition, 10% temporal overlap makes the dataset especially interesting and challenging for studying multiple events occurring at the same time.

4 EVALUATION METRICS

Evaluations performed over machine-generated captions/descriptions of videos can be divided into *Automatic Evaluations* and *Human Evaluations*. Automatic evaluations are performed using six different metrics that were originally designed for machine translation and image captioning. These metrics are BLEU [102], ROUGE$_L$ [82], METEOR [12], CIDEr [142], WMD [76], and SPICE [5]. Below, we discuss these metrics in detail as well as their limitations and reliability. Human

Fig. 9. An example from MSVD [22] 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.

evaluations are performed, too, because of the unsatisfactory performance of automatic metrics given that there are numerous ways to correctly describe the same video.

4.1 Automatic Sentence Generation Evaluation

Evaluation of video descriptions, automatically or manually generated, is challenging, because as there is no specific ground truth or "right answer," that can 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 also in terms of semantic content. Consider a 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 last 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) [102], Recall Oriented Understudy of Gisting Evaluation (ROUGE) [82], and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [12]. Consensus-based Image Description Evaluation (CIDEr) [142] and Semantic Propositional Image Captioning Evaluation (SPICE) [5] are two other recently introduced metrics specifically designed for image captioning tasks that are also being used for automatic evaluation of video description. Table 2 gives an overview of the metrics included in this survey. In addition to these automatic evaluation metrics, human evaluations are also employed to determine the performance of automated video description algorithms.

4.1.1 Bilingual Evaluation Understudy (BLEU, 2002). BLEU [102] is a popular metric used to quantify the quality of machine-generated text. The quality measures the correspondence between a machine and human outputs. BLEU scores take into account the overlap between predicted *unigrams* (single word) or higher order *n-gram* (sequence of *n* adjacent words) and a set of one or more candidate reference sentences. According to BLEU, a high-scoring description should match the ground truth sentence in length, i.e., exact match of words as well as their order. BLEU evaluation will score 1 for an exact match. Note that the higher the number of reference sentences in the ground truth per video, the more the chances of a higher 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,

$$\log \text{BLEU} = \min\left(1 - \frac{l_r}{l_c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n.$$

In the above equation, l_r/l_c is the ratio between the lengths of the corresponding reference corpus and the candidate description, w_n are positive weights, and p_n is the geometric average of the

Table 2. Summary of Metrics Used for Video Description Evaluation

Metric Name	Designed For	Methodology
BLEU [102]	Machine translation	*n-gram* precision
ROUGE [82]	Document summarization	*n-gram* recall
METEOR [12]	Machine translation	*n-gram* with synonym matching
CIDEr [142]	Image captioning	*tf-idf* weighted *n-gram* similarity
SPICE [5]	Image captioning	Scene-graph synonym matching
WMD [76]	Document similarity	Earth mover distance on word2vec

modified n-gram precisions. While the second term computes the actual match score, the first term is a brevity penalty that penalizes descriptions that are shorter than the reference description.

4.1.2 Recall-oriented Understudy for Gisting Evaluation (ROUGE, 2004). ROUGE [82] metric was proposed in 2004 to evaluate text summaries. It calculates recall score of the generated sentences corresponding to the reference sentences using *n-grams*. Similar to BLEU, ROUGE is also computed by varying the *n-gram* count. However, unlike BLEU, which is based on precision, ROUGE is based on recall values. Moreover, other than *n-gram* variants of ROUGE_n, it has other versions known as ROUGE_L (Longest Common Subsequence), ROUGE_W (Weighted Longest Common Subsequence), ROUGE_S (Skip-Bigram Co-Occurrences Statistics), and ROUGE_{SU} (extension of ROUGE_S). We refer the reader to the original paper for details. 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 the generated and each reference sentence. The metric compares common subsequences of words in candidate and reference sentences. The intuition behind this is that longer LCS of candidate and reference sentences correspond to higher similarity between the two summaries. The words need not be consecutive but should be in sequence. ROUGE-N is computed as

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

n being the n-gram length, g_n, and $C_m(g_n)$ represents the highest number of n-grams that are present in candidate as well as ground truth summaries, and R_{Sum} stands for reference summaries.

LCS-based F-measure score is computed to find how similar summary A of length m is to summary B of length n. Where A is a sentence from the ground truth summary and B is a sentence from the candidate-generated summary. The recall R_{lcs}, precision P_{lcs}, and f-score F_{lcs} are calculated as

$$R_{lcs} = \frac{\text{LCS}(A,B)}{m}, \quad P_{lcs} = \frac{\text{LCS}(A,B)}{n}, \quad F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}},$$

where $\text{LCS}(A,B)$ is the length of longest common subsequence between A and B, $\beta = P_{lcs}/R_{lcs}$. The LCS-based F-measure score computed by equation F_{lcs} is known as ROUGE_L score. ROUGE_L is 1 when $A = B$, and zero in case when A and B have no commonalities, i.e., $\text{LCS}(A,B) = 0$.

One of the advantages of ROUGE_L is that it does not consider successive matches of words but employs in-sequence matches within a sentence. Moreover, pre-defining the n-gram length is also not required, as this is automatically incorporated by *LCS*.

4.1.3 Metric for Evaluation of Translation with Explicit Ordering (METEOR, 2005). METEOR [12] was proposed to address the shortcomings of BLEU [102]. Instead of exact lexical match required by BLEU, METEOR introduced semantic matching. METEOR takes WordNet[39], a lexical database of the English language to account for various match levels, including exact words matches, stemmed words matches, synonymy matching, and the paraphrase matching.

METEOR score computation is based on how well the generated and reference sentences are aligned. Each sentence is taken as a set of unigrams, and alignment is done by mapping unigrams of candidate and reference sentences. During mapping, a unigram in candidate sentence (or reference sentence) should either map to unigram in reference sentence (or candidate sentence) or to zero. In case of multiple options available for alignments between the two sentences, the alignment configuration with a lower number of crossings is preferred. After finalizing the alignment process, METEOR score is calculated.

Initially, unigram-based precision score P is calculated using $P = m_{cr}/m_{ct}$ relationship. Here, m_{cr} represents the number of unigrams co-occurring in both candidate as well as reference sentences, and m_{ct} corresponds to total number of unigrams in the candidate sentences. Then unigram-based recall score R is calculated using $R = m_{cr}/m_{rt}$. Here, m_{cr} represents the number of unigrams co-occurring in both candidate as well as reference sentences. However, m_{rt} is the number of unigrams in the reference sentences. Further, precision and recall scores are used to compute the F-score using the following equation:

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

The precision, recall, and F-score measures account for unigram-based congruity and do not cater for *n-grams*. The *n-gram*-based similarities are used to calculate the penalty p for alignment between candidate and reference sentences. This penalty takes into account the non-adjacent mappings between the two sentences. The penalty is calculated by grouping the unigrams into a minimum number of chunks. The chunk includes unigrams that are adjacent in candidate as well as reference sentences. If a generated sentence is an exact match to the reference sentence, then there will be only one chunk. The penalty is computed as

$$p = \frac{1}{2}\left(\frac{N_c}{N_u}\right)^2,$$

where N_c represents the number of chunks and N_u corresponds to the number of unigrams grouped together. The METEOR score for the sentence is then computed as:

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

Corpus-level score can be computed using the same equation by using aggregated values of all the arguments, i.e., P, R, and p. In case of multiple reference sentences, the maximum METEOR score of a generated and reference sentence is taken. To date, correlation of METEOR score with human judgments is better than that of BLEU score. Moreover, Elliot et al. [35] also found METEOR to be a better evaluation metric as compared to contemporary metrics. Their conclusion is based on Spearman's correlation computation of automatic evaluation metrics against human judgments.

4.1.4 Consensus-based Image Description Evaluation (CIDEr, 2015). CIDEr [142] is a recently introduced evaluation metric for image captioning task. It evaluates the consensus between a predicted sentence c_i and reference sentences of the corresponding image. 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 predicted sentence and reference sentence, it measures the co-existence frequency of n-grams in both sentences. Finally, *n-grams* that are very common among the reference sentences of all the images are given lower weight, as they are likely to be less informative about the image content and more biased towards lexical structure of the sentences. The weight for each *n-gram* is computed using Term Frequency Inverse Document Frequency (TF-IDF) [113]. The term TF puts higher weightage on frequently

occurring *n-grams* in the reference sentence of the image, whereas IDF puts lower weightage on commonly appearing *n-grams* across the whole dataset. Finally, CIDEr_n score is computed as

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

where $g^n(c_i)$ is a vector representing all *n-grams* with length n and $\|g^n(c_i)\|$ depicts magnitude of $g^n(c_i)$. Same is true for $g^n(s_{ij})$. Further, CIDEr uses higher-order n-grams (the higher the order, the 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, which incorporates a few modifications in the originally proposed CIDEr to prevent higher scores for the captions that badly fail in human judgments. First, they proposed removal of stemming to ensure correct form of words are used. Otherwise, multiple forms of verbs (singular, plural, etc.) are mapped to the same token producing high scores for incorrect sentences. Secondly, they ensure that if the words of high confidence are repeated in a sentence, a high score is not produced as the original CIDEr produces even if the sentence does not make sense. This is done by introducing a Gaussian penalty over length differences between the candidate and reference sentences and by clipping to the *n-grams* count equal to the number of occurrences in the reference sentence. The latter ensures that the desired sentence length is not achieved by repetition of high-confidence words to get a high score. The aforementioned changes make the metric robust and ensure its high correlation score [142].

4.1.5 Word Mover's Distance (WMD, 2015). The WMD [76] makes use of word embeddings that 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. However, it is possible for multiple captions to have the same attributes, objects, and their relations while still having very different meanings. 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 cast as an Earth Mover's Distance (EMD) [123], typically used in transportation to calculate the travel cost using word2vec embeddings [91]. In this metric, each caption or description is represented by a bag-of-words histogram that includes all but the start and stop words. The magnitude of each bag-of-words histogram is then normalized. To account for semantic similarities that exist between pairs of words, the WMD metric uses the 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 WMD is modelled as a special case of EMD [123] and is then solved by linear optimization. Compared to BLUE, ROUGE, and CIDEr, WMD is less sensitive to word order or synonym swapping. Further, similar to CIDEr and METEOR, it gives high correlation against human judgments.

4.1.6 Semantic Propositional Image Captioning Evaluation (SPICE, 2016). SPICE [5] is the latest proposed evaluation metric for image and video descriptions. SPICE measures the similarity between the *scene graph tuples* parsed from the machine-generated descriptions and the ground truth. The semantic scene graph encodes objects, their attributes, and relationships through a dependency parse tree. A scene graph tuple $G(c)$ of caption c consists of semantic tokens such as

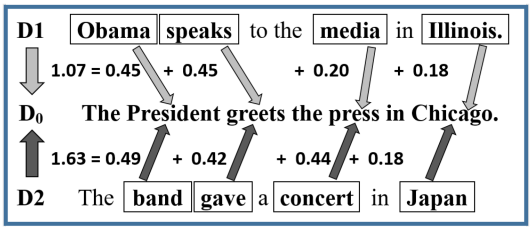

Fig. 10. Components of the WMD metric between a query D_0 and two sentences D_1 and D_2 with the same BOW distance. D_1 with less distance 1.07 matches with query D_0 than D_2 with distance 1.63. The arrows show flow between two words and are labeled with their distance contribution. Figure adapted from Reference [76].

object classes $O(c)$, relation types $R(c)$, and attribute types $A(c)$,

$$G(c) = \langle O(c), R(c), A(c) \rangle.$$

SPICE is computed based on F1-score between the tuples of machine-generated descriptions and the ground truth. Like METEOR, SPICE also uses WordNet to find and treat synonyms as positive matches. Although, in the current literature, the SPICE score has not been employed much; one obvious limiting factor on its performance could be the quality of the parsing. For instance, in a sentence "`white dog swimming through river`," the failure case could be the word "*swimming*" being parsed as "*object*," and the word "*dog*" parsed as "*attribute*," resulting in a very bad score.

4.2 Human Evaluations

Given the lack of reference captions and low correlation with human judgments of automated evaluation metrics, human evaluations are also often used to judge the quality of machine-generated captions. Human evaluations may either be crowd-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 the highest score given to the *"Most Relevant"* and minimum score to the *"Least Relevant."* The score of two sentences cannot be the same 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.

4.3 Limitations of Evaluation Metrics

Like video description, evaluation of the machine-generated sentences is an equally difficult task. There is no metric specifically designed for evaluating video description; instead, machine translation and image captioning metrics have been extended for this task. These automatic metrics compute the score given reference and candidate sentences. This paradigm has a serious problem that there can be several different ways to describe the same video, all correct at the same time, depending upon *"what has been described"* (content selection) and *"how it has been described"* (realization). These metrics fail to incorporate all these variations and are, therefore, far from being perfect. Various studies [63, 154] have examined how metric scores behave under different conditions. In Table 3, we perform similar experiments [63] but with an additional variation of *short length*. First, the original caption was evaluated with itself to analyze the maximum possible score achievable by each metric (first row of Table 3). Next, minor modifications were introduced in the candidate sentences to measure how the evaluation metrics behave. It was observed that all metric

Table 3. Variations in Automatic Evaluation Metric Scores with Four Types of Changes Made to Candidate Sentence, i.e., Words Replaced with Their Synonyms, Added Redundancy to Sentence, Changing Word Order, and Shortening the Sentence Length

Variation	Description	B	M	R	C
reference	an elderly man is playing piano in front of a crowd in an anteroom	1	1	1	10
candidate	an elderly man is showing how to play piano in front of a crowd in a hall room	0.47	0.45	0.70	0.53
synonyms	an **old** man is **demonstrating** how to play piano in front of a crowd in a hall room	0.37	0.40	0.64	0.43
redundancy	an elderly man is showing how to play piano in front of a crowd in a hall room **with a woman**	0.40	0.44	0.65	0.47
word order	an elderly man in front of a crowd is showing how to play piano in a hall room	0.30	0.39	0.57	0.35
short length	a man is playing piano	0.12	0.22	0.39	0.49

The first row shows the upper bound scores of BLEU-4, METEOR, ROUGE, and CIDEr represented by B, M, R, and C, respectively.

scores reduced, BLEU and CIDEr being the most affected, when some words were replaced with their synonyms. This is apparently due to the failure to match synonyms. Further experiments revealed that the metrics were generally stable when the sentence was perturbed with a few additional words. However, changing the word order in a sentence was found to alter the scores of *n-gram*-based metrics such as BLEU, ROUGE, and CIDEr significantly and that of ROUGE to some extent. However, WMD and SPICE were found to be robust to word order changes [63]. Last, reducing the sentence length significantly affected BLEU, METEOR, and ROUGE scores but had little effect on CIDEr score, i.e., the scores were reduced by 74%, 51%, 44%, and 7%, respectively.

4.4 Reliability of Evaluation Metrics

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

Another aspect of the evaluation problem is the syntactic variations in candidate sentences. The same 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 assess the suitability of a caption to the visual input by comparing how well the candidate caption matches reference caption(s). The agreement of the metric scores with human judgments (i.e., the gold standard) improves with the increased number of reference captions [142]. Numerous studies [99, 142, 144, 167] also found that CIDEr, WMD, SPICE, and METEOR have higher correlations to human judgments and are regarded as superior among the contemporary metrics. WMD and SPICE are very recent automatic caption evaluation metrics and had not been studied extensively in the literature at the time of this survey.

5 BENCHMARK RESULTS

We summarize the benchmark results of various techniques on each video description dataset. We group the methods based on the dataset they reported results on and then order them

Table 4. Performance of Video Captioning Methods on MSVD Dataset

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
RBS+RBS & RF-TP+RBS [52]	2012	MSVD	SVO Accuracy			
SVO-LM (VE) [71]	2013	MSVD	0.45+_0.05	0.36+_0.27		
FGM [137]	2014	MSVD	SVOP Accuracy			
LSTM-YT [145]	2015	MSVD	33.3	29.1	-	-
TA [162]	2015	MSVD	41.9	29.6	51.67	-
S2VT [144]	2015	MSVD	-	29.8	-	-
h-RNN [167]	2016	MSVD	49.9	32.6	65.8	-
MM-VDN [159]	2016	MSVD	37.6	29.0	-	-
Glove + Deep Fusion Ensble [143]	2016	MSVD	42.1	31.4	-	-
S2FT [84]	2016	MSVD	-	29.9	-	-
HRNE [99]	2016	MSVD	43.8	33.1	-	-
GRU-RCN [11]	2016	MSVD	43.3	31.6	68.0	-
LSTM-E [100]	2016	MSVD	45.3	31.0	-	-
SCN-LSTM [43]	2017	MSVD	51.1	33.5	**77.7**	-
LSTM-TSA [101]	2017	MSVD	**52.8**	33.5	74.0	-
TDDF [172]	2017	MSVD	45.8	33.3	73.0	69.7
BAE [13]	2017	MSVD	42.5	32.4	63.5	-
PickNet [24]	2018	MSVD	46.1	33.1	76.0	69.2
$M^3 - IC$ [153]	2018	MSVD	**52.8**	33.3	-	-
RecNet_{local} [150]	2018	MSVD	52.3	34.1	**80.3**	69.8
TSA-ED [156]	2018	MSVD	51.7	34.0	74.9	-
GRU-EVE [3]	2019	MSVD	47.9	**35.0**	78.1	**71.5**

Higher scores are better in all metrics. The best score for each metric is shown in bold in green cells.

chronologically. Moreover, for multiple variants of the same model, only their best reported results are reported here. 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 results, as these are the closest to human evaluations. From Table 4, we can see that most methods have reported results on the MSVD dataset, followed by MSR-VTT, M-VAD, MPII-MD, and ActivityNet Captions. The popularity of MSVD can be attributed to the diverse nature of YouTube 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 inclusion in competitions.

Another key observation is that earlier works have mainly reported results in terms of subject, verb, object (SVO) and in some cases place (scene) detection accuracies in the video, whereas more recent works report sentence-level matches using automatic evaluation metrics. Considering the diverse nature of the datasets and the limitations of automatic evaluation metrics, we analyze the results of different methods using four popular metrics, namely, BLEU, METEOR, CIDEr, and ROUGE. Table 4 summarizes results for the MSVD dataset. GRU-EVE [3] achieves the best performance on METEOR and ROUGE_L metrics and the second best on CIDEr metric, whereas LSTM-TSA [101] and M^3-*IC* [153] report the best BLEU scores. RecNet_{local} [150] has the best CIDEr score and second-best BLEU score. Table 5 shows results on the TACoS Multilevel dataset, where h-RNN [167] has the best scores on the three reported metrics (BLEU, METEOR, and CIDEr).

On the more challenging M-VAD dataset, the reported results (Table 6) are overall very poor. Only Temporal-Attention [162] and HRNE [99] reported results using the BLEU metric with a score

Table 5. Performance of Video Captioning Methods on TACoS-MLevel Dataset

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
SMT(SR) + Prob I/P [114]	2014	TACoS MLevel	28.5	-	-	-
CRF + LSTM-Decoder [33]	2015	TACoS MLevel	28.8	-	-	-
h-RNN [167]	2016	TACoS MLevel	**30.5**	**28.7**	**160.2**	-
JEDDi-Net [158]	2018	TACoS MLevel	18.1	23.85	103.98	50.85

Table 6. Performance of Video Captioning Methods on M-VAD Dataset

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
Temporal-Attention (TA) [162]	2015	M-VAD	**0.7**	5.7	**6.1**	-
S2VT [144]	2015	M-VAD	-	6.7	-	-
Visual-Labels [115]	2015	M-VAD	-	6.4	-	-
HRNE [99]	2016	M-VAD	**0.7**	6.8	-	-
Glove + Deep Fusion Ensemble [143]	2016	M-VAD	-	6.8	-	-
LSTM-E [100]	2016	M-VAD	-	6.7	-	-
LSTM-TSA [101]	2017	M-VAD	-	7.2	-	-
BAE [13]	2017	M-VAD	-	**7.3**	-	-

Table 7. Performance of Video Captioning Methods on MPII-MD Dataset

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
S2VT [144]	2015	MPII-MD	-	7.1	-	-
Visual-Labels [115]	2015	MPII-MD	-	7.0	-	-
SMT [116]	2015	MPII-MD	-	5.6	-	-
Glove + Deep Fusion Ensemble [143]	2016	MPII-MD	-	6.8	-	-
LSTM-E [100]	2016	MPII-MD	-	7.3	-	-
LSTM-TSA [101]	2017	MPII-MD	-	**8.0**	-	-
BAE [13]	2017	MPII-MD	**0.8**	7.0	**10.8**	**16.7**

of 0.7 in both cases. All other works that used this dataset reported METEOR scores with BAE [13] achieving the best METEOR score followed by LSTM-TSA [101]. HRNE [99] and Glove+Deep Fusion Ensemble [143] share the third place for METEOR score.

MPII-MD is another very challenging dataset and still has very low benchmark results, as shown in Table 7, similar to the M-VAD dataset. Only BAE [13] has a reported BLEU score for this dataset. LSTM-TSA [101] has achieved the best METEOR score followed by LSTM-E [100] and S2VT [144] at second and third place, respectively. Only BAE [13] reported CIDEr and ROUGE scores on this dataset.

Results on another popular dataset, MSR-VTT, are overall better than the M-VAD and MPII-II datasets. As shown in Table 8, CST-GT-None [104] has reported the highest score on all four metrics, i.e., BLEU, METEOR, CIDEr, and ROUGE. DenseVidCap [126] and HRL [155], respectively, report the second- and third-best scores on BLEU metric. GRU-EVE [3] reports the third-best score in METEOR and CIDEr metrics.

Table 8. Performance of Video Captioning Methods on MSR-VTT Dataset

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
Alto [127]	2016	MSR-VTT	39.8	26.9	45.7	59.8
VideoLab [108]	2016	MSR-VTT	39.1	27.7	44.4	60.6
RUC-UVA [34]	2016	MSR-VTT	38.7	26.9	45.9	58.7
v2t-navigator [59]	2016	MSR-VTT	40.8	28.2	44.8	61.1
TDDF [172]	2017	MSR-VTT	37.3	27.8	43.8	59.2
DenseVidCap [126]	2017	MSR-VTT	41.4	28.3	48.9	61.1
CST-GT-None [104]	2017	MSR-VTT	**44.1**	**29.1**	**49.7**	**62.4**
PickNet [24]	2018	MSR-VTT	38.9	27.2	42.1	59.5
HRL [155]	2018	MSR-VTT	41.3	28.7	48.0	61.7
$M^3 - VC$ [153]	2018	MSR-VTT	38.1	26.6	-	-
$RecNet_{local}$ [150]	2018	MSR-VTT	39.1	26.6	42.7	59.3
GRU-EVE [3]	2019	MSR-VTT	38.3	28.4	48.1	60.7

Table 9. Performance of Video Captioning Methods on ActivityNet Captions Dataset

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
Dense-Cap Model [70]	2017	ActivityNet Cap	**3.98**	9.5	24.6	-
LSTM-A+PG+R [163]	2017	ActivityNet Cap	-	**12.84**	-	-
TAC [107]	2017	ActivityNet Cap	-	9.61	-	-
JEDDi-Net [158]	2018	ActivityNet Cap	1.63	8.58	19.88	**19.63**
DVC [81]	2018	ActivityNet Cap	1.62	10.33	**25.24**	-
Bi-SST [152]	2018	ActivityNet Cap	2.30	9.60	12.68	19.10
Masked Transformer [175]	2018	ActivityNet Cap	2.23	9.56	-	-

Table 10. Performance of Video Captioning Methods on Various Benchmark Datasets

Techniques/Models/Methods	Yr	Dataset	Results			
			BLEU	METEOR	CIDEr	ROUGE
CT-SAN [170]	2016	LSMDC	**0.8**	7.1	**10.0**	**15.9**
GEAN [169]	2017	LSMDC	-	**7.2**	9.3	15.6
HRL [155]	2018	Charades	**18.8**	**19.5**	**23.2**	**41.4**
TSA-ED [156]	2018	Charades	13.5	17.8	20.8	-
Masked Transformer [175]	2018	YouCook-II	**1.13**	**5.90**	-	-

Results of another recent and popular ActivityNet Captions dataset are presented in Table 9. This dataset was primarily introduced for dense video captioning and is gaining popularity very quickly. In this dataset, Dense-Cap Model [70] stands at top in terms of BLEU score. Best METEOR score is reported by LSTM-A+PG+R [163]. Highest scores in CIDEr and ROUGE metrics are achieved by methods DVC [81] and JEDDi-Net [158], respectively. Finally, in Table 10, we report two results for LSMDC and Charades each and only one result for YouCook-II datasets. YouCook-II is also a recent dataset and not reported much in the literature.

We summarize the best reporting methods for each dataset along with their published scores. The tables group methods by the used dataset(s). Hence, one can infer the difficulty level of datasets

by comparing the intra dataset scores of the same methods and the popularity of a particular dataset from the number of methods that have reported results on it.

6 FUTURE AND EMERGING DIRECTIONS

Automatic video description has come very far since the pioneer methods, especially after the adoption of deep learning. Although the performance of existing methods is still far below that of humans, the gap is diminishing at a steady rate, and there is still ample room for algorithmic improvements. Here, we list several possible future and emerging directions that have the potential to advance this research area.

Visual Reasoning: Although video VQA is still in its nascent stage, beyond VQA is the visual reasoning problem. This is a very promising field to further explore. Here the model is made not to just answer a particular question but to reason why it chose that particular answer. For example, in a video where a roadside with parking marks is shown, the question is "*Can a vehicle be parked here?*," and the model answers correctly, "*Yes.*" The next question is "*Why?*" to which the model reasons that there is a parking sign on the road, which means it is legal to park here. Another example is the explanations generated by self-driving cars [64], where the system keeps the passengers in confidence by generating natural language descriptions of the reasons behind its decisions, e.g., to slow down, take a turn, and so on. An example of visual reasoning models is the MAC Network [57], which is able to think and reason giving promising results on CLEVR [60], a visual reasoning dataset.

Visual Dialogue: Similar to audio dialogue (e.g., Siri, Hello Google, Alexa, and ECHO), visual dialogue [30] is another promising and flourishing field, especially in an era where we look forward to interact with robots. In visual dialogue, given a video, a model is asked a series of questions sequentially in a dialogue/conversational manner. The model tries to answer (no matter right or wrong) these questions. This is different from visual reasoning, where the model argues the reasons that lead the model to choose particular answers.

Audio and Video: While the majority of computer vision research has focused on video description without the help of audio, audio is naturally present in most of videos. Audio can help in video description by providing background information; for instance, the sound of a train, the ocean, and traffic when there is no visual cue of their presence. Audio can additionally provide semantic information; for example, who the person is or what they are saying on the other side of the phone. It can also provide clues about the story, context, and sometimes explicitly mention the object or action to complement the video information. Therefore, using audio in video description models will certainly improve the performance [53, 98].

External Knowledge: In video description, most of the time, we are comparing the performance with humans who have extensive out-of-domain or prior knowledge. When humans watch a clip and describe it, most of the time they don't rely solely on the visual (or even the audio) content. Instead, they additionally employ their background knowledge. Similarly, it would be an interesting and promising approach to augment the video description techniques with prior external knowledge [157]. This approach has shown significantly better performance in visual question answering methods and is likely to improve video description accuracy.

Addressing the Finite Model Capacity: Existing methods are performing end-to-end training while using as much data as possible for better learning. However, this approach itself is inherently limited in learning, as no matter how big the training dataset becomes, it will never cover the combinatorial complexity of real-world events. Therefore, learning to use data rather than learning the data itself is more important and may help improve the upcoming system performances.

Automatic Evaluation Measures: So far, video description has relied on automatic metrics designed for machine translation and image captioning tasks. To date, there is no automatic video description (or even captioning) evaluation metric that is purpose-designed. Although metrics designed for image captioning are relevant, they have their limitations. This problem is going to exacerbate in the future with dense video captioning and story-telling tasks. There is a need for an evaluation metric that is closer to human judgments and that can encapsulate the diversities of realizations of visual content. A promising research direction is to use machine learning to learn such a metric rather than hand-engineer it.

7 CONCLUSION

We presented the first comprehensive literature survey of video description research, starting from the classical methods that are based on Subject-Verb-Object (SVO) tuples to more sophisticated statistical and deep learning–based methods. We reviewed popular benchmark datasets that are commonly used for training and testing these models and discussed international competitions/challenges that are regularly held to promote the video description research. We discussed, in detail, the available automatic evaluation metrics for video description, highlighting their attributes and limitations. We presented a comprehensive summary of results obtained by recent methods on the benchmark datasets using all metrics. These results not only show the relative performance of existing methods but also highlight the varying difficulty levels of the datasets and the robustness and trustworthiness of the evaluation metrics. Finally, we put forward some recommendations for future research directions that are likely to push the boundaries of this research area.

From an algorithm-design perspective, although LSTMs have shown competitive caption generation performance, the interpretablity and intelligibility of the underlying models 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 from a frame containing a "white fire hydrant," it is difficult to ascertain 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." Future research must focus on improving diagnostic mechanisms to pinpoint the problematic part of the architectures so it can be improved or replaced.

Our survey shows that a major bottleneck hindering progress along this line of research is the lack of effective and purposely designed video description evaluation metrics. Current metrics have been adopted either from machine translation or image captioning and fall short in measuring the quality of machine-generated video captions and their agreement with human judgments. One way to improve these metrics is to increase the number of reference sentences. We believe that purpose-built metrics that are learned from the data itself are the key to advancing video description research.

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 [141] that are dependent on video-segment lengths. Most feature extractors are suitable only for static or smoothly changing images and hence struggle to handle abrupt scene changes. Current methods rather simplify the visual encoding part by representing holistic videos or frames. Attention models may further need to be explored to focus on spatially and temporally significant parts of the video. Similarly, temporal modeling of the visual features itself is quite rudimentary in

existing methods. Most methods either use mean pooling, which completely discards the temporal information or uses the C3D model, which can only model 15 frames. Future research should focus on designing better temporal modeling architectures that preferably learn in an end-to-end fashion rather than disentangling the visual description from the temporal model and the temporal modeling from language description.

ACKNOWLEDGMENTS

The authors acknowledge Marcus Rohrbach (Facebook AI Research) for his valuable input. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Casting Words transcription service, 2014. Retrieved from: http://castingwords.com/.
[2] Language in Vision, 2017. Retrieved from: https://www.sciencedirect.com/journal/computer-vision-and-image-understanding/vol/163.
[3] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the CVPR*.
[4] J. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the CVPR*.
[5] P. Anderson, B. Fernando, M. Johnson, and S. Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the ECCV*.
[6] B. Andrei, E. Georgios, H. Daniel, M. Krystian, N. Siddharth, X. Caiming, and Z. Yibiao. 2015. A workshop on language and vision at CVPR 2015.
[7] B. Andrei, M. Tao, N. Siddharth, Z. Quanshi, S. Nishant, L. Jiebo, and S. Rahul. 2018. A workshop on language and vision at CVPR 2018. http://languageandvision.com/.
[8] R. Anna, T. Atousa, R. Marcus, P. Christopher, L. Hugo, C. Aaron, and S. Bernt. 2015. The joint video and language understanding workshop at ICCV 2015.
[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. VQA: Visual question answering. In *Proceedings of the ICCV*.
[10] D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. Retrieved from: *arXiv preprint arXiv:1409.0473, (2014)*.
[11] N. Ballas, L. Yao, C. Pal, and A. Courville. 2015. Delving deeper into convolutional networks for learning video representations. Retrieved from: *arXiv preprint arXiv:1511.06432, (2015)*.
[12] S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.
[13] L. Baraldi, C. Grana, and R. Cucchiara. 2017. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the CVPR*.
[14] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi et al. 2012. Video in sentences out. Retrieved from: *arXiv preprint arXiv:1204.2742, (2012)*.
[15] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, and M. I. Jordan. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3 (Feb. 2003), 1107–1135.
[16] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. A. Forsyth. 2004. Names and faces in the news. In *Proceedings of the CVPR*.
[17] A. F. Bobick and A. D. Wilson. 1997. A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 12 (1997), 1325–1337.
[18] G. Burghouts, H. Bouma, R. D. Hollander, S. V. D. Broek, and K. Schutte. 2012. Recognition of 48 human behaviors from video. In *Proceedings of the OPTRO*.
[19] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the CVPR*.
[20] M. Brand. 1997. The "Inverse Hollywood problem": From video to scripts and storyboards via causal analysis. In *Proceedings of the AAAI/IAAI*. Citeseer, 132–137.

[21] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. 2009. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Proceedings of the CVPR*.
[22] D. Chen and W. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL: Human Language Technologies-Volume 1*. ACL, 190–200.
[23] D. Chen, W. Dolan, S. Raghavan, T. Huynh, and R. Mooney. 2010. Collecting highly parallel data for paraphrase evaluation. In *J. Artific. Intell. Res.* 37 (2010), 397–435.
[24] Y. Chen, S. Wang, W. Zhang, and Q. Huang. 2018. Less is more: Picking informative frames for video captioning. Retrieved from: *arXiv preprint arXiv:1803.01457, (2018)*.
[25] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. Retrieved from: *arXiv preprint arXiv:1409.1259, (2014)*.
[26] J. Corso. 2015. GBS: *Guidance by Semantics—Using High-level Visual Inference to Improve Vision-based Mobile Robot Localization*. Technical Report. State University of New York at Buffalo Amherst.
[27] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the CVPR*.
[28] N. Dalal, B. Triggs, and C. Schmid. 2006. Human detection using oriented histograms of flow and appearance. In *Proceedings of the ECCV*.
[29] P. Das, C. Xu, R. F. Doell, and J. J. Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the CVPR*.
[30] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. Moura, D. Parikh, and D. Batra. 2017. Visual dialog. In *Proceedings of the CVPR*.
[31] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei. 2009. Construction and analysis of a large scale image ontology. *Vis. Sci. Soc.* 186, 2 (2009).
[32] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann. 2012. Beyond audio and video retrieval: Towards multimedia summarization. In *Proceedings of the ICMR*.
[33] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. 2015. Long-term RCNN for visual recognition and description. In *Proceedings of the CVPR*.
[34] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. M. Snoek. 2016. Early embedding and late reranking for video captioning. In *Proceedings of the MM*. ACM, 1082–1086.
[35] D. Elliott and F. Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the ACL: Short Papers*, Vol. 452. 457.
[36] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338.
[37] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt et al. 2015. From captions to visual concepts and back. In *Proceedings of the CVPR*.
[38] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the ECCV*.
[39] C. Fellbaum. 1998. WordNet. *Wiley Online Library*. Bradford Books.
[40] P. Felzenszwalb, D. McAllester, and D. Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the CVPR*.
[41] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. 2010. Cascade object detection with deformable part models. In *Proceedings of the CVPR*.
[42] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 9 (2010), 1627–1645.
[43] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the CVPR*.
[44] A. George, B. Asad, F. Jonathan, J. David, D. Andrew, M. Willie, M. Martial, S. Alan, G. Yvette, and K. Wessel. 2017. TRECVID 2017: Evaluating ad hoc and instance video search, events detection, video captioning, and hyperlinking. In *Proceedings of the TRECVID*.
[45] S. Gella, M. Lewis, and M. Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the EMNLP*. 968–974.
[46] S. Gong and T. Xiang. 2003. Recognition of group activities using dynamic probabilistic networks. In *Proceedings of the ICCV*.
[47] A. Graves and N. Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the ICML*. 1764–1772.
[48] A. Graves, A. Mohamed, and G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the ICASSP*. 6645–6649.
[49] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. 2013. Recognizing and describing activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the ICCV*.

[50] S. Guadarrama, L. Riano, D. Golland, D. Go, Y. Jia, D. Klein, P. Abbeel, T. Darrell et al. 2013. Grounding spatial relations for human-robot interaction. In *Proceedings of the IROS*. 1640–1647.
[51] A. Hakeem, Y. Sheikh, and M. Shah. 2004. CASE^E: A hierarchical event representation for the analysis of videos. In *Proceedings of the AAAI*. 263–268.
[52] P. Hanckmann, K. Schutte, and G. J. Burghouts. 2012. Automated textual descriptions for a wide range of video events with 48 human actions. In *Proceedings of the ECCV*.
[53] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the ECCV*.
[54] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the CVPR*.
[55] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.
[56] S. Hongeng, F. Brémond, and R. Nevatia. 2000. Bayesian framework for video surveillance application. In *Proceedings of the ICPR*, Vol. 1. IEEE, 164–170.
[57] Drew A. Hudson, Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *Proceedings of the ICLR*.
[58] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the MM*. ACM, 675–678.
[59] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann. 2016. Describing videos using multi-modal fusion. In *Proceedings of the MM*. ACM, 1087–1091.
[60] J. Johnson, B. Hariharan, L. V. D. Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the CVPR*.
[61] M. U. G. Khan and Y. Gotoh. 2012. Describing video contents in natural language. In *Proceedings of the EACL Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*. ACL, 27–35.
[62] M. U. G. Khan, L. Zhang, and Y. Gotoh. 2011. Human focused video description. In *Proceedings of the ICCV Workshops*.
[63] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. 2016. Re-evaluating automatic metrics for image captioning. Retrieved from: *arXiv preprint arXiv:1612.07600, (2016)*.
[64] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the ECCV*.
[65] W. Kim, J. Park, and C. Kim. 2010. A novel method for efficient indoor-outdoor image classification. *J. Sig. Proc. Syst.* 61, 3 (2010), 251–258.
[66] R. Kiros, R. Salakhutdinov, and R. S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. Retrieved from: *arXiv preprint arXiv:1411.2539, (2014)*.
[67] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, WadeShen, C. Moran, R. Zens et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL, 177–180.
[68] A. Kojima, T. Tamura, and K. Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. J. Comput. Vis.* 50, 2 (2002), 171–184.
[69] D. Koller, N. Heinze, and H. Nagel. 1991. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. In *Proceedings of the CVPR*. 90–95.
[70] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. 2017. Dense-captioning events in videos. Retrieved from: *arXiv:1705.00754*.
[71] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of the AAAI*, Vol. 1. 2.
[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the NIPS*. 1097–1105.
[73] P. Kuchi, P. Gabbur, P. S. Bhat, and S. S. David. 2002. Human face detection and tracking using skin color modeling and connected component operators. *IEEE J. Res.* 48, 3–4 (2002), 289–293.
[74] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the CVPR*.
[75] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Proceedings of the NIPS*. 3675–3683.
[76] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the ICML*.
[77] I. Langkilde-Geary and K. Knight. Halogen Input Representation. [Online]. http://www.isi.edu/publications/licensed-sw/halogen/interlingua.html.
[78] M. W. Lee, A. Hakeem, N. Haering, and S. Zhu. 2008. Save: A framework for semantic annotation of visual events. In *Proceedings of the CVPR Workshops*. 1–8.

[79] L. Li and B. Gong. 2018. End-to-end video captioning with multitask reinforcement learning. Retrieved from: *arXiv preprint arXiv:1803.07950, (2018).*
[80] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the CNLL.*
[81] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the CVPR.*
[82] C. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out.* 74–81.
[83] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the ECCV.*
[84] Y. Liu and Z. Shi. 2016. Boosting video description generation by explicitly translating from frame-level captions. In *Proceedings of the MM.* ACM, 631–634.
[85] D. G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the ICCV.*
[86] I. Maglogiannis, D. Vouyioukas, and C. Aggelopoulos. 2009. Face detection and recognition of natural human emotion using Markov random fields. *Pers .Ubiq. Comput.* 13, 1 (2009), 95–101.
[87] M. Malinowski and M. Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the NIPS.* 1682–1690.
[88] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the ICCV.*
[89] M. Margaret, H. Ting-Hao, F. Frank, and M. Ishan. 2018. In *Proceedings of the First Workshop on Storytelling.* ACL. https://www.aclweb.org/anthology/W18-1500.
[90] C. Matuszek, D. Fox, and K. Koscher. 2010. Following directions using statistical machine translation. In *Proceedings of the HRI.*
[91] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the NIPS.* 3111–3119.
[92] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. 2013. Playing Atari with deep reinforcement learning. Retrieved from: *arXiv preprint arXiv:1312.5602. (2013).*
[93] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassbis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 9529.
[94] D. Moore and I. Essa. 2002. Recognizing multitasked activities from video using stochastic context-free grammar. In *Proceedings of the AAAI/IAAI.* 770–776.
[95] R. Nevatia, J. Hobbs, and B. Bolles. 2004. An ontology for video event representation. In *Proceedings of the CVPR Workshop.* 119–119.
[96] F. Nishida and S. Takamatsu. 1982. Japanese-English translation through internal expressions. In *Proceedings of the COLING, Volume 1.* Academia Praha, 271–276.
[97] F. Nishida, S. Takamatsu, T. Tani, and T. Doi. 1988. Feedback of correcting information in post editing to a machine translation system. In *Proceedings of the COLING, Volume 2.* ACL, 476–481.
[98] A. Owens and A. A. Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the ECCV.*
[99] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the CVPR.*
[100] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the CVPR.*
[101] Y. Pan, T. Yao, H. Li, and T. Mei. 2017. Video captioning with transferred semantic attributes. In *Proceedings of the CVPR.*
[102] K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the ACL.* 311–318.
[103] R. Pasunuru and M. Bansal. 2017. Reinforced video captioning with entailment rewards. Retrieved from: *arXiv preprint arXiv:1708.02300, (2017).*
[104] S. Phan, G. E. Henter, Y. Miyao, and S. Satoh. 2017. Consensus-based sequence training for video captioning. Retrieved from: *arXiv preprint arXiv:1712.09532, (2017).*
[105] C. S. Pinhanez and A. F. Bobick. 1998. Human action detection using PNF propagation of temporal constraints. In *Proceedings of the CVPR.*
[106] C. Pollard and I. A. Sag. 1994. *Head-driven Phrase Structure Grammar.* University of Chicago Press.
[107] S. Chen, Y. Song, Y. Zhao, J. Qiu, Q. Jin, and A. Hauptmann. 2017. RUC-CMU: System descriptions for the dense video captioning task. Retrieved from: *arXiv preprint arXiv:1710.08011, (2017).*

[108] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko. 2016. Multimodal video description. In *Proceedings of the MM*. ACM, 1092–1096.
[109] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. 2013. Grounding action descriptions in videos. *Trans. Assoc. Comput. Ling.* 1 (2013), 25–36.
[110] E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.
[111] M. Ren, R. Kiros, and R. Zemel. 2015. Exploring models and data for image question answering. In *Proceedings of the NIPS*. 2953–2961.
[112] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. Retrieved from: *arXiv preprint arXiv:1704.03899, (2017)*.
[113] S. Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *J. Doc.* 60, 5 (2004), 503–520.
[114] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the GCPR*.
[115] A. Rohrbach, M. Rohrbach, and B. Schiele. 2015. The long-short story of movie description. In *Proceedings of the GCPR*. 209–221.
[116] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. 2015. A dataset for movie description. In *Proceedings of the CVPR*.
[117] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. 2017. Movie description. *Int. J. Comput. Vis.* 123, 1 (2017), 94–120.
[118] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. 2012. A database for fine-grained activity detection of cooking activities. In *Proceedings of the CVPR*.
[119] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. 2013. Translating video content to natural language descriptions. In *Proceedings of the ICCV*.
[120] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. 2012. Script data for attribute-based recognition of composite activities. In *Proceedings of the ECCV*.
[121] D. Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artific. Intell.* 167, 1–2 (2005), 170–205.
[122] D. Roy and E. Reiter. 2005. Connecting language to the world. *Artific. Intell.* 167, 1–2 (2005), 1–12.
[123] Y. Rubner, C. Tomasi, and L. J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* 40, 2 (2000), 99–121.
[124] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252.
[125] M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Sig. Proc.* 45, 11 (1997), 2673–2681.
[126] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y. Jiang, and X. Xue. 2017. Weakly supervised dense video captioning. In *Proceedings of the CVPR*.
[127] R. Shetty and J. Laaksonen. 2016. Frame- and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the MM*. ACM, 1073–1076.
[128] J. Shi and C. Tomasi. 1994. Good features to track. In *Proceedings of the CVPR*.
[129] A. Shin, K. Ohnishi, and T. Harada. 2016. Beyond caption to narrative: Video captioning with multiple sentences. In *Proceedings of the ICIP*.
[130] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the ECCV*.
[131] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. Retrieved from: *arXiv preprint arXiv:1409.1556, (2014)*.
[132] N. Srivastava, E. Mansimov, and R. Salakhudinov. 2015. Unsupervised learning of video representations using LSTMs. In *Proceedings of the ICML*. 843–852.
[133] C. Sun and R. Nevatia. 2014. Semantic aware video transcription using random forest classifiers. In *Proceedings of the ECCV*.
[134] I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence-to-sequence learning with neural networks. In *Proceedings of the NIPS*. 3104–3112.
[135] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the CVPR*.
[136] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI*.

[137] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the COLING* 2, 5 (2014), 9.
[138] C. Tomasi and T. Kanade. 1991. Detection and tracking of point features. Technical Report CMU-CS-91-132. Carnegie Mellon University.
[139] A. Torabi, C. Pal, H. Larochelle, and A. Courville. 2015. Using descriptive video services to create a large data source for video annotation research. Retrieved from: *arXiv preprint arXiv:1503.01070, (2015).*
[140] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. 2003. Context-based vision system for place and object recognition. In *Proceedings of the ICCV*.
[141] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2014. C3D: Generic features for video analysis. Retrieved from: *CoRR* abs/1412.0767, (2014).
[142] R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the CVPR*.
[143] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. 2016. Improving LSTM-based video description with linguistic knowledge mined from text. Retrieved from: *arXiv preprint arXiv:1604.01729, (2016).*
[144] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. 2015. Sequence-to-sequence video to text. In *Proceedings of the ICCV*.
[145] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. Retrieved from: *arXiv preprint arXiv:1412.4729, (2014).*
[146] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. 2017. Feudal networks for hierarchical reinforcement learning. Retrieved from: *arXiv preprint arXiv:1703.01161, (2017).*
[147] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the CVPR*.
[148] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the CVPR*.
[149] Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating New York City through grounded dialogue. Retrieved from: *CoRR* abs/1807.03367 (2018).
[150] B. Wang, L. Ma, W. Zhang, and W. Liu. 2018. Reconstruction network for video captioning. In *Proceedings of the CVPR*.
[151] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. 2009. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the BMVC*. BMVA Press, 124–1.
[152] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the CVPR*.
[153] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan. 2018. M3: Multimodal memory modelling for video captioning. In *Proceedings of the CVPR*.
[154] J. K. Wang and R. Gaizauskas. 2016. Cross-validating image description datasets and evaluation metrics. In *Proceedings of the LREC*. European Language Resources Association, 3059–3066.
[155] X. Wang, W. Chen, J. Wu, Y. Wang, and W. Y. Wang. 2017. Video captioning via hierarchical reinforcement learning. Retrieved from: *arXiv preprint arXiv:1711.11135, (2017).*
[156] X. Wu, G. Li, Q. Cao, Q. Ji, and L. Lin. 2018. Interpretable video captioning via trajectory structured localization. In *Proceedings of the CVPR*.
[157] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the CVPR*.
[158] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko. 2018. Joint event detection and description in continuous video streams. Retrieved from: *arXiv preprint arXiv:1802.10250, (2018).*
[159] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, and K. Saenko. 2015. A multi-scale multiple instance video description network. Retrieved from: *arXiv preprint arXiv:1505.05914, (2015).*
[160] J. Xu, T. Mei, T. Yao, and Y. Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*.
[161] R. Xu, C. Xiong, W. Chen, and J. J. Corso. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI*, Vol. 5, 6.
[162] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the ICCV*.
[163] T. Yao, Y. Li, Z. Qiu, F. Long, Y. Pan, D. Li, and T. Mei. 2017. Trimmed action recognition, temporal action proposals and dense-captioning events in videos. In *Proceedings of the MSR Asia MSM at ActivityNet Challenge 2017.*
[164] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2 (2014), 67–78.
[165] H. Yu and J. M. Siskind. 2013. Grounded language learning from video sentences. In *Proceedings of the ACL* 1. 53–63.

[166] H. Yu and J. M. Siskind. 2015. Learning to describe video with weak supervision by exploiting negative sentential information. In *Proceedings of the AAAI*. 3855–3863.

[167] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the CVPR*.

[168] L. Yu, E. Park, A. C. Berg, and T. L. Berg. 2015. Visual Madlibs: Fill in the blank description generation and question answering. In *Proceedings of the ICCV*.

[169] Y. Yu, J. Choi, Y. Kim, K. Yoo, S. Lee, and G. Kim. 2017. Supervising neural attention models for video captioning by human gaze data. In *Proceedings of the CVPR*.

[170] Y. Yu, H. Ko, J. Choi, and G. Kim. 2016. End-to-end concept word detection for video captioning, retrieval, and question answering. Retrieved from: *arXiv preprint arXiv:1610.02947, (2016)*.

[171] K. Zeng, T. Chen, J. C. Niebles, and M. Sun. 2016. Title generation for user-generated videos. In *Proceedings of the ECCV*.

[172] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian. 2017. Task-driven dynamic fusion: Reducing ambiguity in video description. In *Proceedings of the CVPR*.

[173] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach. 2018. Grounded video description. Retrieved from: *arXiv preprint arXiv:1812.06587 (2018)*.

[174] L. Zhou, C. Xu, and J. J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI*.

[175] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the CVPR*.

[176] S. Zhu and D. Mumford. 2007. A stochastic grammar of images. *Found. Trends Comput. Graph. Vis.* 2, 4 (2007), 259–362.

Received March 2019; accepted August 2019