Natural Language Descriptions for Human Activities in Video Streams

Nouf Al Harbi
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
Department of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia
nmalharbi1@sheffield.ac.uk

Yoshihiko Gotoh
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
y.gotoh@sheffield.ac.uk

Abstract

There has been continuous growth in the volume and ubiquity of video material. It has become essential to define video semantics in order to aid the searchability and retrieval of this data. We present a framework that produces textual descriptions of video, based on the visual semantic content. Detected action classes rendered as verbs, participant objects converted to noun phrases, visual properties of detected objects rendered as adjectives and spatial relations between objects rendered as prepositions. Further, in cases of zero-shot action recognition, a language model is used to infer a missing verb, aided by the detection of objects and scene settings. These extracted features are converted into textual descriptions using a template-based approach. The proposed video descriptions framework evaluated on the NLDHA dataset using ROUGE scores and human judgment evaluation.

1 Introduction

The field of computer vision has advanced to detect humans, identify their activities, or to discriminate between a large number of object classes and assign them attributes. The outcome is usually a compact semantic representation that encodes activities associated with object categories. Such representations could be easily processed and interpreted by automatic systems. However, the natural way to convey this kind of information to humans is through natural language. Thus, this paper addresses the issue of producing textual descriptions for human activities in videos. This task has a range of applications, such as human-computer/robot interaction, video summarising, indexing and retrieval. Furthermore, translation between visual video content and language provides a solid foundation for understanding relations between vision and linguistics, as they are the closest modalities to interact with humans.

Generating textual descriptions of visual content is an intriguing task that requires a combination of two major research aspects: visual recognition approaches and natural language generation (NLG) techniques. To generate descriptions for videos and images, a template-based approach is a powerful tool though one which needs to be manually identified (Kulkarni et al., 2011; Barbu et al., 2012; Gygli et al., 2014a; Khan et al., 2015). An alternative approach is to retrieve descriptive sentences from a training corpus based on visual similarity, or to utilise externally textual-based corpora to help rank the visual detections (Farhadi et al., 2010; Kuznetsova et al., 2012; Mitchell et al., 2012; Hanckmann et al., 2012; Das et al., 2013b).

The most relevant researches to us are the (Khan et al., 2015) and (Barbu et al., 2012). Both of these approaches identify high-level features (HLFs) such as humans, chairs, and so forth, and generate textual descriptions using a template-based approach. (Khan et al., 2015) propose a method that relies on treating a video as a sequence of frames, and performs image detection for each frame independently, to identify HLFs without exploiting the temporal domain. Alternatively, (Barbu et al., 2012) have used a dataset with simple video settings where only one action is performed. Consequently, their natural language descriptions consist of one sentence.

In contrast, this study focuses on generating de-

Proceedings of The 10th International Natural Language Generation conference, pages 85–94,
Santiago de Compostela, Spain, September 4-7 2017.

scriptions of human activities in videos sequences at a shot-based level, relying mainly on visual detections. Specifically, objects tracks and their visual attributions are extracted from each shot, along with their spatial and temporal relations. In cases of zero-shot action recognition, where no verb (action class) is assigned for a given track, the detected objects classes are used to mine the relative verb from web-scale textual corpora via incorporated text-mined likelihoods. Structuring videos at shot-level enables us to utilise the temporal information associated with video data. Finally, the set of detected HLFs will be used to generate the final description for the video using a template-based approach.

2 Related Work

Video data introduces the additional dimension of time, with an associated set of challenges, such as temporal continuity. The majority of the literature pertaining to video descriptions has centred around two fundamental themes: deriving the description from semantic visual content and/or mining the relevant description from text-based corpora.

(Barbu et al., 2012) demonstrate a method whereby a single sentential description of a short video is generated by visual recognition techniques to render the language entities; specifically an event recognition approach is utilised to identify object tracks, role assignment and body posture variability. Finally, generation is achieved by pre-defined templates for each event class, in the form of subject-action-object. (Khan et al., 2015) and (Hanckmann et al., 2012) introduce a video description framework which starts with the extraction of the set of HLFs by the implementation of conventional image processing techniques. Context-free grammar (CFG) is used next to convert the extracted concepts into natural language descriptions. The drawback of these techniques is that they rely on only a limited set of high-level concepts, without exploiting text mined from text-based corpora. Moreover, videos are manipulated as sequences of images; hence no interaction between objects is considered over the time domain.

(Guadarrama et al., 2013) introduce a new framework that addresses the challenges associated with describing activities 'in-the-wild'. The method encompasses a wide range of verbs, objects and functions in an out-of-domain manner that does not necessitate videos consisting of the precise activity. If it is unable to provide a precise prediction by using the pre-trained model, it will generate a more concise and credible answer. The semantic hierarchies are learned from web-based corpora in order to decide upon the most suitable degree of generalisation. However, this work focuses on short videos clips that depict one activity; hence the resulting descriptions consist of single sentences, without investigation of any temporal associations between objects.

(Gygli et al., 2014b) describe a novel way to carry out video summarisation, the process of which is initiated by segmenting the video via the use of a 'super-frame'. Then, the degree to which the visuals are appealing is approximated for every super-frame with the use of low-, mid- and high-level characteristics. On the basis of this scoring method, an ideal subset of super-frames is chosen to produce an informative summary. However, this approach concentrates mainly on subject, verb, object (SVO) triples, without taking into account the spatial and temporal associations between objects.

(Thomason et al., 2014) integrate the use of linguistics and computer vision techniques in order to enhance the description of objects in real-life videos. They propose a method through which textual descriptions of videos could be generated by combining visual detections with language statistics, via the use of a factor graph model. A conventional visual detection system was used to detect and score objects, activities and scenes involved in the video. Then, the factor graph model combines these detection confidences with probabilistic knowledge mined from text corpora to estimate the most likely subject, verb, object, and place. Again, this study targets videos with single activity without identification of spatial and temporal relations.

In contrast to earlier researches, through which individual presences have been determined through the use of the DPM model (Felzenszwalb et al., 2010) at a frame-based level, our approach is different in several important ways. We consider the video as 3D (x, y, t), and consequently individual detection is achieved by the recent human body segmentation approach introduced in (Al Harbi and Gotoh,

verbs:	clap, wave, jog, run, walk, dive, kick, lift, ride, skate, swing, answer phone, drive, eat, fight, kiss, hug, sit down, sit up, stand up, get out, hand shake, approach, carry, catch, collide, drop, high five, depart and touch
nouns:	man, women, baby, child , person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, phone, TV/monitor, home, road, bedroom, park, hotel , kitchen, living room , office, restaurant and shop
prepositions:	in, on, next to, to the left, to the right, under, beside, above and inside
conjunctions:	and, after, before, while, later, then, next, finally
adverbs:	away and toward
adjectives:	small, big, young, old, angry, happy, sad, surprised, serious and disgust
pronouns:	he, she, they, him and her
articles:	a, an, the
auxiliary:	is

Table 1: The set of vocabulary used to produce textual descriptions of video.

2015b). This approach is designed for video data, to alleviate the shortcomings of the DPM model, such as partial occlusion, background noise and temporal variation. As a result it provides reliable physical interpretations. Visual attributes for regions of detected salience are extracted, along with their spatial and temporal relations, to avoid generating long, complex and unnatural textual descriptions. The video in this approach is structured as a sequence of shots, to preserve the order of activities, combining the sentence description of each shot to generate a coherent multi-sentence video description at the required level of detail. Additionally, our work utilises a language model trained on text-based corpora only in cases of zero-shot action recognition, where no action class is detected, drawing on detected object tracks and scene setting information.

3 Framework for Generating Textual Video Description

Figure 1 shows the overall approach for the video description task, while Table 1 illustrates the set of vocabulary used to generate textual descriptions of video. The generating of video descriptions task basically includes two main modules: content planning and a surface realizer. In our system, the content planning is mainly accomplished by improved visual recognition techniques, with the exception of the case of zero-shot action recognition, where language statistics are utilised to infer the verb class, given the detected subject and object classes. For the surface realizer stage, the template-based approach is used to generate a single sentence shot-based description. The following describes each of these components in turn.

3.1 Visual recognition of Subjects

As humans are the main participants in the video activities, in this study the role of subject is assigned to human objects if they are present. A recent model that detects and segments human body regions across video frames is utilised (Al Harbi and Gotoh, 2015b), rather than using the human detector of (Felzenszwalb et al., 2010), which is used by all previous works in generating video descriptions. This approach improves visual detection by focusing only on human regions rather than on holistic features (*e.g.*dense trajectories). As a result, a list of human objects tracks is extracted which will be used for further processing to identify their adjective attributes, such as gender (Bekios-Calfa et al., 2011), age (Horng et al., 2001) and emotion (Garg and Choudhary, 2012), using conventional image processing techniques.

3.2 Visual recognition of Objects

We used the discriminatively trained part-based models from (Felzenszwalb et al., 2010) in order to detect the non-human objects present in each video, creating a store of twenty object classes: bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, phone and TV/monitor. As these object detectors are mainly designed for images, they are applied to each keyframe, in order to obtain the maximum scores allocated to each objects, and top two objects are chosen per frame to reduce the false positive detections.

3.3 Visual recognition of Verbs

We aim to process and represent complex actions that are difficult to track efficiently using conventional descriptors. To this end a recent model for action representation that relies on extracted human regions is used from (Al Harbi and Gotoh, 2015b). It formulates a descriptor that encompasses the static and dynamic features of detected segments. After several trials the classifier is applied every ten frames, to assign a human objects track with the appropriate action class. In our experiment, 30 different action classes are used to train the model, with an extra negative class that is assign to any action that doesnt appear in the training data.

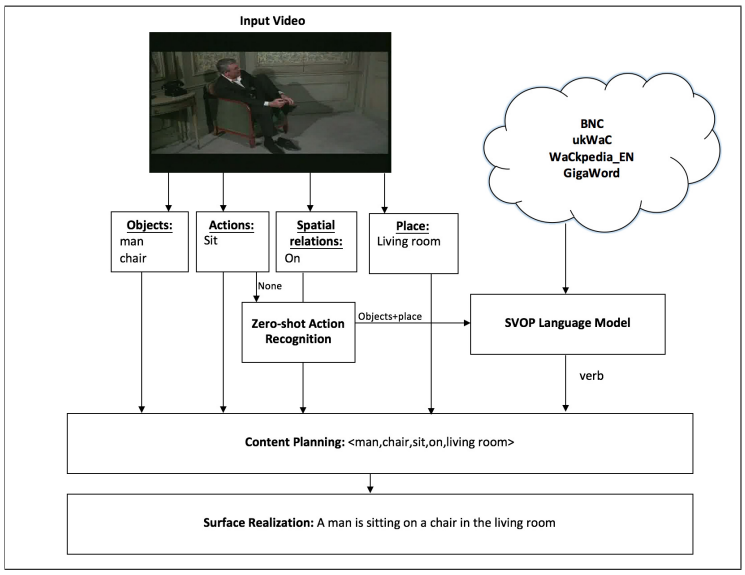

Figure 1: Summary of proposed framework of generation of video description.

3.4 Visual recognition of Prepositions

Generating elaborate textual descriptions demands more than simply applying object detection and event recognition. Producing a sentence with the embedding of spatial relations as a prepositional phrase requires the extraction of spatial relations between the detected interacting objects. To efficiently and accurately represent the relationships between the interacting objects present in a video stream, the AngledCORE-9 is adopted in (Al Harbi and Gotoh, 2015a) is utilised. Firstly, an approximated region of OBB is replaced with a space-time volume for detected objects and for each extracted region a tight OBB is drawn. Finally, the compact CORE-9 representation is used to extract the spatial and temporal aspects for multiple inter-related object bodies by analysing the nine cores and six intervals in each binary relation. Compared to the commonly used representation CORE-9, the object-volume based method has a higher chance of generating reliable results regarding the direction of objects, topologies, size, distances and temporal changes. Symmetric relations are not allowed between any pairs, to eliminate the redundancy. In this study, the following prepositions are identified, including in, on, away from, next to, to the left, to the right, under, toward, beside, above and inside.

3.5 Visual recognition of Scene Settings

In order to accurately identify the scene featured in the corpus for this study, the environment recognition method suggested by (Zhou et al., 2014) was employed. The method was used to identify the scene setting of the first frame in each shot whether it was an indoor or an outdoor scene, with a ranked list of the five most likely place categories. For this experiment, 12 different scenes settings are exist and recognised for both between indoor and outdoor settings, for each of which the associated preposition is assigned manually.

3.6 Zero-shot Language Statistics

Our approach to generating a textual video description relies mainly on visual semantic content. However, there is a case called zero-shot action recognition where is the action recognition system is unable to identify the performed action, as the action has not previously appeared in the training data; in this case a negative class is assigned. Subsequently, language statistics will be used to predict the missing verb (action class), given a detected objects classes and recognised scene settings.

Language statistics are mined from four large text-based English corpora. As in (Thomason et al., 2014) the dependency parser[1] is used to parse

[1]The spacy's API: https://spacy.io

text from the following corpora: English Gigaword, British National Corpus (BNC), ukWac and WaCkypedia-EN. The quadruple of SVOP (subject, verb, object, place) are extracted using the dependency parser. The subject-verb relations are extracted on the basis of nsubj dependencies, while the verb-object relations are identified by dobj and prep dependencies (prep dependencies are used in order to account for intransitive verbs that occur with prepositional objects). Object-place relations are extracted by utilising the prep dependencies where the noun affected by the preposition belong to the recognisable places list.

The quadruple frequency of SVOP are maintained and if no object or place is present in the sentence, their values in the quadruple are None. For the best performance, the frequency counts are a python dictionary with verbs as keys, and for each verb we keep the count of each context (subject, object, place) that co-occurs with that verb. To propose the best verb for a given context, the conditional probability $P(V|S,O,P)$ is calculated by maximum likelihood estimate (MLE) as follows:

$$P(V|S,O,P) = \frac{P(V,S,O,P)}{P(S,O,P)} = \frac{Count(V,S,O,P)}{Count(S,O,P)} \quad (1)$$

The verb with high probability given the context of subject, object and place is chosen to generate the sentence.

3.7 Sentence Generation

Finally, the extracted information from previous stages will be used to generate informative descriptions for each shot. For this purpose the template-based approach will be used. The same template will be used to create a description for each human track present in the video shot, if no human track is detected the object is considered as a subject and described in term of it motion. The list of generated sentences will be further processed to generate a coherent description. Like (Thomason et al., 2014), the following template will be used for the generation task:

'Determiner (A, The) - Adjective (optional)- Subject - Verb (Present Continuous) - Preposition (optional) - Determiner (A, The) - Adjective (optional) - Object (optional) - Preposition (optional) - Determiner (A,The) - Place (optional)'.

For implementation purposes, the surface realizer simpleNLG is utilised (Gatt and Reiter, 2009). This package also provides some extra processing applied automatically to the generated sentence: (1) the first letter is capitalised for each sentence; (2) -ing is attached to the verb if the progressive tense is chosen; (3) the words are assembled in the correct grammatical order; (4) white spaces are automatically added to separate words; and (5) at the end of each sentence a full stop is inserted.

3.8 Creating Cohesive Descriptions

Our system independently describes each video shot. The generated multi-sentence descriptions for the video as a whole tend to be a 'list of sentences' rather than a coherent 'text'. Generating coherent natural language descriptions requires linking sentences at a surface level without any need for deep understanding of the text produced. Hence, the generated list of sentences for each video is automatically post-processed at two levels shot-level and video-level in order to create more cohesive and informative descriptions. First, each human track in each shot will be described independently in a complete sentence, which results in a list of sentences describing a given shot. The following set of rules is applied in order to generate compact and coherent sentence:

1. When multiple subjects perform the same action at the same time, the subjects of these sentences are combined by 'and'. (*e.g.*If (i) 'A man is eating.' and (ii) 'A woman is eating.' these are combined to become (iii) 'A man and woman are eating.')

2. If multiple subjects perform different action simultaneously, they will be combined using 'while'. (*e.g.*in Figure 2 (a)(b)).

3. In the case where multiple subjects interact to create certain common actions (*e.g.*hug or fight), one is considered as the subject while the other(s) serve as objects in the sentence. (*e.g.*If (i) 'A man is fighting.' and (ii) 'A man is fighting.' these are combined to become (iii) 'A man is fighting another man.')

4. Proper pronouns (co-reference) are added if multiple verbs are allocated to the same subject during the same video shot. In this case, when a subject is mentioned again after its debut, a proper pronoun is used to improve the sentences concision. (*e.g.*in Figure 2 (c)(d)).

Secondly, shot-based descriptions are combined to produce the final video description. For this purpose the following rules are applied:

1. Temporal adverbials (*e.g.*next, then and finally) are incorporated between subsequent sentences as a powerful device for conserving the logical order of events performed over different shots.

2. Scene-setting information is added only to the leading sentence and discarded from subsequent sentences if the event take place in the same setting to eliminate redundancy.

3. The phrase 'In this video,' is added to the leading sentence of each video description.

4 Experiments and results

This section presents the evaluation procedure of our video description framework on the NLDHA dataset introduced in (Al Harbi and Gotoh, 2016). First, a brief overview of the baseline approach used to provide a comparison with our system is presented. Next, the results of quantitative evaluation with the ROUGE Metric, along with qualitative human judgements, are discussed.

4.1 Frame-based Video Description Baseline

To put our performance in perspective, we compare our proposed approach against the baseline video description framework of (Khan et al., 2015). This approach is chosen as the baseline as it augments the sentence components largely on the basis of semantic video content by applying conventional image processing techniques. Additionally, in order to make a fair comparison, the same set of detected objects are used for both systems. However, we advanced the detection to accommodate temporal information from the videos. The baseline approach processes the video as a sequence of frames. For each frame, conventional image processing methods are implemented to extract a set of high-level visual features (*e.g.*humans and their activities). A limited set of spatial relations are calculated between the extracted HLFs geometric features, though no temporal information is considered. These HLFs are translated into sentential descriptions utilising the SimpleNLG, a template-based approach with a context free grammar.

4.2 Evaluation with ROUGE Metric

The complexity of evaluating video textual descriptions comes from the fact that defining the criteria is a challenging task. To evaluate our method, we examine the metrics commonly used for this purpose in machine translation. These metrics include the BLEU (bilingual evaluation understudy) (Papineni et al., 2002) and ROUGE (Recall Oriented Understudy for Gisting Evaluation) (Lin, 2004) metrics, among others. The BLEU score calculates precision on a word basis or n-grams, and for this reason is not suitable for our task of lingual video description, as has already suggest by (Mitchell et al., 2012) and (Das et al., 2013a).

By contrast, ROUGE score is an n-gram recall oriented measure of the information coverage of human annotation references compared to automatic summaries produced by a system. A higher ROUGE score denotes a higher degree of match between them. In general, a score of '1' indicates a perfect match whereas a score close to '0' means the match occurs in only a small portion of the data. Four different ROUGE scores are used in this experiment, ROUGE-1 (unigram) recall is the perfect option to compare descriptions based on predicted keywords only (Das et al., 2013a). ROUGE-2 (bigram) and ROUGE-SU4 (skip-4 bi-gram) scores are best to evaluate lingual video descriptions for coherence and fluency, whereas ROUGE-L scores depend on the longest common subsequence. ROUGE metrics are chosen for this study following (Das et al., 2013a) who used it to evaluate lingual video summarisation.

Table 2 present the average ROUGE scores achieved between the automatic descriptions produced by the baseline and our system, averaged over all twelve different human action categories, with respect to manual annotations. Manual annotations tend to be subjective as they depend on the annota-

Figure 2: Example of applying post-processing rules to the system-generated description of 'actionclipautoautotrain00463' video from the AnswerPhone category, with two shots.

		Baseline	Our approach
ROUGE-1	R	0.2480	**0.3513**
	P	**0.3443**	0.2474
	F	0.2749	**0.2806**
ROUGE-2	R	0.0532	**0.0737**
	P	**0.0801**	0.0500
	F	**0.0592**	0.0577
ROUGE-L	R	0.2353	**0.33365**
	P	**0.3275**	0.2354
	F	0.2609	**0.26689**
ROUGE-SU4	R	0.0939	**0.1526**
	P	**0.1745**	0.0951
	F	0.1064	**0.1098**

Table 2: ROUGE scores calculated for the baseline and our approach, with respect to hand annotations. For each ROUGE metric, the recall (R), precision (P), and F-measure (F) are averaged over all twelve categories from the NLDHA dataset.

tors perception and understanding. Moreover, this subjectivity might be affected by personal education level, interests, background and experiences. As a result, the ROUGE metric inevitably penalises many automatically generated sentences where these do not match the manual annotations, despite being technically correct.

Clearly, the best results were obtained by ROUGE-1, as our method involves an extended language vocabulary compared to the baseline. This richness comes from a varied set of verbs included along with their scene setting, especially when the language model is involved for the case of zero-shot action recognition. (*e.g.*When 'person' and 'TV' are detected in the scene without a connected verb, the language model will infer the verb 'watch' to complete the sentence.) Additionally, ROUGE-L results confirm the efficiency of our approach as

	Grammar	Correctness	Relevance
Baseline	3.40	3.40	2.25
Our approach	3.54	3.75	3.74

Table 3: Human evaluation for the baseline and our approach, with respect to three aspects: grammatical correctness, cognitive correctness, and relevance.

it captures similarity at sentence-level between the automatic generated descriptions and hand annotations. There is also an observable improvement for ROUGE-2 and ROUGE- SU4. This is not surprising since attributes (such as adjectives and prepositions) and co-reference enhance the quality of description by generating richer and less verbose descriptions. However, this kind of improvement in quality does not usually contribute considerably to the ROUGE score, which is based on n-gram comparisons.

4.3 Human Evaluation

The ROUGE metrics produce only a rough estimate of the informativeness of an automatically produced summary, as it does not consider other significant aspects, such as readability or overall responsiveness. To evaluate these types of aspects there is an urgent need for manual evaluation. For this task Amazon Mechanical Turk was used to collect human judgements of automatic video descriptions. We follow (Kuznetsova et al., 2012) and asked 10 Turk workers to rate video descriptions generated by the baseline and our description. Each worker watched each video and rated the description on a scale of 1 to 5, where 5 means 'perfect description', and 1 indicates 'bad description'.

The description rating was based on three different criteria: grammar, correctness, and relevance.

For both the correctness and relevance aspects, the video was displayed with its description. The correctness evaluates to what extent the textual description depicted the video semantic content, while the relevance rates if the sentence captures the most salient actions and objects. For the grammar correctness, only lingual descriptions were presented to the worker, without the video, to evaluate the sentence. Table 3 shows the results of human evaluation of both the baseline and our approach. It can be observed that our system improves on the baseline in all three aspects. However, the relevance score significantly outperforms the baseline with margin of 1.61. This indicates that our approach is able to describe much more semantic video content, especially in terms of activities, attributes and scene setting.

4.4 Discussion

The majority of previous works, including the baseline system, rely on the image-based detector deformable parts model (DPM) (Felzenszwalb et al., 2010) which is applied to each frame to augment a store of detected objects, without preserving any temporal dependency between video frames. As a result, the descriptions generated using this detector suffer from several weaknesses, mainly redundancy and lack of coherence. The redundancy issue basically results from applying the object detector at each frame without maintaining any temporal correlation; hence if the object changes its position gradually between frames it will be considered as a new detection.

Moreover, consistent co-reference of pronouns to visual objects across multiple sentences cannot be reliably identified for image-based detections, as prior information is required from the preceding frames to prove the previous detection. As a result, the generated description will be verbose, unnatural and contain irrelevancies. The Figure 3(c) shows an example of co-reference identification achieved successfully by the proposed system in '*she* is sitting next to *him*', while the baseline was unable to identify such information as its detection based on individual frames rather than tracking the detection over video frames and exploiting the temporal continuity. See Figure 3 for some examples of automatic video descriptions.

Generating elaborate textual descriptions demands more than action recognition and object detection. Identifying spatial and temporal relations between entities allows them to be mapped onto prepositions and adverbs in the output description. Figure 3(b) shows an example of improvement over the baseline as the proposed system was able to identify the scene layout by formalising spatial relations in 'a man is standing *next to* a car; while a woman is standing *to the right of* him'. Additionally, temporal relations are captured by the system in Figure 3(c) 'a woman is walking *toward* a man' and in Figure 3(a) 'a man is walking away from her' as this relation is calculated by comparing the distance between two objects over sequence of frames.

The proposed framework is applicable to any video genre with human actions and even if no human is detected, the video will be described based on detected non-human objects and scene setting. Although this framework produces a syntactically and grammatically correct description, the current immaturity of computer vision techniques can lead to false positive detections or missing information. As a result, the generated description can be inaccurate and mismatch the real action performed in the video sequences. There is a room for improvement, especially in object detections and their associated attributes, such as actions, colour and dress, which can significantly enhance the accuracy and quality of automatically generated description.

5 Conclusion

This paper has introduced a framework that produces textual descriptions of video based on extracted semantic video content. In an extensive experimental evaluation we show the improvements of our framework compared to the recent baseline frame-based video description system. The improvements are consistent among both automatic evaluation with ROUGE metrics and manual human evaluations of correctness and relevance. This improvement offered by the proposed system stems from the fact that the main sentence components are extracted by visually parsing the video content with respect to temporal information.

Figure 3: Sample of textual video descriptions along with their video shots from different categories from the NLDHA dataset.

References

Nouf Al Harbi and Yoshihiko Gotoh. 2015a. Describing spatio-temporal relations between object volumes in video streams. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Nouf Al Harbi and Yoshihiko Gotoh. 2015b. A unified spatio-temporal human body region tracking approach to action recognition. *Neurocomputing*, 161:56–64.

Nouf Al Harbi and Yoshihiko Gotoh. 2016. Natural language descriptions of human activities scenes: Corpus generation and analysis. pages 39–47.

Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, et al. 2012. Video in sentences out. *arXiv preprint arXiv:1204.2742*.

Juan Bekios-Calfa, Jose M Buenaposada, and Luis Baumela. 2011. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):858–864.

Pradipto Das, Rohini K Srihari, and Jason J Corso. 2013a. Translating related words to videos and back through latent topics. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 485–494. ACM.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013b. A thousand frames in just a few

words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Akshat Garg and Vishakha Choudhary. 2012. Facial expression recognition using principal component analysis. *Int. J. Sci. Eng. Res. Technol.*

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the 14th International Conference on Computer Vision (ICCV-2013)*, pages 2712–2719, Sydney, Australia, December.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014a. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014b. Creating Summaries from User Videos. In *ECCV*.

Patrick Hanckmann, Klamer Schutte, and Gertjan J Burghouts. 2012. Automated textual descriptions for a wide range of video events with 48 human actions. In *European Conference on Computer Vision*, pages 372–380. Springer.

Wen-Bing Horng, Cheng-Ping Lee, and Chun-Wen Chen. 2001. Classification of age groups based on facial features. *Tamkang Journal of Science and Engineering*, 4(3):183–192.

Muhammad Usman Ghani Khan, Nouf Al Harbi, and Yoshihiko Gotoh. 2015. A framework for creating natural language descriptions of video streams. *Information Sciences*, 303:61–82.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Coling*, volume 2, page 9.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.