Natural Language Descriptions of Human Activities Scenes: Corpus Generation and Analysis

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. Although the method of annotating this data with keywords is relatively well researched, the quality can be improved through describing videos with natural language. We are exploring approaches to generating natural language descriptions of inter-relations between human activities in a video stream. This paper focuses on creation of a dataset that can be used for development and evaluation. To this end a corpus of video clips, manually selected from the Hollywood2 dataset, and their natural language descriptions has been generated. Analysis of the hand annotation presents insights into human interests and thoughts. Such resource can be used to evaluate automatic natural language generation systems for video.

1 Introduction

Video synopses can be created by converting video summaries using natural language. They serve to generate a multimedia archive where video analysis, retrieval and summarisation can be developed. The majority of previous research, in particular for video description tasks, has relied upon short video clips. They typically presented one subject performing one action, hence a single sentence was often sufficient to annotate them. By contrast reality-based video scenarios incorporate various camera shots depicting a range of actions.

We are exploring approaches to generating natural language description for inter-relations of humans and their activities within video streams. The first step of the study was to create a dataset that could be used for development and evaluation, as we did not find publicly available resource that suitably considered the spatial and temporal relations between individual entities. Initially, from the Human Actions and Scenes dataset (Hollywood2 dataset[1]), 120 video segments were selected, 10 for each of the twelve categories. They were relatively long videos ranging from 1 to 3 minutes, selected based on a number of criteria, such as the number of camera shots and the variety of human actions. For selected video clips, a dataset was then created, comprising hand annotations with natural language descriptions. We refer to this dataset as NLDHA[2] Corpus.

The contributions of the work presented in this paper include the following two aspects:

- A total of 12 participants manually annotated this dataset in two ways: a brief synopsis (title) consisting of a single phrase or sentence, and a full explanation in everyday language, set out using a number of sentences.
- An action classification experiment based on hand annotations was performed to demonstrate the application of the corpus with natural language descriptions.

2 Related Work

There are a variety of corpora in the video processing studies, ranging from basic object recognition to analysis of complex scenes. Unfortunately most video corpora for visual event recognition are not suitable for evaluating their natural language description. For example the KTH dataset (Schuldt

[1] www.di.ens.fr/~laptev/actions/hollywood2

[2] 'NLDHA' stands for Natural Language Descriptions for Human Activities in videos.

Proceedings of the 5th Workshop on Vision and Language, pages 39–47,
Berlin, Germany, August 12 2016.

et al., 2004) and the Weizmann dataset (Blank et al., 2005) facilitate depicting events with a single human, thus there is no interaction with other individuals or objects. Recently a number of video corpora have been created, aiming at annotation with natural language. They are designed with certain prerequisites or constraints to fulfil the specific task or the purpose. Some of these corpora are reviewed in the following.

ACL2013 dataset[3]**.** A methodology was proposed by (Yu and Siskind, 2013) to learn word meanings from video that was coupled with sentences. A range of combined situations could be compiled into a dataset of 61 short filmed video clips, each with 3-5 seconds and 640×480 resolution at 40 fps (frames per second). Every clip was made up of a combination of a number of synchronous instances, which could involve a subset of up to four different entities: a chair, a garbage can, a backpack and a person. The corpus of 159 training examples coupled up videos with more than one sentence and sentences with more than one video — on average there were 2.6 sentences per video. Some of these video clips depicted non-human objects' activities without human presence, such as an airplane landing, which makes this dataset not suitable for our task.

TACoS Cooking dataset[4]**.** This dataset was created for addressing the issue of grounding sentences to describe actions in visual information extracted from videos (Regneri et al., 2013). 127 videos with 26 basic cooking tasks were included and 22 subjects were used for recording a corpus in the kitchen environment. 20 different textual descriptions were collected for each video, resulting in 2540 annotation assignments. This corpus was designed for the specific purpose of cooking and, as a result, all actions were centred on the kitchen environment, which makes it not suitable for a general video description task.

SumMe dataset[5]**.** SumMe was a new benchmark proposed for the task of summarizing video (Gygli et al., 2014). There were in total 25 videos included in the SumMe dataset, covering sports, events and holidays. The video length varied between 1 and 6 minutes. The study included a total of 41 participants (19 males and 22 females) that had different educational backgrounds, for summarizing the videos' visual content. Around 15 to 18 people summarised each video. Since there is no human present in some videos, this dataset is inappropriate for our task.

3 Corpus Generation

We have created the NLDHA Corpus of English language, describing 12 action classes from real-life video scenes, observed in the manually selected subset of the Hollywood2 dataset which was collected from 69 Hollywood films. This dataset was selected as it had realistic and generic video settings including human subjects with various activities, emotions and interactions with others. We have selected 10 video clips for each of 12 action classes, resulting in 120 video clips in total. The selected clips contained either (1) multiple camera shots of human activities to incorporate temporal and spatial association of human activities, or (2) a single shot consisting of a variety of actions, performed either by one or multiple persons. The intention was to develop a compact dataset to study approaches for translating video contents of human interaction and their temporal and spatial relations to natural language descriptions. For each of 120 video clips, NLDHA consists of 12 descriptions obtained via Amazon Mechanical Turk (MTurk)[6].

The majority of selected segments contained multiple camera shots, with 6 shots on average, varying between indoor and outdoor scene-settings. The total length of the selected clips was 225,000 frames, with a frame rate of 25 fps and the average length of 1875 frames for each video. Videos span between 1 and 3 minutes, with the average length of 75 seconds. Human interactions may be classified into two themes:

human-human interaction: This involves multiple humans, including categories such as 'FightPerson', 'HandShake', 'HugPerson', 'SitUp', 'Run' and 'Kiss'.

human-object interaction: A human performes some action with an object (*e.g.*, car, chair or dining table), such as '*driving*', '*sitting*' or '*eating*'. This includes the following categories: 'AnswerPhone', 'DriveCar', 'SitDown', 'StandUp', 'Eat' and 'GetOutCar'.

All categories involved human activities, expressions and emotions. A sequence of actions was

[3] haonanyu.com/research/acl2013

[4] www.coli.uni-saarland.de/projects/smile/page.php?id=tacos

[5] www.vision.ee.ethz.ch/gyglim/vsum

[6] www.mturk.com

performed by one person, depicted in one shot, whereas multiple shots presented relation and interaction between multiple humans. Some videos depicted humans' interaction with other objects in a variety of indoor and outdoor settings.

3.1 Collecting Textual Video Descriptions

Amazon Mechanical Turk (MTurk) was used to collect video descriptions. A Human Intelligence Task (HIT) was created and published on MTurk, using an adapted version of the annotation tool Vatic (Vondrick et al., 2013). For each video we collected 12 different textual annotations, leading to 1440 annotation assignments. Each annotator prepared manual descriptions for 120 video segments in two different types: title assignment (a single phrase) and full description (multiple sentences). A title, in some sense, can be considered as a summary provided in the most compact form, which includes the essential themes, or contents of a video in a short phrase. In contrast, full description is detailed and comprises of a number of sentences with in-depth description of objects, their activities and interactions. In the rest of this paper they are referred to as 'hand annotation'. A valuable resource for text-based video retrieval and summarisation can be created through the combination of titles and full descriptions.

For each assignment one video was shown to the annotator, who was then requested to provide a title for the video in one phrase, highlighting the main theme and explaining human activities observed in the video. The annotator was also asked to provide a description of minimum 5 and maximum 15 complete English sentences for explaining the events in the video. In order to help annotators understand the task, they were presented with a sample video segment, as well as possible textual annotations, *i.e.*, a title and a complete description. Instructions were provided, allowing an open vocabulary, meaning that annotators had the freedom to use any English word. However annotators were asked not to use (1) computer codes or symbols, (2) proper nouns (*e.g.*, person's name), and (3) information identified through audio, since they could affect the quality of descriptions for semantic video content.

4 Corpus Analysis

With 12 annotators describing each of 120 videos, there are 1440 documents in this corpus. The total number of words is 67080, hence the average length of one document is roughly 47 words. There are 5136 unique words and 2336 keywords consisting of nouns and verbs. Figure 1 presents six annotation examples for one of video clips from the 'DriveCar' category. This video segment consisted of four different shots depicting multiple actions performed by four humans, with the two main activities, '*smashing*' and '*driving*'.

Hand annotation 1
(title) Furious man crushing a window on the car;
(description) Furious man is crushing window on the car with iron stick and screaming; After that, we see him and the other two men driving in the car; All except the driver are eating sandwiches; Then we see driver sticking nails in wooden lath;

Hand annotation 2
(title) Beating a car and running;
(description) Angry man begins beating up a car; He breaks the windshield and windows; Then there are three men riding in another car; They are eating and riding somewhere; Then a man is beating nails into a board;

Hand annotation 3
(title) Smashing a car window;
(description) A man is smashing the window of a parked car with a sledge hammer at night; Next, the man who was speaking on the phone is driving a car with two other men as passengers at night; Later, another man is speaking on the phone while hammering nails in a board;

Hand annotation 4
(title) Broken windshield of a car;
(description) A man breaks the front windshield of a car using a rod; Then the same car is driven by three men; Later, one among them makes a call inside the car;

Hand annotation 5
(title) A windshieldless car driving;
(description) A man smashing in a windshield with a bat and later three men go for a drive in the car;

Hand annotation 6
(title) Car wrecking;
(description) A man wrecks a car and then people drive in it;

Figure 1: A montage of a 3-minute video segment and six sets of the hand annotations. This clip was extracted from the 'DriveCar' category in the Hollywood2 dataset, 'actionclipautoautotrain00094', depicting a sequence of actions performed by four humans in an outdoor scene.

The hand annotations have been made in two types: 'title' and 'description'. A title often consists of only a couple of words that do not constitute a complete sentence. Verbs are often used to express the main theme of the video, *e.g.*, '*family eating dinner*', '*men fighting*' and '*three people driving*'. The average length of titles is three words. An extensive analysis on titles indicates that, for each video, the same theme was identified by most annotators, though there were differences between them in the words used to express the theme. Figure 1 clearly illustrates that six

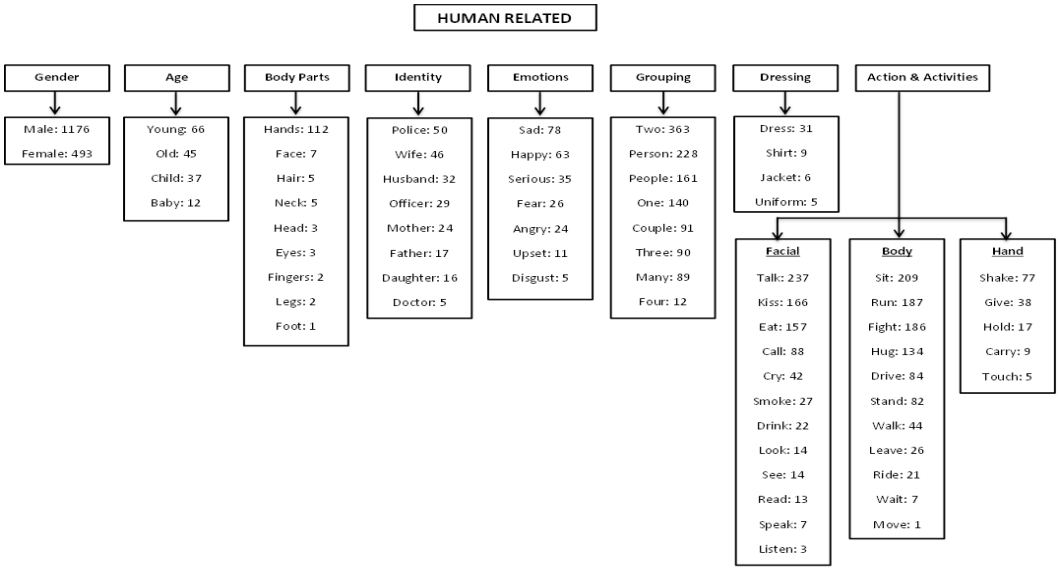

Figure 2: Human related features and their occurrences found in the hand annotations.

annotators were expressing the same theme using different words — '*crushing*', '*beat up*', '*smash*', '*break*' and '*wreck*'.

On the other hand, full descriptions on average contain four to six phrases or sentences; typically each camera shot is described by one sentence. Most sentences are concise, ranging between six and eight words. Descriptions for human, gender, emotion and actions, with their temporal order, are commonly observed. Minor details for objects, dressing and location are only occasionally stated, unless these objects participate in the event. Annotations vary in a wide range from highly abstract to very detailed descriptions, although they typically preserved the temporal order of activities performed in the video clip. The amount of detail included in full descriptions can be observed in examples presented in Figure 1. They vary between the very compact (*e.g.*, annotation 6), to the very detailed (*e.g.*, annotation 3). Nevertheless almost all annotations maintain the same temporal order of activities performed in the video.

4.1 Human Related Features

Figure 2 illustrates the human-related information that is highlighted in the hand annotations. Full attention was paid to the human presence in the video by the annotators, in particular gender specification for female and male are most frequently observed. Note that in the '*female*' category, related words indicating female, such as '*lady*' and '*woman*' are also included; and so are in the '*male*' category. This supports that humans and their attributes which identify as high level visual features (HLFs) are the most important and interesting information for annotators. By contrast some factors such as identity (*e.g.*, '*police officer*', '*father*') and age information (*e.g.*, '*young*', '*old*', '*child*') are not observed very often. Human body parts have mixed occurrences, ranging from high ('*hand*') to low ('*foot*').

Six basic emotions were presented in (Ekman, 1992); they relate to the most frequent facial expressions, including fear, anger, sadness, surprise, disgust and happiness. Another interesting feature is dressing; when an individual has dressed in a unique manner, for instance wearing a formal suit, an army, a police uniform or a coloured jacket, it was described; otherwise dressing information was not stated frequently. Scenes with multiple humans were also very common, and therefore, grouping information were frequently stated. Human activities were identified through the involvement of body parts, including actions such as '*walking*', '*running*', '*sitting*', '*fighting*' and '*standing*'. It was also observed among the majority of annotators that they liked to describe using general terms (*e.g.*, '*female*') rather than their specific identities (*e.g.*, '*doctor*').

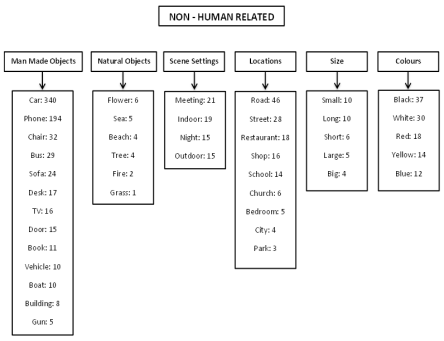

Figure 3: Non-human features and their occurrences in the hand annotations.

4.2 Objects and Scene Settings

Figure 3 illustrates the hierarchy of visual features that are not found in Figure 2. Many of them denote artificial objects, and interaction between humans and these objects are stated to complete activities, *e.g.*, '*man is sitting on chair*', '*he is driving a car*' and '*she is talking on the phone*'. Other important information is location (*e.g.*, '*restaurant*', '*shop*', '*school*'), which identifies object's position in the scene (*e.g.*, '*people are eating in the restaurant*', '*there is a car on the road*').

When identifying individual high level features (HLFs), colour information often plays an essential role — *e.g.*, '*she is wearing a white uniform*', '*a man in a black shirt is walking with a woman with a green jacket*'. Considering the great number of colour occurrences, it is evident that humans have an interest in observing the colour scheme in visual scenes, along with the objects. We are able to observe individual annotators' interest in foreground/background. Some annotators also paid attention to outdoor/indoor scene settings. Details for prominent objects in a visual scene was demonstrated by some annotations — *e.g.*, '*two boys are seated on a small boat*', '*a lady with long hair is walking on the road*'. Natural objects were rarely described in the hand annotations.

4.3 Spatial Relations

Visual scenes in filming are best described in terms of spatial relations, which can define how objects are located spatially in relation to some reference object. In a video stream this reference object is usually in the foreground. The competent use of prepositions, such as *on*, *at*, *inside* or *above*, can facilitate the creation of smooth and concise descriptions when presenting the spatial relations between objects. For example '*three people are swimming in the canal*' provides more descriptive detail than '*three people are swimming*' and '*there is a canal in the background*' separately. There are a variety of expressions that can be used to gain accurate spatial representations, *e.g.*, direction ('*left*', '*under*'), distance ('*far*', '*near*'), or topology ('*touch*', '*inside*') (Cohn et al., 2008).

in: 653; on: 335; with: 235; at: 121; between: 36; around: 26; behind: 25; touch: 23; middle: 21; together: 20; inside: 17; far: 16; in front of: 13; beside: 11; on the right: 10; on the left: 8, near: 6; under: 5; in the middle: 3

Figure 4: List of frequent spatial relations and their frequency counts, manually collated from the hand annotations.

A list of the most frequent words in the corpus concerning spatial relations are presented in Figure 4. Frequent occurrences of these words indicate people's regular use when describing visual scenes. Semantics of the visual scenes are better understood through the use of these words with which we are able to identify connections between various HLFs. For various reasons they had to be manually counted. Firstly, some words in the list may have a multitude of alternative uses in addition to spatial relations. The following three phrases demonstrate how the word '*in*' can be used for different purposes: '*three people are sitting in a car*' represents a spatial relation, whilst '*the dog in the last shot*' depicts a relationship between various scenes, and '*two people in a dialogue*' augments the ease with which the description can be read. Secondly, the spatial word can be a preposition by itself; *e.g.*, '*in*' or syntactically overlapped with another preposition such as '*three persons are talking in front of shops at night*'. Finally, there are some preposition words that can be used for both spatial and temporal relations; *e.g.*, '*at*' in the following example, '*a man is smashing the window of a parked car with a sledge hammer at night*' presents the temporal relation, whereas '*at*' in '*there are three people eating dinner at home*' indicates the spatial relation.

4.4 Temporal Relations

When something happens, temporal expressions, such as *before*, *long*, *awhile* or *during*, describe

Single human
then: 125; after: 60; afterwards: 44; before: 42; later on: 32; throughout: 32; start: 27; end: 25; next: 25; finish: 25;

Multiple humans
while: 87; meeting:71; during: 27; overlap: 12; meanwhile: 12; throughout: 12; then: 11; equals: 4,

Figure 5: List of frequent temporal relations and their frequency counts in the NLDHA Corpus.

the duration or how often it occurs (Pustejovsky et al., 2003). Temporal and spatial relations are combined in videos as time series data using highly sophisticated multi-dimensional contents. A complete video sequence is created by linking individual scenes. Annotators use temporal relations to combine the narratives for a sequence of scenes and produce a complete account of the video content. In the following example, three separate scenes can be connected using two temporal relations, *then* and *later*:

> '*A man and woman are talking and the woman walks out of the house;* ***then*** *she sees him through the door as he is passing in the street;* ***later****, another man enters the house.*'

A total of thirteen relations (*overlaps*, *overlapped-by*, *starts*, *started-by*, *meets*, *meet-by*, *finishes*, *finished-by*, *equals*, *after*, *before*, *contains* and *during*) make up a temporal logic, as identified in (Allen and Ferguson, 1994), who also proposed that scenarios could be more often described using time intervals than time points. Analysis of the NLDHA Corpus indicates that temporal relations can be classified into two types: activities of a single human and multiple humans interacting with each other. Figure 5 presents a list of the most frequent temporal relations found in the hand annotations. Clearly keywords, connecting numerous human activities, are important. According to Allen's algebra (Allen and Ferguson, 1994), '*meet*' and '*met by*' are keywords, indicating important temporal relations. This kind of relation occurs frequently in meeting scenes where there are multiple humans present, thus a specific action is performed once another action is completed. '*While*' is also a commonly used temporal keyword as it describes actions carried out simultaneously, *e.g.*, '*a man is eating while his friend is drinking*'.

Our observation indicates that, for activities by a single human, temporal relations are typically used in the chronological order of actions, *e.g.*, '*a man comes into the room a little awkwardly; then he sits on the chair*'. On the other hand, for the multiple humans scenes, corpus analysis shows that the annotators were likely to pay much more attention to the actions carried out simultaneously by different people, rather than describing individual human activities. Some of video scenes incorporated both types, thus their occurrences had to be counted manually.

4.5 Similarity between Descriptions

Cohen's kappa coefficient (κ) is widely used for calculating the inter-annotator agreement (Cohen, 1960). However, for measuring the similarity in the NLDHA Corpus, a kappa coefficient may not be suitable because of the large variation in the description length among individual annotators. For such situation, a so-called cosine similarity may be more effective because it works independent of document lengths as one of its important properties. The similarity between two documents can be quantified as the cosine of the angle between the vectors when the documents are represented as term vectors.

Let $D = d_1, \ldots, d_n$ denote a set of documents and $T = t_1, \ldots, t_m$ be a set of distinct terms occurring in D. A document is then represented as an m-dimensional vector $\vec{t_d}$. Let $tf(d,t)$ stand for the frequency of term $t \in T$ in document $d \in D$. Then the vector representation of a document d is given by

$$\vec{t_d} = (tf(d,t_1), \ldots, tf(d,t_m)) \qquad (1)$$

and the cosine similarity is defined by

$$SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|} \qquad (2)$$

where $\vec{t_a}$ and $\vec{t_b}$ are m-dimensional vectors over the term set $T = t_1, \ldots, t_m$. The numerator represents the dot product of two vectors, while the denominator is the product of their Euclidean lengths. Each dimension is used for representing a term with its weight in the document which is non-negative, due to which the cosine similarity is non-negative and bounded between $[0, 1]$. It is 0 where two documents are totally different, and 1 where they are identical.

To evaluate the similarity between hand annotations, a number of standard text processing filtering techniques are applied. The first is the removal of stop words (Flood, 1999), which are non-descriptive for the purpose of these documents.

The second measure involves stemming, which is reducing words into their base forms using the Porter stemmer (Porter, 1980). Finally it is usually helpful to minimise the vocabulary by substituting words with their common synonyms without affecting meaning, which can be achieved by using the NLTK WordNet interface[7]. Synonyms will reduce the annotators' variation and subjectivity caused by their use of different words for the same concept, and will also increase the occurrence of significant collocations.

The average similarity scores within 12 hand annotations for each of 120 video across 12 categories are shown in Table 1. Individual description scores were used for calculating the average, which was compared with the remaining descriptions in the same category. They were calculated in three conditions: (A) raw hand annotations, (B) applying Porter Stemmer and removing stop words, without replacing synonyms, and (C) without removing stop words, but applying Porter Stemmer and replacing synonyms. The table indicates that condition (C) resulted in the better similarity. In other words, the similarity has increased by replacing some words with their synonyms, indicating that we are expressing the same concept using different terms.

Table 1 also shows that the similarity scores for 'DriveCar', 'AnswerPhone' and 'Eat' categories were higher than the rest. Each of these three categories appeared to have some common factors among hand annotations, resulting from existence of important objects associated with humans and their actions, such as a car, a phone, and a dining table. Most annotators paid attention to such objects, hence common concepts were used for their description, leading to higher similarity scores than others. Conversely for the rest of categories, a broader range of concepts were incorporated in their hand annotations, although they still maintained the similarity by focusing on the same actions (thus using the same verbs).

5 Video Classification Experiments

This section uses an action classification task for demonstrating the application of the NLDHA Corpus with natural language descriptions.

[7] www.nltk.org

	(A)	(B)	(C)	Average
AnswerPhone	0.5294	0.5236	0.5446	0.5325
DriveCar	0.5564	0.5587	0.5632	0.5594
Eat	0.5272	0.5386	0.5386	0.5348
FightPerson	0.4010	0.4104	0.4245	0.4119
GetOutCar	0.4679	0.4607	0.4707	0.4664
HandShake	0.3955	0.4034	0.4187	0.4058
HugPerson	0.4036	0.4216	0.4236	0.4162
Kiss	0.3868	0.4065	0.4187	0.404
Run	0.3996	0.4056	0.4076	0.4042
SitDown	0.3925	0.4065	0.4158	0.4049
SitUp	0.3898	0.3952	0.4023	0.3958
StandUp	0.4043	0.4074	0.4274	0.4130

Table 1: Similarity scores within 12 hand annotations using the cosine similarity. For each class, scores are calculated in three conditions: (A) raw hand annotations; (B) applying Porter Stemmer and removing stop words, without replacing synonyms; (C) without removing stop words, but applying Porter Stemmer and replacing synonyms.

5.1 Experimental Setup

Textual document features can be expressed through *tf-idf* scores (Dumais et al., 1998). The importance of a term t within a particular document d can be measured by

$$tfidf(t,d) = tf(t,d) \cdot idf(d) \tag{3}$$

The term frequency $tf(t,d)$ is given by

$$tf(t,d) = \frac{N_{t,d}}{\sum_k N_{k,d}} \tag{4}$$

where the number of occurrences of t in d is presented by $N_{t,d}$, while the denominator is the size of the document $|d|$. Further, the inverse document frequency $idf(d)$ is

$$idf(d) = \log \frac{N}{W(t)} \tag{5}$$

where N is the total number of documents in the corpus and $W(t)$ is the total number of documents containing the term t. A term-document matrix is presented by $T \times D$ matrix $\{tfidf(t,d)\}$.

When conducting the experiment, stop words were removed and stemming was applied. For the action classication task, the most frequent 1000 words were used. We applied the Naive Bayes probabilistic supervised learning algorithm from the *Weka* machine learning library (Hall et al., 2009). Ten-fold cross validation was performed and the outcome was measured using precision, recall and F1-measure.

	Precision	Recall	F1-measure
AnswerPhone	0.836	0.850	0.843
DriveCar	0.803	0.850	0.826
Eat	0.855	0.883	0.869
FightPerson	0.786	0.858	0.821
GetOutCar	0.791	0.725	0.757
HandShake	0.817	0.783	0.800
HugPerson	0.921	0.775	0.842
Kiss	0.783	0.783	0.783
Run	0.939	0.900	0.919
SitDown	0.623	0.675	0.648
SitUp	0.686	0.583	0.631
StandUp	0.483	0.575	0.525
Average	0.777	0.770	0.772

Table 2: Outcomes for the action classification experiment using the Naive Bayes classifier.

5.2 Results

Table 2 presents the outcomes of the monitored classification assessment using *tf-idf* characteristics. The F1 scores for certain categories, such as 'AnswerPhone', 'Eat', 'DriveCar' and 'Run', were greater than some others. For these categories, description concerning humans and the important objects (*e.g.*, dining table, car, phone) were found in most of hand annotations thus classification was not too difficult. In general, F1 scores were higher for categories where human's interaction with an object was evident.

In comparison some categories, such as 'SitDown', 'SitUp' and 'StandUp', had the substantially lower F1 scores than the rest. There were two potential reasons why the annotators did not pay sufficient attention to these actions. Firstly, these actions were performed very quickly in the context of some videos. For example, when a person sat down or stood up during an eating scene, the annotators would have focused on eating (rather than sitting down or standing up) in their description. Secondly, these actions were often overlapped with another action by different humans in the video, which the annotators might have found more important for description. Overall outcome of the classification experiment indicates that the corpus is a reliable tool for assessing natural language description of video streams.

6 Findings from the Corpus Analysis

The corpus is important for the following reasons: (1) limiting this study to a clearly defined and manageable domain; (2) identifying the most important HLFs that should be extracted by image processing techniques in order to describe semantic content of videos; and (3) providing development and test dataset. They should also serve as the ground truths for evaluation.

We have obtained a few insights into the dataset based on the analysis of hand annotations. Annotators are most interested in presence of humans and their attributes in videos, especially their gender, emotions, actions and their interaction with other humans and objects. Based on these observations, we derive a list of HLFs for automatic extraction, consisting of humans and their age, gender, emotion, action, the number of humans, objects, scene setting, spatial and temporal relations. Hand annotation of one visual scene can vary substantially due to the subjectivity of individuals. It can be argued that the dissimilarity lies in the choice of words and that the similarity can be found in the contents that are described. Hand annotations can be used as a reference to evaluate the information content of machine generated descriptions.

7 Conclusion and Future Work

We have developed a new corpus, consisting of natural language descriptions for video data. 12 annotators produced a title and a full description for each of 120 video segments, derived from a subset of Hollywood2 dataset. They are much longer streaming videos than existing ones that were previously annotated with natural language descriptions. As a consequence each segment contains numerous instances of a variety of actions that may overlap in time and occur at various spatial positions within the frame, hence providing a challenge in processing the contents spatially and temporally. The accompanied annotation delineates not only a type of action but also its spatial position and temporal extent. Analysis of this corpus presents insights into human interests and thoughts in such visual scenes. Important visual entities have been identified, aiming at future use for automatic extraction of visual features, which are then used for automatic generation of natural language descriptions for that visual scene.

Acknowledgements. The first author would like to thank Taibah University, Medina, Saudi Arabia for funding this work as part of her PhD scholarship program.

References

James F Allen and George Ferguson. 1994. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579.

Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. 2005. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Anthony G Cohn, Jochen Renz, et al. 2008. Qualitative spatial representation and reasoning. *Handbook of knowledge representation*, 3:551–596.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Barbara J Flood. 1999. Historical note: the start of a stop list at biological abstracts. *Journal of the American Society for Information Science*, 50(12):1066–1066.

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

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.

C. Schuldt, I. Laptev, and B. Caputo. 2004. Recognizing human actions: a local SVM approach. In *Proceedings of ICPR*, volume 3, pages 32–36.

Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *ACL (1)*, pages 53–63.