
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/112427/

Version: Accepted Version

Article:
Khan, M.U.G. and Gotoh, Y. (2017) Generating natural language tags for video information management. Machine Vision and Applications, 28 (3-4). pp. 243-265. ISSN: 0932-8092

https://doi.org/10.1007/s00138-017-0825-7

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



Generating Natural Language Tags for Video Information Management

Muhammad Usman Ghani Khan[1]

Department of Computer Science, University of Engineering & Technology, Lahore, Pakistan

Yoshihiko Gotoh[2]

Department of Computer Science, University of Sheffield, United Kingdom

Abstract

This exploratory work is concerned with generation of natural language descriptions that can be used for video retrieval applications. It is a step ahead of keyword based tagging as it captures relations between keywords associated with videos. Firstly we prepare hand annotations consisting of descriptions for video segments crafted from a TREC Video dataset. Analysis of this data presents insights into human's interests on video contents. Secondly we develop a framework for creating smooth and coherent description of video streams. It builds on conventional image processing techniques that extract high level features from individual video frames. Natural language description is then produced based on high level features. Although feature extraction processes are erroneous at various levels, we explore approaches to putting them together to produce a coherent, smooth and well phrased description by incorporating spatial and temporal information. Evaluation is made by calculating ROUGE scores between human annotated and machine generated descriptions. Further we introduce a task based evaluation by human subjects which provides qualitative evaluation of generated descriptions.

Keywords: video information management, video annotation, natural language generation

1. Introduction

In recent years video has established its dominance in communication and has become an integrated part of our everyday life ranging from hand-held videos to broadcast news video (from unstructured to highly structured). There is need for formalising video semantics to help users gain useful information relevant to their interests. One approach to explaining semantics and contents of videos is to convert them into some

[1]Corresponding author. email: usmanghanikhan@gmail.com (M.U.G. Khan).
[2]email: y.gotoh@dcs.shef.ac.uk (Y. Gotoh).

other modality such as text. It is not a difficult task for humans to describe a video using their language. On the other hand, machines can only identify and recognise some objects and certain activities. Most of the previous studies were related to semantic indexing of video using keywords (10, 14). However it is often difficult with keywords alone to represent relations between various entities and events in video. Natural language description of videos can be desirable as it is more human friendly and is able to capture relationships between keywords, thus clarifying the context between individual keywords. They can guide generation of summaries by converting a video sequence to natural language. They can provide basis for generating a multimedia repository for video analysis, retrieval and summarisation tasks.

We start the work with manual development of a dataset, consisting of natural language descriptions of video segments crafted from a small subset of TREC Video[3] data. In a broad sense the task may be considered one form of machine translation as it translates video streams into textual descriptions. To date the number of studies in this field is relatively small partially because of lack of appropriate dataset for such a task. Another obstacle may be inherently larger variation for descriptions that can be produced for videos than a conventional translation from one language to another. Indeed humans are very subjective when annotating video streams, *e.g.*, two humans may produce quite different descriptions for the same video. Based on these descriptions we are interested to identify the most important and frequent high level features (HLFs); they may be 'keywords', such as a particular object and its position/move, used for a semantic indexing task in video retrieval.

This work encompasses two disciplines, namely image processing and natural language processing. Image processing techniques lead to identification of HLFs such as humans, objects, their moves and properties (*e.g.*, gender, emotion and action) (38). In this work a human is considered as a subject, performing some action, and objects may be affected by human actions or activities. Natural language processing deals with merging these HLFs into syntactically and semantically correct textual presentations, which can help a user in understanding the visual scene and summarising the video contents. The frame based natural language generation procedure results in many identical descriptions produced from adjacent frames. Hence simple concatenation of descriptions may lead to redundancy, lacking coherency. Additionally visual feature extraction processes are erroneous at various levels. We explore approaches to putting them together to produce a coherent, smooth and well phrased description by incorporating spatial and temporal information. Evaluation is made by calculating ROUGE scores between human annotated and machine generated descriptions. Further we introduce a task based evaluation by human subjects which provides qualitative evaluation of generated descriptions.

1.1. Related Work

On the whole, previous approaches followed the similar strategy for converting images to natural language. To begin with, the image content was represented by features, such as colour information (33, 41, 51), texture (41, 51), and detected edges

[3]http:trecvid.nist.gov

(33). Image features were then replaced with an abstract representation, essentially a set of description words based on a visual-to-textual representation dictionary. For certain applications, objects were detected and recognised using some prior knowledge to supply higher level features (33, 41, 51, 1).

Thomason et al. (41) presented description of objects based on semantic relations. For example, '*orange ball*' indicated a single object using the relationship between orange and ball while isolated words '*orange*' and '*ball*' presented two separate objects, an orange and a ball, rather than a single object (an orange ball). Initially a dictionary of objects was created based on image signature consisting of features such as object's colour, texture, name and category. Secondly images were segmented into regions and corresponding signatures were fetched from the database by comparing the region features with entries in the dictionary. Finally description was generated based on the retrieved signatures using manually defined templates.

Rohrbach et al. (33) presented a method for describing human activities based on a concept hierarchy for actions. They described head, hands and body movements using natural language texts. Firstly human poses and head moves along with their trajectories were identified in the form of numerical values. Secondly numerical values were converted into actions, such as '*enter*', '*carry*', '*turn*', '*exit*', based on a manually created concept dictionary. Finally these actions were combined to generate natural language descriptions using predefined grammars. Kojima et al. (21) further improved their method by incorporating more objects and interaction between human and non human objects. They also extended the concept hierarchy of actions related to human body and their interaction with other objects in office environments.

For a traffic control application, Yang et al. (50) investigated automatic visual surveillance systems where human behaviour was represented by scenarios, consisting of predefined sequences of events. A scenario was automatically translated into a text by analysing contents of the image over time, deciding on the most suitable event. Lee et al. (24) introduced a framework for semantic annotation of visual events in three steps; image parsing, event inference and language generation. Instead of humans and their activities, they focused on object detection, their inter-relations and events in videos. Baiget et al. (6) performed human identification and scene modelling manually and focused on human behaviour description for crosswalk scenes.

Yao et al. (51) introduced a framework for video to text description which was dependent on the significant amount of annotated data. Main building blocks of this framework were an image parser, visual knowledge representation, the semantic web and a text generation modules. Firstly images were hierarchically decomposed into their constituent visual patterns and presented using an And-Or graph. Secondly graphs were converted into structured representations with specified semantic relations (*e.g.*, categorical, spatial and functional relations) using a visual knowledge database. Finally, aided by the semantic web, natural language descriptions were generated using templates or grammars-that were manually defined for specific applications such as video surveillance. Both the parser and visual semantic representation were built on a large scale, manually annotated groundtruth image database.

The several approaches outlined above were able to generate grammatical natural language sentences for images or videos by analysing the image content. They typically relied on the large amounts of manually created resources, including (1) annotation of

an image database for a training purpose, (2) construction of a visual-textual translation dictionary or ontology, and (3) engineering of application specific sentence templates or grammars. Unfortunately most of such resources cannot be reused in cross domains or applications. The amount of manual effort is an important factor however it is costly and time consuming.

More recently the research community has shifted its focus towards exploiting textual data and natural language processing techniques for generating image descriptions. Main interest for combining vision and natural language is to investigate any significant improvement towards producing readable and descriptive sentences compared to naive strategies that use vision alone. Li et al. (25) proposed a three-step framework for generating textual descriptions of images. Image processing was the first step, consisting of identification of objects, their visual attributes and their spatial relationships. These three types of visual outputs were presented in the form of tuples (one triple for each pair of detected objects). Finally smooth and well phrased sentences were generated using web-scale n-grams that provided the frequency count of each possible n-gram sequence for $1 \leq n \leq 5$. Yang et al. (49) presented a framework for converting static images to textual descriptions by predicting nouns, verbs and prepositions that made up the core sentence structure. Initially nouns and scenes were detected using image processing methods. Secondly, to estimate the verbs and prepositions, a language model was trained from the English Gigaword corpus. A hidden Markov model was used for sentence generation, with sentence components as the hidden nodes and images as the emissions. In (18), we presented an approach for creation of textual descriptions for video sequences. Main focus of that work was accommodation of errors generated during image processing steps. This work is an extension is a sense that it provides analysis of human generated descriptions for a specific dataset and then based on those findings we provide a complete framework for generation of natural language tags for complete video sequences.

Donahue et al. (9) introduced a long-term recurrent convolution network (LRCN) for visual recognition and description which combines layers and long-range temporal recursion and is end-to-end trainable. Their work was straightforward to fine-tune end-to-end for any computer vision application. Further, their work was not confined to fixed length inputs or outputs allowing simple modeling for sequential data of varying lengths, such as text or video. Our approach is rule based which is simple, easier to implement and more modular; The model based approaches needs to be trained for any new objects that occur in the unknown dataset. Finally, they evaluated their dataset for only one category of videos while our work shows results against seven video categories.

Guadarrama et al. (15) presented an approach which works on out-of-domain actions which do not require training of the exact activities. If the system is unable to find an accurate prediction for a pre-trained model, it will give a less specific answer. A language model from web-scale text corpora is learned and used it as a prior on triplets to infer verbs missing from the vocabulary. Though, the activities and objects list is broad, still the training data for each label is scarce or unavailable. In addition, large vocabulary video activity description presents challenges such as modeling dynamics and actor-action object relationships from limited data. Finally, large vocabulary also tends to have the problem of polysemy and ambiguity.

Rohrbach et al. (34) generated a rich semantic representation of the visual content that includes objects and their relevant activities. Then a Conditional Random Field (CRF) model was employed to predict the relationships between different components of the visual input. Generation of natural language was considered as a machine translation problem that uses the semantic representation as source language and the generated sentences as target language. Again the applicability of their work was limited to only Kitchen scene settings, while our approach works on diverse settings like indoor, outdoor and traffic. The language model used in that work was limited to n-grams whereas in our approach we have used combination of language model, paraphrasing and removal of redundant sentences by using greedy string tiling algorithm to refine the description.

Tan et al. (40) extracted three kinds of audio-visual features from video streams; 2D static SIFT, 3D spatial-temporal interest points (STIPs), and MFCC audio descriptors. These features extract local information from the image. One of the key problem in dealing with local features is that the numbers of feature points in each image would differ hence complicating the process of comparison among the images. Our approach extracts individual high level features (HLFs), such as for humans: age, gender, emotions and actions are mined. Classifiers are applied to 10-second clips of videos to generate audio-visual concept classification scores. In our approach the video clips are 10 to 30 seconds in duration and deals with single and multiple activities. Finally, there is no utilization of natural language processing (NLP) techniques for generating well-designed textual descriptions while in our approach we have used a language model, paraphrasing and removal of redundant sentences.

Thomason et al. (42) proposed a factor graph model to perform content selection by integrating visual and linguistic information for selecting the best subject-verb-object-place description of a video. This approach seems to be lacking in image processing part, *i.e.*, in detecting activities, objects, and scenes with high precision. Further, video clips are 10 to 25 seconds in duration and typically consist of a single activity. This needs to be extended for multiple activities. Our approach is robust enough as it works for single and multiple activities.

This work contrasts to previous approaches in several aspects: first, sentences are generated from scratch, instead of retrieving (11, 27) or summarising existing text fragments associated with an image (3, 7). Second, textual descriptions are generated for specific and real contents of videos, whereas related (but subtly different) work in automatic caption generation created news text (12) or encyclopedic text (2) that was contextually relevant but not closely pertinent to the specific content of images. Third, many studies in image (video) to text conversion required domain specific hand-written grammar rules (51). This work also focuses on a relatively narrow domain, however we present an idea to produce a description even if a video is outside of that domain. Fourth, we allow for some creativity in the generation process which produces more humanlike descriptions than a closely related recent approach that derived annotation directly from computer vision inputs (22). Fifth, we presented our work for variety of scene settings instead of single scene setting (9, 34). Finally, many recent studies concerned description of a single image frame (36, 48). In this work, textual expressions are generated for a sequence of video frames, resulting in richer descriptions.

2. Corpus Development

We are exploring approaches to natural language descriptions of video data. The step one of the study is to create a dataset that can be used for development and evaluation because we could not identify a suitable dataset. Textual annotations are manually generated in three different flavors, *i.e.*, selection of HLFs (keywords), title assignment (a single phrase) and full description (multiple phrases). Keywords are useful for identification of objects and actions in videos. A title, in a sense, is a summary in the most compact form; it captures the most important content, or the theme, of the video in a short phrase. On the other hand, a full description is long, comprising of several sentences with details of objects, activities and their interactions. Combination of keywords, a title, and a full descriptions will create a valuable resource for text based video retrieval and summarisation tasks. Finally, analysis of this dataset provides an insight into how humans generate natural language description for video (Section 3).

2.1. Categories

In this study we select video clips from TREC Video benchmark. They include categories such as news, meeting, crowd, grouping, indoor/outdoor scene settings, traffic, costume, documentary, identity, music, sports and animals. The most important and probably the most frequent content in these videos appears to be a human (or humans), showing their activities, emotions and interactions with other objects. We do not intend to derive a dataset with a full scope of video categories, which is beyond our work. Instead, to keep the task manageable, we aim to create a compact dataset that can be used for developing approaches to translating video contents to natural language description.

Annotations were manually created for a small subset of data prepared form the rushes video summarisation task and the HLF extraction task for the 2007 and 2008 TREC Video evaluations. It consisted of 140 segments of videos — 20 segments for each of the following seven categories:

Action videos: Human posture is visible and human can be seen performing some action such as 'sitting', 'standing', 'walking' and 'running'.

Close-up: Human face is visible. Facial expressions and emotions usually define mood of the video (*e.g.*, happy, sad).

News: Presence of an anchor or reporters. Characterised by scene settings such as weather boards at the background.

Meeting: Multiple humans are sitting and communicating. Presence of objects such as chairs and a table.

Grouping: Multiple humans interaction scenes that do not belong to a meeting scenario. A table or chairs may not be present.

Traffic: Presence of vehicles such as cars, buses and trucks. Traffic signals.

Indoor/Outdoor: Scene settings are more obvious than human activities. Examples may be park scenes and office scenes (where computers and files are visible).

Each segment contained a single camera shot, spanning between 10 and 30 seconds in length. Two categories, 'Close-up' and 'Action', are mainly related to humans' activities, expressions and emotions. 'Grouping' and 'Meeting' depict relation and interaction between multiple humans. 'News' videos explain human activities in a constrained environment such as a broadcast studio. Last two categories, 'Indoor/Outdoor' and 'Traffic', are often observed in surveillance videos. They shows humans' interaction with other objects in indoor and outdoor settings.

2.2. Annotation Process

A total of 13 annotators were recruited to create texts for the video corpus. They were undergraduate or postgraduate students and fluent in English. It was expected that they could produce descriptions of good quality without detailed instructions or further training. A simple instruction set was given, leaving a wide room for individual interpretation about what might be included in the description. For quality reasons each annotator was given one week to complete the full set of videos.

Each annotator was presented with a complete set of 140 video segments on the annotation tool. For each video annotators were instructed to provide

- selection of high level features (*e.g.*, male, female, walk, smile, table);
- a title of one sentence long, indicating the main theme of the video;
- description of four to six sentences, related to what are shown in the video.

The annotations are made with open vocabulary — that is, they can use any English words as long as they contain only standard (ASCII) characters. They should avoid using any symbols or computer codes. Annotators were further guided not to use proper nouns (*e.g.*, do not state the person name) and information obtained from audio. They were also instructed to select all HLFs appeared in the video.

3. Corpus Analysis

13 annotations were created for 140 videos, resulting in 1820 documents in the corpus. They are referred to as **hand annotations** in the rest of this paper. The total number of words is 30954, hence the average length of one document is 17 words. We counted 1823 unique words and 1643 keywords (nouns and verbs).

Figure 1 shows a video segment for a meeting scene, sampled at 1 fps (frame per second), and three examples for hand annotations. For keywords, a similar set of HLFs were selected in most hand annotations although there was some differences in human's age and emotion information. For titles, some annotators provided a main theme based on semantic interpretation of the video scene (*e.g.*, hand annotations 1 and 2 in Figure 1), while others stated the visual information without much context (*e.g.*, hand annotation 3). Full descriptions typically contained two to five phrases or sentences. Most sentences were short, ranging between two to six words. Descriptions for human, gender, emotion and action were commonly observed. Occasionally minor details for objects and events were also stated. Descriptions for the background were often associated with objects rather than humans. It is interesting to observe the subjectivity

Hand annotation 1
(keywords) male, adult, old, sit, serious, table, chair, indoor, tv presenter, interview, papers, formal clothes
(title) interview in the studio;
(description) three people are sitting on a red table; a tv presenter is interviewing his guests; he is talking to the guests; he is reading from papers in front of him; they are wearing a formal suit;

Hand annotation 2
(keywords) male, old, sit, happy, serious, table, chair, indoor, tv presenter, host, guests
(title) tv presenter and guests;
(description) there are three persons; the one is host; others are guests; they are all men;

Hand annotation 3
(keywords) male, old, adult, sit, serious, table, chair, indoor
(title) three men are talking;
(description) three people are sitting around the table and talking each other;

Figure 1: A montage showing a meeting scene in a news video and three sets of hand annotations. In this video segment, three persons are shown sitting on chairs around a table — extracted from TREC Video '*20041116_150100_CCTV4_DAILY_NEWS_CHN33050028*'.

with the task; the variety of words were selected by individual annotators to express the same video contents. Figure 2 shows another example of a video segment for a human activity and hand annotations[4].

After removing function words, the frequency for each word was counted in hand annotations (full descriptions). Following two classes are manually defined:

1. A class, relating directly to humans, their body structure, identity, action and interaction with other humans;
2. Another class, representing artificial and natural objects and scene settings (*i.e.*, all the words that are not directly related to humans, although they are important for semantic understanding of the visual scene).

Note that some related words (*e.g.*, '*woman*' and '*lady*') were replaced with a single concept ('*female*'); concepts were then built up into a hierarchical structure.

Hand annotation 1
(keywords) male, female, adult, young, sit, stand, sad, serious, chair, outdoor, park
(title) outdoor talking scene of a man and woman;
(description) young woman is sitting on chair in park and talking to man who is standing next to her;

Hand annotation 2
(keywords) male, female, adult, sit, walk, stand, serious, chair, bus, outdoor, formal suit, people, street, taxi
(title) a couple is talking;
(description) two person are talking; a lady is sitting and a man is standing; a man is wearing a black formal suit; a red bus is moving in the street; people are walking in the street; a yellow taxi is moving in the street;

Hand annotation 3
(keywords) male, female, sit, stand, serious, chair, outdoor, dark clothes, talking
(title) talk of two persons;
(description) a man is wearing dark clothes; he is standing there; a woman is sitting in front of him; they are saying to each other;

Figure 2: A montage of video showing a human activity in an outdoor scene and three sets of hand annotations. In this video segment, a man is standing while a woman is sitting in outdoor — from TREC Video '*20041101_160000_CCTV4_DAILY_NEWS_CHN_41504210*'.

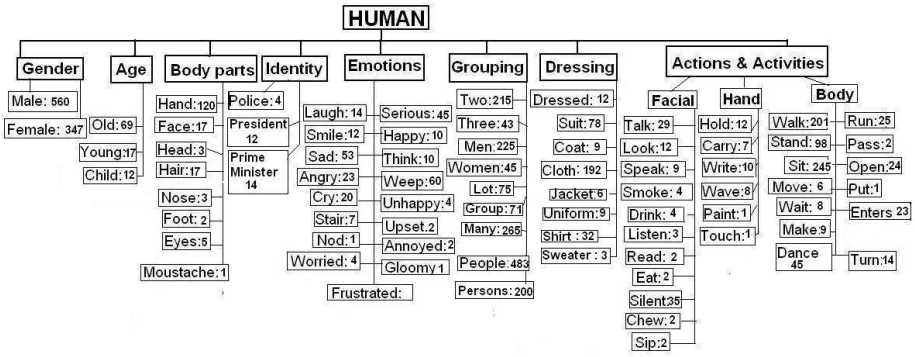


Figure 3: Human related information found in 13 hand annotations. Information is divided into structures (gender, age, identity, emotion, dressing, grouping and body parts) and activities (facial, hand and body). Each box contains a high level concept (*e.g.*, '*woman*' and '*lady*' are both merged into '*female*') and the number of its occurrences.

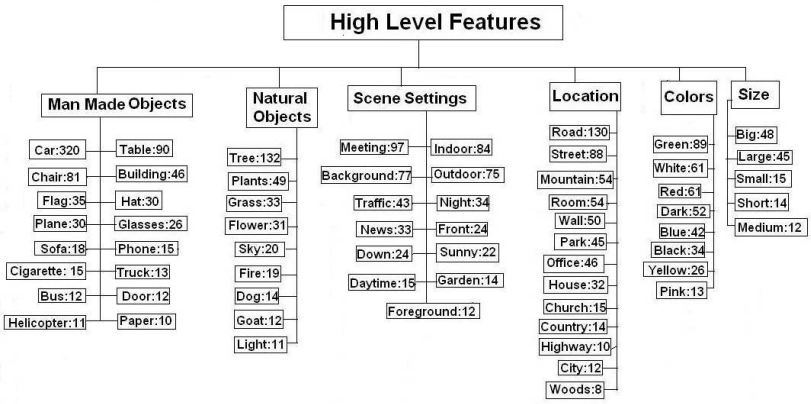


Figure 4: Artificial and natural objects and scene settings were summarised into six groups.

3.1. Human Related Features

Figure 3 presents human related information observed in hand annotations. Annotators paid full attention to human gender information as the number of occurrences for '*female*' and '*male*' is the highest among HLFs. This supported our prediction that most interesting and important HLF was humans when they appeared in a video. On the other hand age information (*e.g.*, '*old*', '*young*', '*child*') was not identified very often. Names for human body parts had mixed occurrences ranging from high ('*hand*') to low ('*moustache*'). Six basic emotions — anger, disgust, fear, happiness, sadness, and surprise as discussed by Paul Ekman[5] — covered most of facial expressions.

Dressing became an interesting feature when a human was in a unique dress such as a formal suit, a coloured jacket, an army or police uniform. Videos with multiple humans were common, and thus human grouping information was frequently recognised. Human body parts were involved in identification of human activities; they included actions such as standing, sitting, walking, moving, holding and carrying. Actions related to human body and posture were frequently identified. It was rare that unique human identities, such as police, president and prime minister, were described. This may indicate that a viewer might want to know a specific type of an object to describe a particular situation instead of generalised concepts.

3.2. Objects and Scene Settings

Figure 4 shows the hierarchy created for HLFs that did not appear in Figure 3. Most of the words were related to artificial objects. Humans interacted with these objects

[4]Although annotations were also provided by TREC Video for these two video segments they were not used for this study. TREC Video annotations differ from our hand annotations to some extent; they are shot based, created for one camera take. Multiple humans performing multiple actions in different backgrounds can be shown in one shot. Descriptions for human, gender and action are observed. Additionally camera motion and angle, ethnicity information and human's dressing are frequently stated, however there are not much details for events or objects.

[5]en.wikipedia.org/wiki/Paul_Ekman

on: 237; in: 121; around: 53; with: 44; near: 43; at: 41; on the left: 35; in front of: 24; together: 24; behind: 22; between: 18; beside: 16; on the right: 16; on the left: 12, in the middle: 10; inside: 7; middle: 7; under: 7

Figure 5: List of frequent spatial relations with their frequency counts.

to complete activities — '*man is sitting on a chair*', '*she is talking on the phone*', '*he is wearing a hat*'. Natural objects were usually in the background, providing the additional context of a visual scene — '*human is standing in the jungle*, '*sky is clear today*'. Place and location information (*e.g.*, room, office, hospital, cafeteria) were important as they showed the position of humans or other objects in the scene — '*there is a car on the road*, '*people are walking in the park*'.

Colour information often played an important part in identifying separate HLFs — *e.g.*, '*a man in black shirt is walking with a woman with green jacket*', '*she is wearing a white uniform*'. The large number of occurrences for colours indicated human's interest in observing not only objects but also their colour scheme in a visual scene. Some hand descriptions reflected annotator's interest in scene settings shown in the foreground or in the background. Indoor/outdoor scene settings were also interested in by some annotators. These observations demonstrated that viewers were interested in high level details of a video and relationships between different prominent objects in a visual scene.

3.3. Spatial Relations

Spatial relation specifies how some object is spatially located in relation to some reference object. A reference object is usually a part of foreground in a video stream. Spatial relations are important when explaining visual scenes. Prepositions (*e.g.*, '*on*', '*at*', '*inside*', '*above*') can present the spatial relations between objects. Their effective use helps in generating smooth and clear descriptions, *e.g.*, '*man is sitting on the chair*' is more descriptive than '*man is sitting*' and '*there is a chair*'. Spatial relations can be categorised into

static: relations between not moving objects;

dynamic: direction and path of moving objects;

inter-static and dynamic: relations between moving and not moving objects.

Static relations can establish the scene settings (*e.g.*, '*chairs around a table*' may imply an indoor scene). Dynamic relations are used for finding activities of moving objects present in the video (*e.g.*, '*a man is running with a dog*'). Inter-static and dynamic relations are a mixture of stationary and non stationary objects; they explain semantics of the complete scene (*e.g.*, '*persons are sitting on the chairs around the table*' indicates a meeting scene). For this study videos containing humans are considered candidates for dynamic and inter-static and dynamic relations. Videos having little motion information are candidates for static relations.

Figure 5 presents a list of most frequent words in the corpus expressing the spatial relations. They were manually counted because not all appearances of these words

Single human: then: 25; end: 24; before: 22; after: 16; next: 12; later on: 12; start: 11; previous: 11; throughout: 10; finish: 8; afterwards: 6; prior to: 4; since: 4;
Multiple humans: meeting:114; while: 37; during: 27; at the same time: 19; overlap: 12; meanwhile: 12; throughout:7; equals: 4,

Figure 6: List of frequent temporal relations with their frequency counts.

indicated spatial relations. For example '*in*' can be used for several different purposes; a sentence, such as '*a man is sitting in the car*', indicates the spatial relation, while '*there is a man in the video*' improves the readability of description, or '*the man in the previous video*' explains a link between various scenes.

3.4. Temporal Relations

Video is a class of time series data formed with highly complex multi dimensional contents, involving not only spatial but also temporal relations. Individual frames are connected together to form a complete and coherent video sequence. To generate a full description of video contents, annotators can use temporal information to join descriptions for sequential frames. The following example uses two temporal relations, *i.e.*, '*after*' and '*later on*', for connecting descriptions of three individual frames:

> *a man is walking;* ***after*** *sometime he enters the room;* ***later on*** *he is sitting on the chair.*

Vallacher and Wegner (43) suggested that it was more common to describe scenarios by time intervals rather than by time points, and listed thirteen relations formulating a temporal logic (*before*, *after*, *meets*, *meet-by*, *overlaps*, *overlapped-by*, *starts*, *started-by*, *finishes*, *finished-by*, *during*, *contains*, *equals*). Temporal relations play a major role in identifying activities in videos. According to Allen's temporal logic (4), most common relations in video sequences are '*before*', '*after*', '*start*' and '*finish*' for single humans. For this corpus, '*overlap*' and '*during*' are also frequently observed.

Based on analysis of the corpus, we describe temporal information in two flavors; (1) temporal information extracted from activities by a single human, and (2) interactions between multiple humans. Figure 6 presents a list of most frequent words in the corpus related to temporal relations. As can be seen, annotators put much focus on keywords related to activities of multiple humans. Keyword '*meeting*' had the highest frequency because annotators usually considered most scenes involving multiple humans as the meeting scene. Keyword '*while*' was typically used for presenting separate activities by multiple humans such as '*a man is walking while a woman is sitting.*'

3.5. Findings from the Corpus Analysis

This section presented analysis of video annotation dataset[6]. The corpus is important for the following reasons:

1. limiting this study to a manageable and defined domain;

[6]We plan to make this dataset public with the following structure, video ID, start time, end time, set of keywords, title, description and annotator ID.

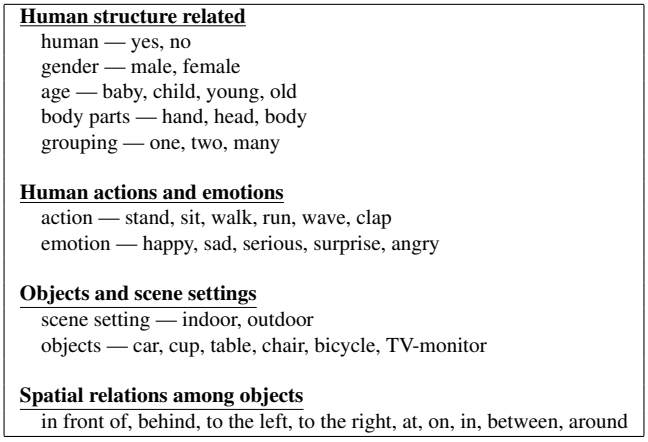


Figure 7: List of HLFs to be extracted by image processing techniques.

2. decision of HLFs that should be extracted by image processing;
3. preparation for development/test data and groundtruths for evaluation.

Concerning 2. above, several conclusions can be drawn based on the analysis of hand annotations. Annotators are most interested in human emotions, actions and their interaction with other humans and objects. Natural objects play a important role for identification of scene settings (*e.g.*, presence of trees indicates '*park scene*', presence of sky generates '*outdoor scene*'). Artificial objects are mostly attached with humans and their activities (*e.g.*, '*man is sitting on the chair*'). Colour information is important to distinguish one object from others. Humans are normally considered a part of foreground while other objects constitute the background. Figure 7 presents a list of HLFs that we are interested in. They roughly cover most important HLFs identified by the analysis. Natural language description of video streams starts with identification of visual features.

4. Describing Individual Video Frames in Natural Language

We explore a bottom up approach to smooth, coherent and well structured descriptions of video sequences. Based on analysis of hand annotations (Section 3) we focus on description humans and their activities in video streams. This section outlines the creation of a compact, sentence length description for each video frame. An approach to generating a full length description for a video stream will be argued in Section 5.

4.1. High Level Features Extraction from Video

Identification of human face or body can prove the presence of human in a video. The method by Hu et al. (16) is adopted for face detection using colour and motion information. The method works against variations in lightning conditions, skin colours, backgrounds, face sizes and orientations. When the background is close to the skin colour, movement across successive frames is tested to confirm the presence of a human

face. Facial features play an important role in identifying age, gender and emotion information (46). Human emotion can be estimated using eyes, lips and their measures (gradient, distance of eyelids or lips). The same set of facial features and measures can be used to identify a human gender[7].

To recognise human actions the approach based on a star skeleton and a hidden Markov model (HMM) is implemented (37). Commonly observed actions, such as '*walking*', '*running*', '*standing*', and '*sitting*', can be identified. Human body is presented in the form of sticks to generate features such as torso, arm length and angle, leg angle and stride (8). Further Haar features are extracted and classifiers are trained to identify non-human objects (45). They include car, bus, motor-bike, bicycle, building, tree, table, chair, cup, bottle and TV-monitor. Scene settings — indoor or outdoor — can be identified based on the edge oriented histogram (EOH) and the colour oriented histogram (COH) (19).

Detailed description about extraction and performance of the HLF extraction task is summarised in Khan and Gotoh (17). Although erroneous at various levels, image processing techniques are able to produce HLFs that can be used as predicates for the natural language generation task. Finally, spatial relations are estimated using positions of humans and other objects (or their bounding boxes, to be more precise). Following relationships can be recognised between two or three objects: '*in front of*', '*behind*', '*to the left*', '*to the right*', '*beside*', '*at*', '*on*', '*in*', and '*between*'. For example Bai et al. (5) illustrated how to calculate the three-place relationship '*between*'.

Predicates. HLFs listed in Figure 7 can be seen as predicates for natural language generation. Some predicates are derived by combining multiple HLFs extracted, *e.g.*, '*boy*' may be inferred when a human is a '*male*' and a '*child*'. Apart from objects, only one value can be selected from candidates at one time, *e.g.*, gender can be male or female, action can be only one of those listed. Note that predicates listed in Figure 7 are for describing single human scenes; combination of these predicates may be used if multiple humans are present.

4.2. Natural Language Generation

HLFs acquired by image processing require abstraction and fine tuning for generating syntactically and semantically sound natural language expressions. Further humans and objects need to be assigned proper semantic roles. In this study, a human is always treated as a subject, performing a certain action. Other HLFs are treated as objects, affected by human's activities. These objects are usually helpful for description of background and scene settings. A template filling approach is applied for sentence generation. A template is a pre-defined structure with slots for user specified parameters. Each template requires three components: lexicons, template rules and grammar. Lexicon is a vocabulary containing HLFs extracted from a video stream (see Figure 8). Grammar assures syntactical correctness of the sentence. Template rules are defined for selection of proper lexicons with a well defined grammar.

Template rules. Template rules are employed for selection of appropriate lexicons. Following are some template rules used in this work.

[7] www.virtualffs.co.uk/In_a_Nutshell.html

Noun	→	man | woman | car | cup | table | chair | cycle | head | hand | body
Verb	→	stand | walk | sit | run | wave
Adjective	→	happy | sad | serious | surprise | angry | one | two | many | young | old
Pronoun	→	me | i | you | it | she | he
Determiner	→	the | a | an | this | these | that
Preposition	→	from | on | to | near | while
Conjunction	→	and | or | but

Figure 8: Lexicons and their part of speech (POS) tags.

If (gender == male) then *man* **else** *woman*
Select 1 (Action == *walk*, *run*, *wave*, *clap*, *sit*, *stand*)
Select n (Object == *car*, *chair*, *table*, *bike*)
Elaboration (**If** '*the car is moving*' and '*person is inside the car*') then '*person is driving the car*'

Figure 9: Template rules applied for creating a sentence '*man is driving the car*'.

Base returns a pre-defined string (*e.g.*, when no HLF is detected);

If is the same as an if-then statement of programming languages, returning a result when the antecedent of the rule is true;

Select 1 is same as a condition statement of programming languages, returning a result when one of antecedent conditions is true;

Select n is used for returning a result while more than one antecedent conditions is true;

Concatenation appends the the result of one template rule with the results of another rule;

Alternative is used for selecting the most specific template when multiple templates are available;

Elaboration evaluates the value of a template slot.

Figure 9 illustrates template rules selection procedure. This example assumes human presence in the video. **If-else** statements are used for fitting proper gender in the template. Human can be performing only one action at a time referred by **Select 1**. There can be multiple objects which are either part of background or interacting with humans. Objects are selected by **Select n** rule. These values can be directly attained from HLFs extraction step. **Elaboration** rule is used for generating new words by joining multiple HLFs. '*Driving*' is achieved by combing '*person is inside car*' and '*car is moving*'.

Grammar. Grammar is the body of rules that describe the structure of expressions in any language. We make use of context free grammar (CFG) for the sentence generation task. CFG based formulation enables us to define a hierarchical presentation for sentence generation; *e.g.*, a description for multiple humans is comprised of single human actions. CFG is formalised by 4-tuple:

$$G = (T, N, S, R)$$

S → NP VP	*man is walking*
S → NP	*man*
NP → Pronoun	*he*
NP → Det Nominal	*a man*
Nominal → Noun	*man*
Nominal → Adjective nominal	*old man*
VP → Verb	*wave*
VP → Verb NP	*wave hand*
VP → Verb PP NP	*sitting on chair*
PP → Preposition NP	*on chair*

Figure 10: Grammar for lexicons shown in Figure 8, with an example phrase for each rule.

where T is set of terminals (lexicon) shown in Figure 8, N is a set of non-terminals (usually POS tags), S is a start symbol (one of non-terminals). Finally R is rules / productions (Figure 10) of the form $X \rightarrow \gamma$, where X is a non-terminal and γ is a sequence of terminals and non-terminals which may be empty.

Procedure for sentence generation. Figure 11 outlines the procedure for generating natural language descriptions for individual frames. First, subject(s) should be identified; there can be one, two or many (more than two) humans present in the frame. Determiners and cardinals (*e.g.*, '*the*', '*an*', '*a*', '*two*', '*many*') are selected based on the number of subjects. Age, gender and emotion are selected as an adjective for each subject. Action and pose (verb) is also identified. In the presence of human(s), non-human objects are considered either as objects operated by a human or as a part of background. The most likely preposition (spatial relations) is calculated and inserted between the subject, verb and objects.

Suppose that human is absent in the video, a non-human object may be used as a subject. If they are moving, verb ('*moving*') will be attached. If one is moving and the other is static, the verb is attached with the moving object and the static one is considered as a part of background. In case no object is identified, we try to find the scene settings (*i.e.*, indoor or outdoor) and express the scene using a fixed template. Finally, if the scene setting is not identified, we try to detect any motion and express the scene using a fixed template.

5. Describing a Video Stream in Natural Language

When describing a video sequence, simply joining frame based descriptions will have several shortcomings. Descriptions of individual frames are crude, repeated and in some cases missing useful information due to sparseness of HLFs that can be produced by current technologies. Image processing errors can be accumulated. Lack of temporal information may cause a further problem. In this section we first introduce a 'unit', aiming to create smooth and coherent descriptions while alleviating the effects of these shortcomings. By structuring a video sequence based on units, we are able to remove redundancy caused by repeated expressions and to accommodate temporal information into a description. We further explore an approach to paraphrasing unit based descriptions aiming at creating compact and coherent natural language. Temporal information is also incorporated during this process.

Input: video stream, E (initially empty sentence)
Output: F (populated final sentence)

(1) Find subject of the sentence:
— if one human is present — add one subject to E
— if two humans are present — add two subjects to E
— if more than two humans — add multiple subjects to E
(1.1) Find age, gender, emotion (adjective) for subject(s):
— if age is identified — add age to the subject in E
— if gender is identified — add gender to the subject in E
— if emotion is identified — add emotion to the subject in E
(1.2) Find actions (verb) for subject(s):
— if action is identified — add action to the subject in E

(2) Find other HLFs (object) in the video sequence:
— add the object to E
— find the spatial relation between human(s) and HLFs and add keywords to E
— **transfer** $E \rightarrow F$ **and clear** E

(3) If no human is identified in the video — find other HLFs and add these HLF(s) as subject(s) to E
— if HLF is moving — attach '*moving*' (verb) in the E
— if one HLF is moving and the other is static — attach '*moving*' with the moving HLF in the E, and static HLF is considered a part of background
— **transfer** $E \rightarrow F$ **and clear** E

(4) If no HLF is identified in the video — find scene settings (indoor, outdoor)
— if scene settings identified — use the fixed template (*e.g.*, '*this is an outdoor scene*')
— if scene settings are not identified — find any motion in the video and use the fixed template (*e.g.*, '*there is movement in the scene*', or '*this is a static scene*')
— **transfer** $E \rightarrow F$ **and clear** E

Figure 11: Procedure for generating natural language descriptions for individual frames.

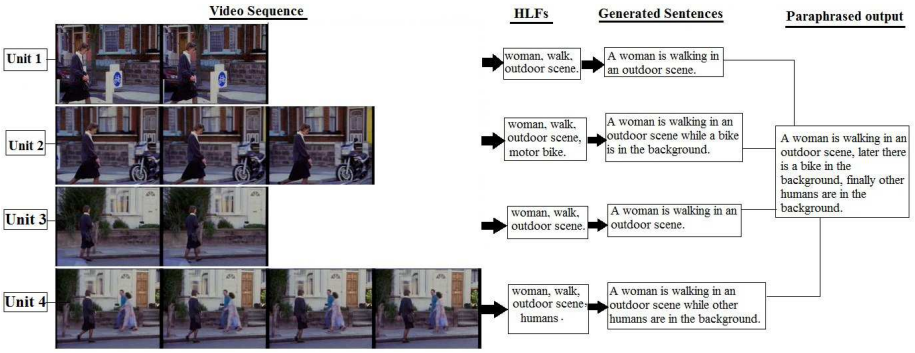


Figure 12: Four frame sequences extracted from a scene, '*a woman walking on a road*' — seen in video '*MS212890*' from the 2007 TREC Video rushes summarisation task. Each row represents a single unit.

5.1. Identifying Units for Description

Definition. We consider a sequence of video frames from which some HLFs (*e.g.*, human, objects or their moves) can be identified. For example in a scene where a man walks out from a room after desk work, the following actions may be identified: '*sitting*' (in front of a desk), '*standing up*' (from a chair), and '*exiting*' (from a room), each of which can span over multiple frames. It may also contain another identifiable features such as facial expressions (*e.g.*, serious in the beginning and smiling later) and some objects in the background. In this work we refer to a sequence of frames with an identical set of visual HLFs as a unit. Using this definition, the length of individual units may be affected by the availability, as well as the quality, of HLF extraction techniques.

Examples. Each row in Figure 12 represents a distinct unit, consisting of a frame sequence of variable length, extracted from a single scene where a human is walking on a road. We assume that image processing techniques resulted in an identical set of visual HLFs for frames in each row. A set of HLFs (for each row of Figure 12) could lead to each line of the following expressions:

> *a woman is walking;*
> *a woman is walking while a bike is in the background;*
> *a woman is walking;*
> *a woman is walking while other humans are in the background.*

A full description of the scene can be derived from the above four lines. However a simple concatenation of the four is crude because the same statement (*i.e.*, '*a woman is walking*') is repeated, hence paraphrasing technique may be explored. Further, consideration of temporal information means that, as soon as a woman and her action is identified, this particular expression ('*a woman is walking*') is no longer required in the rest (until some change happens). The better description of the scene in Figure 12 may be

> *a woman is walking; then a bike is in the background; later other humans are in the background.*

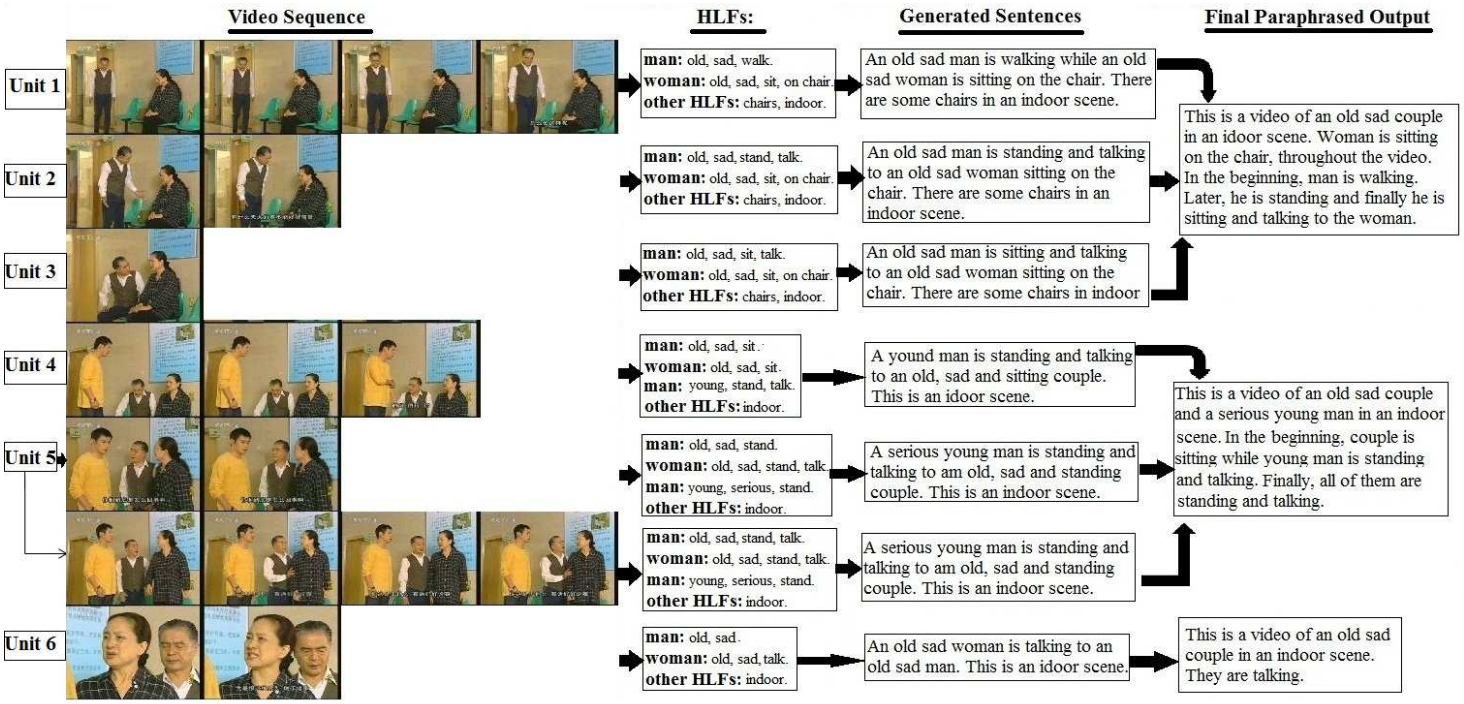


Figure 13: Seven frame sequences extracted from a scene, '*indoor scene with a man and a woman*' — seen in video '*20041101_160000_CCTV4_DAILY_NEWS_CHN*' from the 2005 TREC Video search and retrieval task.

where '*then*' and '*later*' indicate the order of occurrences[8].

Figure 12 further elaborates the concept of unit in videos. Each row presents a single feature unit where HLFs remain constant for a fixed number of frames. Natural language description is created for each unit of frame sequence (*i.e.*, each row in the figure). The next step is to derive a full description of the scene based on individual descriptions. A paraphrasing operation is outlined at this stage.

5.2. Paraphrasing Unit-based Descriptions

By identifying description units we should be able to reduce dramatically, although not fully, redundant and repeated expressions, resulting from simply concatenating frame based descriptions. In order to further improve compactness and coherency, we consider joining multiple unit-based descriptions into a single sentence. We refer to this operation as paraphrasing. The following two problems should be addressed:

- creation of paraphrasing candidates from original descriptions;
- decision of whether to keep the originals, or to replace them with one of paraphrases.

We use a few rules to derive paraphrasing candidates, then calculate statistical language modelling and probabilistic parsing scores to choose the most syntactically appealing expression (13). For example, '*a woman is smiling*' and '*a woman is walking*' can be paraphrased into a sentence '*a woman is smiling and walking*' assuming that the paraphrase has the higher syntactic score.

Creating paraphrasing candidates. Suppose that two original descriptions are given. We aim to create multiple, if possible, paraphrasing candidates by applying a relatively straightforward set of rules:

1. Simply joining both descriptions using conjunction, *i.e.*, '*and*' operator;

   (example) '*a man is walking*' + '*a woman is happy*' ⇒ '*a man is walking and a woman is happy*';

2. Joining originals using most frequent function words (preposition), *e.g.*, '*while*', '*then*', '*after*', '*for*', '*on*', '*with*', '*but*', '*by*', '*because*', '*then*', '*only*', '*between*', '*though*', and '*more*'. The choice is made by calculating the language modelling (LM) score;

   (example) '*a man is walking*' + '*a woman is happy*' ⇒ '*a man is walking while a woman is happy*';

3. Rephrasing sentences:

   (a) Finding the same phrase between both sentence. If it occurs, keep it in the first sentence and discard from the second;

   (example) '*a happy old man is walking*' + '*a happy old man is standing*' ⇒ '*a happy old man is walking and standing*';

[8] One of the hand annotation for this video clip is as follows: '*A woman appears from left. She is walking while a bike in the background. Later she comes across other humans.*'

(b) Dealing with adjectives: if an adjective is found, preference is to place it at the beginning of the sentence;

(example) '*a man is walking*' + '*a man is happy*' ⇒ '*a happy man is walking*';

(c) Dealing with verbs: verbs are joined by the conjunction operator.

(example) '*a man is walking*' + '*a man is talking*' ⇒ '*a man is walking and talking*'.

For 3.(a) above, greedy string tiling (GST) algorithm[9] can be applied to find similar phrases between two sentence (47). Common tiles[10] between both sentences are rephrased together. Rule 3(a) normally will be used in conjunction with other rules, such as 3(b) or 3(c).

Examples. Here are a few examples for creating paraphrasing candidates. In the first example, rule 3(a) and 3(c) result in the same paraphrase:

originals:

a man is walking;
a man is happy;

candidates:

a man is walking and a man is happy; (rule 1)
a man is walking then a man is happy; (rule 2)
a man is walking and happy; (rule 3(a), 3(c))
a happy man is walking; (rule 3(b))

And here is the second example:

originals:

an old man is sitting on the chair;
a man is smiling;

candidates:

an old man is sitting on the chair and a man is smiling; (rule 1)
an old man is sitting on the chair while a man is smiling; (rule 2)
an old man is sitting on the chair and smiling; (rule 3(a), 3(c))
a smiling old man is sitting on the chair; (rule 3(b))

The last example consists of four sentences:

originals:

[9]The advantages of using GST, in comparison to alternative string similarity algorithms such as a longest common subsequence or an edit distance, is its ability to detect block moves: treating the transposition of a substring of contiguous words as a single move instead of considering each word separately.

[10]A tile is a consecutive subsequence of the maximal length that occurs as one-to-one pairing between two input sentences.

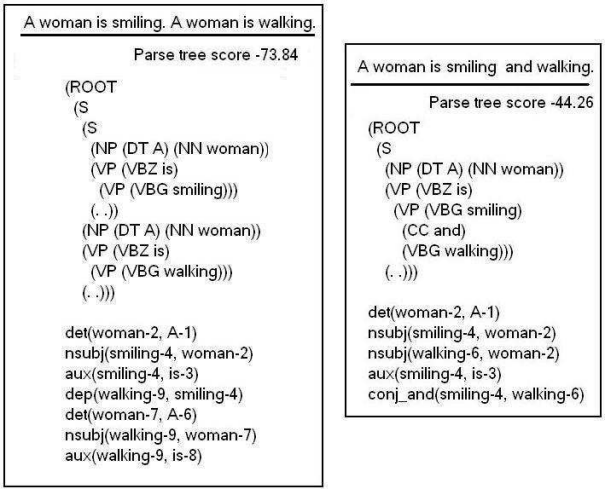


Figure 14: The paraphrase had the higher parsing score than its original.

a woman is walking;
a woman is walking while a bike is in the background;
a woman is walking;
a woman is walking while other humans are in the background;

where a phrase '*a woman is walking*' appears in all sentences, hence it is kept in the first sentence and dropped from the rest. The modified originals and the paraphrasing candidates are the following. Only rules 1 and 2 are able to produce the paraphrase:

modified originals:

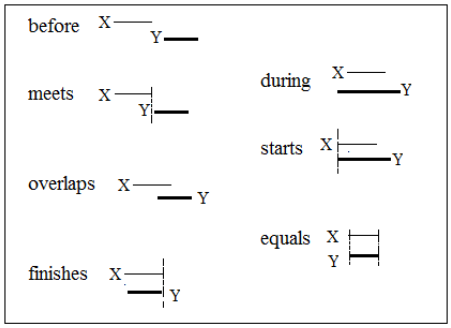
a woman is walking;
a bike is in the background;
other humans are in the background;

candidates:

a woman is walking and a bike is in the background and other humans are in the background; (rule 1)
a woman is walking, then a bike is in the background, later other humans are in the background; (rule 2)

Language modelling. A statistical language model assigns a probability to a sequence of words. A language modelling score can indicate which one is more syntactically likely between the original description and its paraphrases. In the experiments (Section 6) we derived a trigram language model from the Penn Treebank (28), using *SRILM* toolkit (39). As an example, using this language model, the paraphrase:

a woman is smiling and walking.



[t]

Figure 15: Temporal relations between two events X and Y. This figure is from Muller and Reymonet (29).

has the greater log likelihood score than its original in two phrases:

a woman is smiling; a woman is walking.

As a consequence the paraphrase is chosen.

Probabilistic parsing. The syntactic structure of a natural language description can be parsed using a probabilistic parser. Because a parse tree determines the terminal yield it is sufficient to calculate the probability of the tree. A probabilistic parser is able to score sentences before and after the paraphrasing operation. An example in Figure 14 was calculated using a probabilistic parser by Klein and Manning (20).

5.3. Incorporating Temporal Information

TimeML specification language is a standard for presenting temporal information in a natural language text (31). It includes temporal expressions, events, and the relationships they share. The following four tags are mainly used:

- '`TIMEX3`' tag is used for temporal expressions such as date, time, duration and sets. In general they are hard facts presented by specific phrases such as '*12th June*', '*12:10*', '*20 mins*' and '*two days every week*'.
- '`EVENT`' tag usually presents verbs. Nouns, adjectives, and even some prepositions can also be described by this tag. There are seven event categories, namely, '`REPORTING`', '`PERCEPTION`', '`ASPECTUAL`', '`I_ACTION`', '`I_STATE`', '`STATE`' and '`OCCURRENCE`'.
- '`SIGNAL`' tag is used to relate temporal objects to each other by an additional word present, such as '*at*', '*on*', and '*between*', whose function is to specify the nature of that relationship.
- '`LINK`' tag presents relationship between times, events, or between times and events.

Details of *TimeML* and its tags can be found in Pustejovsky et al. (32). Note that '`LINK`' tag in *TimeML* is based on Allen's relation algebra (23). Relations between two time

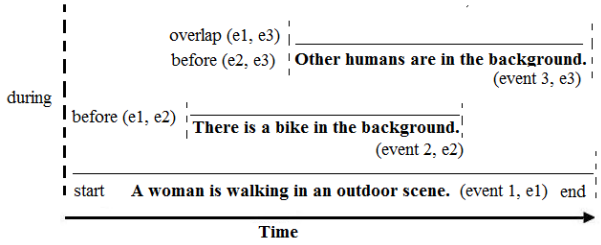


Figure 16: *TimeML* applying temporal relations to machine generated descriptions for Figure 12. Three events are identified and relations between events are shown as '*start*', '*before*', and '*overlap*'.

after	then, after, later, next, thereafter
during	while, at the same time, meanwhile, throughout
before	previous, afterwards, prior to, since
start	initiation, at the beginning, at the start

Figure 17: Keywords for temporal relations between sentences. These keywords are defined for explaining relations based on Allen algebra.

intervals are based on position of the interval endpoints (before, after or simultaneous). Combination of these intervals results in 13 relations; some of which are shown Figure 15.

Nevatia et al. (30) introduced Event Recognition Language based on Allen's algebra for identification of composite events. Rosani et al. (35) presented an approach to automatic composite actions recognition based on context free grammar. Both approaches focused on recognising human activities based on the relations among atomic-level actions. Our work is concerned with the detection of relations between atomic-level actions and events. Use of these relations in our work is twofold; first, we aim to incorporate temporal information into description of activities by a single human, using expressions such as '*before*' and '*after*'. Second, interactions between multiple humans can also be presented, using expressions such as '*meets*', '*overlaps*', '*equals*' and '*during*'.

TARSQI toolkit (44) is used to find these relations between sentences. Figure 16 shows temporal relations for the video sequence in Figure 12. There are three events identified — '*walking*', '*walking while a bike is in the background*' and '*walking while other humans are in the background*'. '*Duration*' relations is common among these three events. First two events have '*before*' and '*after*' relation between them. Last event has two sets of subjects involved; firstly a walking woman and secondly other humans. There is '*overlap*' relation between woman walking and other people walking. Most frequent relations in video sequences are '*before*', '*after*', '*start*' and '*finish*' for single humans, while '*overlap*', '*during*' and '*meeting*' are common for multiple humans. Once these relations are identified between sentences, keywords are manually defined to present them. Figure 17 shows a list of keywords used for *TimeML* to represent relations. Finally they are put into templates to produce temporally coherent description of video sequences.

Hand annotation:
A woman is crying then looks serious. Finally she looks happy and smiling.
Machine generated original:
A woman is crying. A woman is serious. A woman is smiling.
Paraphrased:
A woman is crying, then she is serious. Later on she is smiling.

Figure 18: Three description units in one emotional scene — seen in video '*MRS144765*' from the 2007 TREC Video rushes summarisation task. Shown under the montage are one of hand annotations, machine generated original description and its paraphrase.

6. Experiments

In recent work (18), we presented a framework for investigating quality of descriptions, affected by the number of successfully extracted visual features. We further presented scalability study of our proposed framework and showed that our work can be extended for any type of videos with minor additions in natural language generation templates. Results for evaluation were shown for five more categories, *i.e.*, costume, crowd, sports, violence and animals in addition to seven categories discussed in this paper. That paper focused on the framework for handling potentially missing visual features. Another important issue relating to that work was existence of erroneously identified features (e.g., identification of 'male' instead of 'female').

On the other hand, evaluation in this paper are more focussed towards a framework for creating descriptions based on visual HLFs extracted from a video stream. It was built for a specific genre of videos, incorporating spatial and temporal information in a natural language generation framework. We showed in those experiments that a full description was much more functional than a set of keywords for representing the video content. To this end the task-based evaluation is conducted, and critical observations are made for various categories of videos.

We start this section by presenting a relatively simple scene with three emotion units identified in a single camera shot (Figure 18). Each row represents a unit, indicating changes of the emotional states. It is needless to say that the actual number of

frames in one unit was many more than what is shown. The figure also includes hand annotation, the original machine generated description and its paraphrase More complex example in Figure 19 is a scene with a mixture of human emotions, expressions, background and presence of multiple humans. In the following we aim to evaluate the machine generated descriptions using the hand annotation as a groundtruth.

6.1. Automatic Evaluation using ROUGE

Difficulty in evaluating natural language descriptions stems from the fact that it is not a simple task to define the criteria. We adopted ROUGE, widely used for evaluating automatic summarisation (26), to calculate the overlap between machine generated and hand annotations. Table 1 presents the results where a higher ROUGE score indicates closer match between them. In seven categories it compares (1) simple concatenation of machine generated original descriptions and (2) their paraphrases with temporal information incorporated.

Hand annotations were often subjective, and dependent on one's perception and understanding, that could be affected by educational and professional background, personal interests and experiences. Still there was reasonable similarity between machine generated descriptions and hand annotations as depicted by ROUGE scores. 'Action', 'Close-up' and 'News' videos had higher scores, probably because of the presence of humans with well defined activities and emotions. 'Indoor/Outdoor' videos showed the poorest results, clearly due to the limited capability of image processing techniques. The similarity score was improved in many cases after the paraphrasing operation. In many videos, paraphrases were much shorter than their ground truth hand annotation. Limited number of extracted HLFs had the worst effect on meeting videos, since hand annotations had variety of HLFs attached with this category videos. Paraphrasing further dropped the similarity score since while paraphrasing, certain sentences are merged to generate lesser number of sentences thus losing some HLFs which were part of original annotations. This means that, although handy, ROUGE might not have been the most suitable measure for evaluation. To remedy this, a task based evaluation strategy was explored in the section below.

In order to evaluate usefulness of proposed approach, we used some state of the art datasets. TACoS (34) build the corpus on top of the 'MPII Cooking Composite Activities' video corpus which contains videos of different activities in the cooking domain, *e.g.*, preparing carrots or separating eggs. MPII Composites contains 212 high resolution video recordings of 1-23 minutes length (4.5 minutes on average). 41 basic cooking tasks such as cutting a cucumber were recorded, each between 4 and 8 times. TACoS multi-Level provides multiple sentence descriptions and longer videos; however they are restricted to the cooking scenario.

Figure 20 presents a montage of video taken from TACoS video dataset and generated description using our framework. It can be seen that for each of the three images, a useful and correct description was generated. Although last two images have limited description due to distance of person from the camera.

6.2. Task Based Evaluation

To shed light on the advantage of description for video retrieval application, two sets of task based evaluations were performed. Firstly human subjects were provided

Machine generated original:

A serious man is speaking while there are humans in the background. A sad man is speaking while there are humans in the background. It is an outdoor scene. A sad woman is speaking while there is a man in the background. It is an outdoor scene. There are many humans. There are many men present. One of the men is old and two are young. It is an indoor scene. There are many human. There are two women, one child and other humans. A sad woman is silent and looking into camera. A child is looking into camera. A sad woman is silent and looking into camera. A child is looking into camera. A man is present while there is another human in front of him. A serious woman is speaking.

Paraphrased:

A sad man is speaking with a sad woman while there are humans in the background in an outdoor scene. One old man is present with two young men and some other humans in an indoor scene. A sad woman is silent and looking into camera. A child is looking into camera. A man is present while there is another human in front of him. A serious woman is speaking.

Figure 19: This video montage is seen in video '*MS2063001*' from the 2008 TREC Video rushes summarisation task. Due to space limitation hand annotation is not shown for this example.

		-1	-2	-3	-L	-S	-SU
Action	original	0.4369	0.3087	0.2994	0.4369	0.3563	0.3686
	paraphrase	0.4839	0.3327	0.3191	0.4919	0.4123	0.4486
Close-up	original	0.5385	0.3109	0.2106	0.4110	0.4193	0.4413
	paraphrase	0.5787	0.3202	0.2198	0.4622	0.4587	0.4713
News	original	0.4814	0.3627	0.2712	0.3852	0.3618	0.3712
	paraphrase	0.4839	0.3327	0.3191	0.4919	0.4123	0.4486
Meeting	original	0.3330	0.2462	0.2400	0.3330	0.2648	0.2754
	paraphrase	0.3216	0.2154	0.2096	0.3187	0.2543	0.2544
Grouping	original	0.3067	0.2619	0.1229	0.3067	0.2229	0.3067
	paraphrase	0.3213	0.2703	0.1312	0.3315	0.2492	0.3188
Traffic	original	0.3121	0.1268	0.1250	0.3121	0.3236	0.3407
	paraphrase	0.3121	0.1268	0.1250	0.3121	0.3236	0.3407
Indoor/Outdoor	original	0.2544	0.1877	0.1302	0.2544	0.2302	0.2544
	paraphrase	0.2544	0.1877	0.1302	0.2544	0.2302	0.2544

Table 1: ROUGE scores between machine generated descriptions and 13 hand annotations. ROUGE 1-3 shows *n*-gram overlap similarity between reference and model descriptions. ROUGE-L is based on longest common subsequence. ROUGE-S skips bigram co-occurrence without gap length. ROUGE-SU shows results when skipping bigram co-occurrence with unigrams.

(a:) A person is walking. This is an indoor scene.
(b:) A man is standing. This is an indoor scene.
(c:) A person is present. This is an indoor scene.
(d:) A person is standing. This is an indoor scene.

Figure 20: Comparison against TACoS dataset (34).

with a description and required to find a matching video. Secondly human subjects were provided with a video clip and required to find a matching description.

The evaluation strategy for **Task 1** was designed as follows: human subjects were instructed to find a video that corresponded to a natural language description. Each subject was provided with one textual description and 20 video segments at one time. The same set of 20 video clips were repeatedly used, a half of which were selected from the 'Close-up' category and the rest were from the 'Action' category. This resulted in a pool of candidates, consisting of clearly distinctive videos (between categories) and videos with subtle differences (within a single category). Once a choice was made, each subject was provided with the correct video stream and the following questionnaire:

question 1: how well the video stream were explained, rating from 'explained completely', 'satisfactorily', 'fairly', 'poorly', or 'does not explain';

question 2: fluency, rating from 'very fluent', 'satisfactory', 'fair', 'poor', to 'does not make sense';

question 3: usefulness for including the following visual contents into descriptions, ratings from 'most useful', 'very useful', 'useful', 'slightly useful' to 'not useful':

- scene description in a text form;
- references to humans, age, gender, emotion and expression;
- references to objects and relationship with humans;
- background, colour information or scene settings.

Five human subjects conducted this task searching a corresponding video for each of five descriptions. They did not involve creation of the dataset, hence they saw these videos for the first time. A baseline performance was measured by replacing a description with keywords. Keywords consisted of a complete set of HLFs that were used for deriving the natural language description. For fairness, subjects were provided with keywords and descriptions for different videos. This arrangement was needed because use of the same video for both keywords and descriptions almost always affected the performance when they saw the same video for the second time.

For **Task 2**, subjects were provided with a video segment at each time and instructed to choose the corresponding description out of ten candidates, consisting of five from the 'Action' category and another five from the 'Close-up' category. The remaining procedure was the same as Task 1: five human subjects conducted this task, searching a corresponding description for each of five videos. None of them involved in dataset creation nor Task 1. A baseline performance was measured by replacing a description with keywords.

Figure 21 presents results for correctly identified videos for both evaluation tasks. For both Tasks 1 and 2, description based retrieval performed better than the keyword baseline by a clear margin (roughly 20% absolute or more), indicating that transforming keywords into more verbose descriptions was a valuable exercise. Task 2 resulted in higher performance for both the keyword based baseline (46% correct) and description based evaluation (68%). It was probably because Task 2 was inherently the simpler

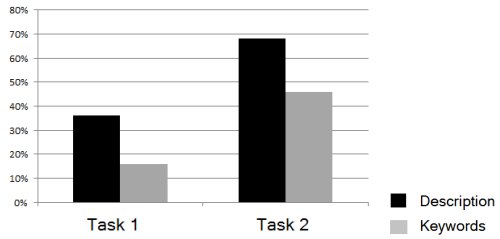


Figure 21: Correctly identified videos based on description and keywords alone.

task, not only because the number of candidates were less (ten as opposed to 20 in Task 1) but also comparison of descriptions was more efficient than comparison between video streams. In the following we summarise the outcomes from the questionnaire collected in Task 1. The same set of questionnaire was set for Task 2 however, the outcomes were similar to those for Task 1, hence we do not present them in this paper.

How well the video stream were explained. This question measures the scale for using natural language to describe videos. Figure 22(a) shows more than 50% of subjects responded that natural language descriptions provided satisfactory explanation (or better) of the video. Only 20% of subjects stated that keywords were satisfactory or better, and more than 40% considered keywords explained the video poorly or worse.

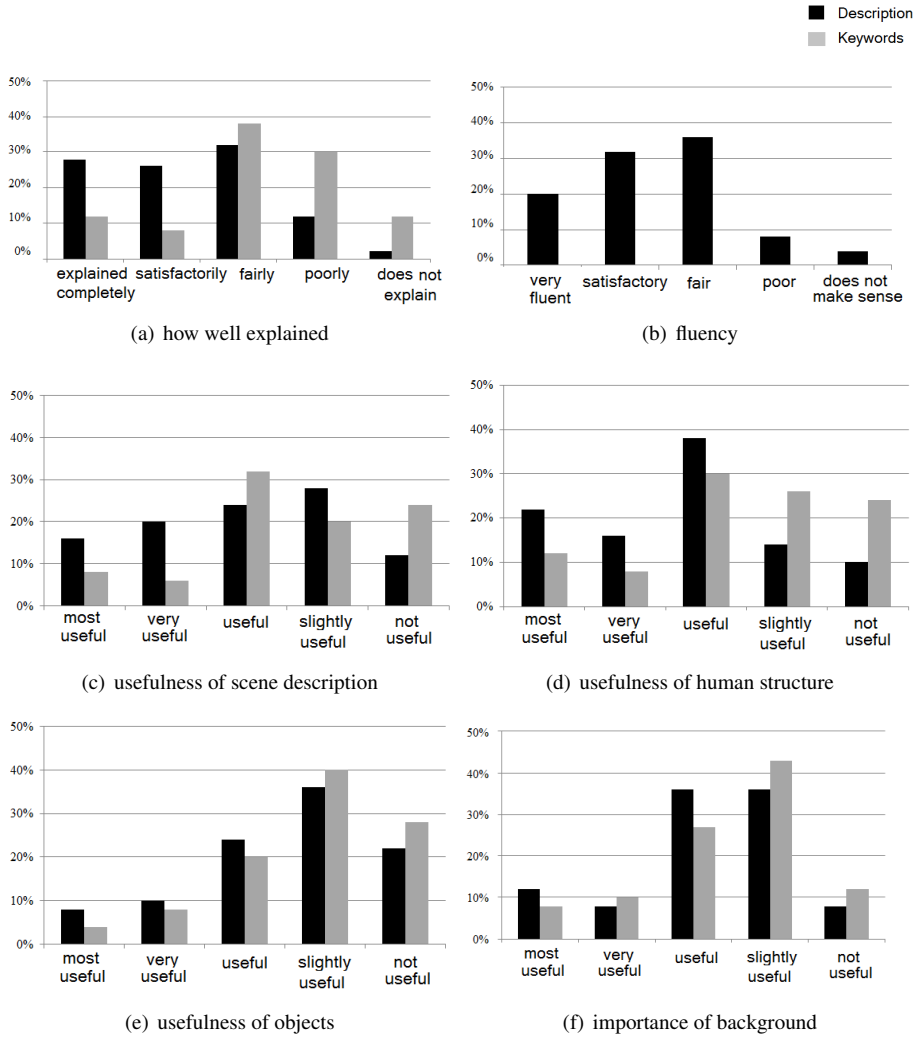
Fluency. This question sheds light on the fluency factor of generated descriptions[11]. Figure 22(b) indicates that about 50% of subjects said that fluency in natural language descriptions was at least satisfactory, while 10% of cases they noticed poorly fluent expressions.

Usefulness of scene description. This questions finds the effect of scene description presented in a text form. The result was encouraging, as shown in Figure 22(c), since roughly 60% of subjects found scene description was useful for a video retrieval task. The number dropped to 40% for keyword based identification.

Usefulness of human structure related information. As most videos in the dataset were related to humans, their emotions and activities, this question was very important. Much emphasis was placed on the effect of human structure related information such as age, gender, emotions (facial expressions) and body gestures. The question aimed to find the effect of human related descriptions for correct identification of videos. Figure 22(d) shows that more than 70% of subjects considered this information was useful, of which nearly 40% said it was very useful. Even for keywords based evaluation, roughly 50% of subjects found this information useful.

Usefulness of objects and their relations with humans. For better understanding of visual scene and its semantics, non-human objects play a very important role. As presented in Figure 22(e), the outcome for this question was not very encouraging; roughly 60% of participants were unable to find well formed explanation of objects

[11]No comparison is made against keywords since measuring fluency with keywords does not make sense.



(a) how well explained

(b) fluency

(c) usefulness of scene description

(d) usefulness of human structure

(e) usefulness of objects

(f) importance of background

Figure 22: Outcomes from the questionnaire collected in Task 1 for the task based evaluation, comparing natural language descriptions and keywords based evaluation.

and their relations with humans in descriptions generated although descriptions were still considered better than keywords.

Importance of background, colour information and scene settings. Often, background and scene settings help in better understanding of the visual scene when humans and other objects are present. However humans and non-human objects are not always present (or failed to be identified all together by the image processing techniques). In such cases, it is particularly useful if we are able to identify visual scene setting and colour information of a video. This question measures the importance of background and scene setting information. Figure 22(f) shows that most subjects thought they were useful (or slightly so) but not too much. Similar results were obtained for keywords based evaluation. This might have been the consequence of the dataset biased towards videos with human centred contents.

7. Conclusion

This paper explored the bottom up approach to describing video contents in natural language. Conversion from quantitative information to qualitative predicates was suitable for conceptual data manipulation and natural language generation. The outcome of the experiments indicates that the natural language formalism makes it possible to generate fluent, rich descriptions, allowing for detailed and refined expressions. Future work may include detection of groups, more complex interactions among humans and other objects and extension of behavioural models.

References

[1] A. Abella, J.R. Kender, and J. Starren. Description generation of abnormal densities found in radiographs. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 542, 1995.

[2] A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1250–1258, 2010.

[3] A. Aker and R. Gaizauskas. Generating descriptive multi-document summaries of geo located entities using entity type models. *Association for Information Science and Technology*, 66(4):721–738, 2015.

[4] J.F. Allen. Towards a general theory of action and time. *Artificial intelligence*, 23 (2):123–154, 1984.

[5] L. Bai, K. Li, J. Pei, and S. Jiang. Main objects interaction activity recognition in real images. *Neural Computing and Applications*, pages 1–14, 2015.

[6] P. Baiget, C. Fernández, X. Roca, and J. Gonzàlez. *Trajectory-based Abnormality Categorization for Learning Route Patterns in Surveillance*. Springer Berlin Heidelberg, 2012.

[7] Claudia Cruz-Perez, Oleg Starostenko, Vicente Alarcon-Aquino, and Jorge Rodriguez-Asomoza. Automatic image annotation for description of urban and outdoor scenes. In *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*, pages 139–147. Springer, 2015.

[8] D. Das. Human gait classification using combined hmm & svm hybrid classifier. In *IEEE International Conference on Electronic Design, Computer Networks & Automated Verification (EDCAV)*, pages 169–174, 2015.

[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[10] G. Erozel, N. K. Cicekli, and I. Cicekli. Natural language querying for video databases. *Information Sciences*, 178(12):2534–2552, 2008.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[12] Y. Feng and M. Lapata. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249, 2010.

[13] S. Filice, G. Da San Martino, and A. Moschitti. Structural representations for learning relations between pairs of texts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China, July. Association for Computational Linguistics*, 2015.

[14] M. Gitte, H. Bawaskar, S. Sethi, and A. Shinde. Content based video retrieval system. *International Journal of Research in Engineering and Technology*, 3(6), 2014.

[15] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Sarad Venugopalan, Randy Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2712–2719. IEEE, 2013.

[16] W. C. Hu, C. Y. Yang, D. Y. Huang, and C. H. Huang. Feature-based face detection against skin-color like backgrounds with varying illumination. *Journal of Information Hiding and Multimedia Signal Processing*, 2(2):123–132, 2011.

[17] M. U. G. Khan and Y. Gotoh. Describing video contents in natural language. In *Proceedings of the EACL workshop*, Avignon, 2012.

[18] M. U. G. Khan, N. Al Harbi, and Y. Gotoh. A framework for creating natural language descriptions of video streams. *Information Sciences*, 303:61–82, 2015.

[19] W. Kim, J. Park, and C. Kim. A novel method for efficient indoor–outdoor image classification. *Journal of Signal Processing Systems*, pages 1–8, 2010.

[20] D. Klein and C.D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, 2003.

[21] A. Kojima, M. Takaya, S. Aoki, T. Miyamoto, and K. Fukunaga. Recognition and textual description of human activities by mobile robot. In *Proceedings of the 3rd International Conference on Innovative Computing Information and Control*, pages 53–53, 2008.

[22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1608, 2011.

[23] H. Lee, V. Morariu, and L. S. Davis. Clauselets: leveraging temporally related actions for video event analysis. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1161–1168, 2015.

[24] M.W. Lee, A. Hakeem, N. Haering, and S.C. Zhu. Save: A framework for semantic annotation of visual events. In *Proceedings of the Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[25] S. Li, G. Kulkarni, T.L. Berg, A.C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the 15th Conference on Computational Natural Language Learning*, pages 220–228, 2011.

[26] C.Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop*, 2004.

[27] Dahua Lin, Chen Kong, Sanja Fidler, and Raquel Urtasun. Generating multi-sentence lingual descriptions of indoor scenes. *arXiv preprint arXiv:1503.00064*, 2015.

[28] M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993.

[29] P. Muller and A. Reymonet. Using inference for evaluating models of temporal discourse. *12th International Symposium on Temporal Representation and Reasoning*, 2005.

[30] R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, volume 4, pages 39–39, 2003.

[31] J. Pustejovsky, J. Castano, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics*, 2003.

[32] J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. The specification language TimeML. *The Language of Time: A Reader. Oxford University Press, Oxford*, 2004.

[33] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. *Pattern Recognition, Lecture Notes in Computer Science, Springer International Publishing*, 8753:184–195, 2014.

[34] Marcus Rohrbach, Wei Qiu, Igor Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 433–440. IEEE, 2013.

[35] A. Rosani, N. Conci, and F. G. B. De Natale. Human behavior understanding for assisted living by means of hierarchical context free grammars. In *IS&T/SPIE Electronic Imaging (pp. 90260E-90260E). International Society for Optics and Photonics*, 2014.

[36] Bharat Singh, Xintong Han, Zhe Wu, Vlad I Morariu, and Larry S Davis. Selecting relevant web trained concepts for automated event retrieval. *arXiv preprint arXiv:1509.07845*, 2015.

[37] D. Singh, A. K. Yadav, and V. Kumar. Human activity tracking using star skeleton and activity recognition using hmms and neural network. *International Journal of Scientific and Research Publications*, 4(5), 2014.

[38] A.F. Smeaton, P. Over, and W. Kraaij. High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. *Multimedia Content Analysis*, pages 1–24, 2009.

[39] A. Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, 2002.

[40] Chun Chet Tan, Yu-Gang Jiang, and Chong-Wah Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 655–658. ACM, 2011.

[41] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014.

[42] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING), August*, 2014.

[43] R. R. Vallacher and D. M. Wegner. *A theory of action identification*. Psychology Press, 2014.

[44] M. Verhagen, I. Mani, R. Sauri, R. Knippen, S.B. Jang, J. Littman, A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating temporal annotation with TARSQI. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 81–84, 2005.

[45] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 2001.

[46] Torsten Wilhelm, Hans-Joachim Böhme, and Horst-Michael Gross. Classification of face images for gender, age, facial expression, and identity. In *Artificial Neural Networks: Biological Inspirations–ICANN 2005*, pages 569–574. Springer, 2005.

[47] M. J. Wise. String similarity via greedy string tiling and running karp-rabin matching. *Online Preprint, Dec*, 1993.

[48] Fei Yan and Krystian Mikolajczyk. Leveraging high level visual information for matching images and captions. In *Computer Vision–ACCV 2014*, pages 613–627. Springer, 2015.

[49] Y. Yang, C.L. Teo, H. Daumé III, C. Fermüller, and Y. Aloimonos. Corpus-guided sentence gereration of natural images. In *Proceedings of the EMNLP*, 2011.

[50] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos. A cognitive system for understanding human manipulation actions. *Advances in Cognitive Systems*, 3:67–86, 2014.

[51] B.Z. Yao, X. Yang, L. Lin, M.W. Lee, and S.C. Zhu. I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.