Article:
Khan, M., AlHarbi, N. and Gotoh, Y. (2015) A framework for creating natural language descriptions of video streams. Information Sciences, 303. 61 - 82. ISSN: 1872-6291

https://doi.org/10.1016/j.ins.2014.12.034

A Framework for Creating Natural Language Descriptions of Video Streams

Muhammad Usman Ghani Khan[1]

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

Nouf Al Harbi[2], Yoshihiko Gotoh[3]

Department of Computer Science, University of Sheffield, United Kingdom

Abstract

This contribution addresses generation of natural language descriptions for important visual content present in video streams. The work starts with implementation of conventional image processing techniques to extract high-level visual features such as humans and their activities. These features are converted into natural language descriptions using a template-based approach built on a context free grammar, incorporating spatial and temporal information. The task is challenging particularly because feature extraction processes are erroneous at various levels. In this paper we explore approaches to accommodating potentially missing information, thus creating a coherent description. Sample automatic annotations are created for video clips presenting humans' close-ups and actions, and qualitative analysis of the approach is made from various aspects. Additionally a task-based scheme is introduced that provides quantitative evaluation for relevance of generated descriptions. Further, to show the framework's potential for extension, a scalability study is conducted using video categories that are not targeted during the development.

Keywords: video retrieval, video annotation, natural language generation

1. Introduction

Humans can describe a video scene in natural language without much effort. However what is simple for a human may not always be easy for a machine. To a certain extent machines are able to identify visual content in videos [1] but only a small number of works exist towards automatic description of visual scenes. Most studies in video retrieval have been based on keywords [2]. Although important concepts in a visual scene can be presented by keywords, they lack context information which is needed for detailed explanation of the video sequences. An interesting extension to the keyword based scheme is natural language textual description of video streams. They are human friendly and are able to clarify context between keywords by capturing their relations. Descriptions can guide generation of video summaries by converting a video to natural language and provide a basis for creating a multimedia repository for video analysis, retrieval and summarisation tasks.

This work. This paper presents a bottom-up approach to describing video contents in natural language, with a particular focus on humans, their activities and interaction with other objects. Conventional image processing techniques are applied to extract high-level features (HLFs) from individual video frames. Natural language generation is performed using extracted visual features as predicates that are fed to the templates based on a context free grammar (CFG).

In particular this paper focuses on one important issue that has not been addressed in recent work; we aim to establish a framework for accommodating processing errors, specifically those from the image processing stage.

[1]Corresponding author. email: usmanghanikhan@gmail.com (M.U.G. Khan). The work was conducted while the first author was in the University of Sheffield.

[2]email: nmalharbi1@sheffield.ac.uk (N. Al Harbi).

[3]email: y.gotoh@dcs.shef.ac.uk (Y. Gotoh).

Progress made in image processing technologies in recent years has been substantial, nevertheless we are able to extract a limited number of visual features, most of which are below humans' ability. This manuscript addresses the effect of missing or erroneously identified features, then presents a framework whereby a number of sentence templates are prepared, each of which incorporates a different combination of visual features. Given this framework the approach selects the most suitable template that accommodates visual features that are successfully extracted.

Using a dataset, consisting of natural language descriptions of video segments crafted from a small subset of TREC Video[4] data [3], we first study the image processing errors (Section 2). We then develop the framework for natural language generation that is robust to a number of image processing errors (Sections 3 and 4). The experiments consist of an automatic scheme and a task-based evaluation by human subjects, showing that the framework is robust against missing visual features (Section 5). A scalability study is also conducted, illustrating that the framework does not fail with a broader range of video contents for which only a small number of visual features are identified (Section 6). The outcome indicates that, although the amount of image processing errors can vary, the framework is able to produce syntactically correct expressions. The additional benefit is that the scheme can handle a video stream in a different genre from those considered for development of the framework.

Related work. There have been an increasing number of efforts made in recent years towards description of videos. Baiget *et al.* manually performed human identification and scene modelling, focusing on human behaviour description of crosswalk scenes [4]. Lee *et al.* introduced a framework for semantic annotation of visual events in three steps; image parsing, event inference and language generation [5]. Instead of humans and their activities, they focused on detection of objects, their inter-relations and events in videos. Yao *et al.* presented their work on video-to-text description [6]; this work was dependent on a significant amount of annotated data, a requirement that is avoided in this paper. Yang *et al.* developed a framework for static image to textual descriptions where they dealt with images with up to two objects [7]. Krishnamoorthy *et al.* presented triplet (subject, verb and object) based sentence generation where image processing techniques were applied for extraction of subjects and their activities [8]. For presenting context information web-scale corpora were used. However their work did not handle complex textual properties such as adjectives, adverbs, multiple objects and multi-sentence descriptions of long videos where various activities were observed. Their approach was further extended by Guadarrama *et al.* who employed a rich set of content words (218 verbs and 241 different objects) [9]. Direct manipulation of visual contents was not considered, but they made use of textual corpora when generating descriptions.

More recently Metze *et al.* presented a topic oriented multimedia summarisation (TOMS) system which was able to generate a paragraph description of multimedia events using important information in a video belonging to a certain topic [10]. Their feature sets included objects, actions, environmental sounds and speech recognition transcripts. Rather than generating descriptions of videos using natural language, their major focus was on the event detection and retrieval of specific events based on user queries. Yu *et al.* generated sentences for video sequences which were comprised of nouns, verbs, prepositions, adjectives and adverbs [11]. Their test set was limited in the sense that the focus was on humans performing some action in outdoor environments. They further generated sentences given a scenario in which two humans were participating in some combined actions, though a scenario with more than two humans was missing from their investigation. Section 5 accommodates additional introduction of related work, including those by Das *et al.* [12] and by Barbu *et al.* [13], where we plan to make some comparison with the framework presented in this paper.

2. Visual Feature Extraction

A dataset was manually created for a small subset prepared form the rushes video summarisation task and the high-level features (HLF) extraction task for the 2007 and 2008 TREC Video evaluations [3]. It consisted of 140 segments of videos; each segment contained a single camera shot, spanning between 10 and 30 seconds in length. There were 20 video segments for each of the following seven categories:

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

[4] trecvid.nist.gov

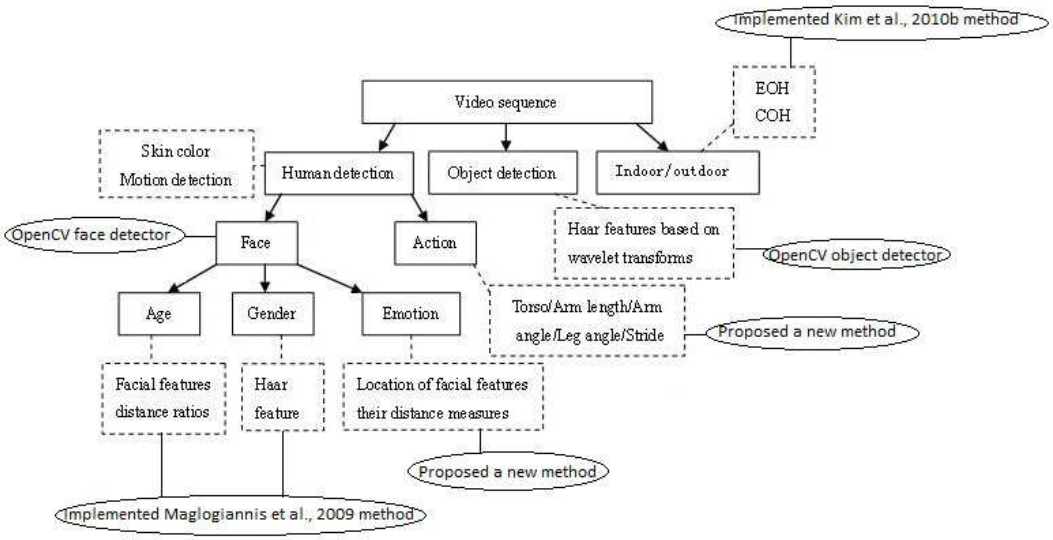


Figure 1: Visual features are extracted from video streams using conventional image processing techniques. Closed rectangles present the HLFs, while dashed rectangles represent lower level features used for identification of the HLFs.

Close-up: A human face is visible in the large part of the screen. Facial expressions sometimes define their emotional status (*e.g.*, happy, sad).

News: An anchor person or a reporter is present. News videos are often characterised by scene settings such as a weather board at the background.

Meeting: Multiple humans are sitting and communicating. Objects such as a table and chairs are present.

Grouping: Multiple humans' interaction is seen but they do not belong to a meeting scenario. A table or chairs may not be present.

Traffic: Vehicles (*e.g.*, cars, buses and trucks) are visible. Traffic signals can be seen.

Indoor/Outdoor: Scene settings are more obvious than human activities. Examples may include a park scene (trees, pond) and an office scene (computer display, desk).

13 human annotators individually created descriptions for the video data. They are referred to as manual annotations in the rest of this paper. Further detail of the dataset and its analysis are presented in [14].

Note that the above seven categories partly overlap. Although an overlap can occur with any categories, it is particularly so between the 'News', the 'Meeting' and the 'Grouping' categories. For example, there can be a meeting scene in a TV programme which is typically classified into the 'Meeting' category rather than to the 'News'. There can be a group of people walking on the street beside the vehicle traffic. We may classify the scene into either the 'Grouping' or the 'Traffic' videos, depending on the core theme of that video clip. The impact of the partial overlap on the theme was studied in the experiment section of [14].

2.1. High-Level Features (HLFs)

Figure 1 illustrates a list of high-level visual features together with the lower level features used for their identification. Detection of a human face or a body can prove the presence of a human in a video. The method by Kuchi *et al.* [15] is adopted for face detection using the colour and motion information. The method works against variations in the lighting conditions, the skin colours, the backgrounds, the face sizes and the orientations. When the background colour 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 the age, gender and the emotion information [16]. Human

	(groundtruth)	
	exist	not exist
exist	1795	29
not exist	95	601

(a) human detection

	(groundtruth)	
	male	female
male	911	216
female	226	537

(b) gender identification

Table 1: Confusion tables for (a) human detection and (b) gender identification. Columns show the groundtruth, and rows indicate the automatic recognition results. The human detection task is biased towards existence of human, while in the gender identification males and females are more balanced.

	(groundtruth)					
	stand	sit	walk	run	wave	clap
stand	98	12	19	3	0	0
sit	0	68	0	0	0	0
walk	22	9	105	8	0	0
run	4	0	18	27	0	0
wave	2	5	0	0	19	2
clap	0	0	0	0	4	9

Table 2: Confusion table for human action recognition. Columns show the groundtruth, and rows indicate the automatic recognition results. Some actions (*e.g.*, '*standing*') were more commonly seen than others (*e.g.*, '*waving*').

emotion can be estimated using eyes, lips and their measures (the gradient, the distance for eyelids or lips). The same set of facial features and measures can be used to identify a human gender[5].

To recognise human actions an approach based on the star skeleton and a hidden Markov model (HMM) is implemented [17]. Commonly observed actions, such as '*walking*', '*running*', '*standing*', and '*sitting*', can be identified. A human body is presented in the form of sticks to generate features such as the torso, the arm length and angle, the leg angle and the stride [18]. Further Haar features are extracted and classifiers are trained to identify non-human objects [19]. 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) [20]. In the following, we review the implemented approaches and the outcomes for extracting various visual features.

2.2. Extracting HLFs in Video

Conventional image processing techniques were able to identify HLFs only to a certain extent, depending on a nature of visual HLF and the image quality, hence resulting in erroneous or potentially missing information from videos. In all the experiments, video frames were extracted using *ffmpeg*[6], sampled at 1 fps (frame per second), resulting in 2520 frames in total. Most of HLFs required one frame to evaluate. Human activities were shown in 45 videos and they were sampled at 4 fps, yielding 3600 frames. Following several trials we decided to use eight frames (roughly two seconds) for human action recognition. Consequently tags were assigned for each set of eight frames, totalling 450 sets of actions.

Table 1(a) presents a confusion table for human detection. It was a heavily biased dataset where human(s) were present in 1890 out of 2520 frames. Of these 1890, misclassification occurred on 95 occasions. On the other hand gender identification is not always an easy task even for humans. Table 1(b) shows a confusion table for gender identification. Out of 1890 frames in which human(s) were present, frontal faces were shown in 1349 images. The total of 3555 humans were present in 1890 frames (1168 frames contained multiple humans), however the table shows the results when at least one gender is correctly identified. Female identification was often more difficult due to make-up, a variety of hair styles and wearing hats, veils and scarves.

Table 2 shows the human action recognition performance tested with a set of 450 actions. It was difficult to recognise '*sitting*' actions, probably because HMMs were trained on postures of a complete human body, while a

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

[6] www.ffmpeg.org — *ffmpeg* is a command line tool composed of a collection of free software and open source libraries. It can record, convert and stream digital audio and video in the numerous formats. The default conversion rate is 25 fps.

	(groundtruth)				
	angry	serious	happy	sad	surprised
angry	59	0	0	15	16
serious	0	661	0	164	40
happy	0	35	427	27	8
sad	61	13	0	281	2
surprised	9	19	0	0	53

Table 3: Confusion table for human emotion recognition. Columns show the groundtruth, and rows indicate the automatic recognition results.

complete posture was often not available when a person was sitting. '*Hand waving*' and '*clapping*' were related to movements in the upper body parts, and detection of '*walking*' and '*running*' relied on the lower body movements. In particular '*waving*' appeared an easy action to identify because of significant movement of the upper body parts. Table 3 shows the confusion for human emotion recognition. '*Serious*', '*happy*' and '*sad*' were the most common emotions in this dataset; among which '*happy*' emotion was the most correctly identified.

Non-human objects. Haar features [19] are extracted in order to identify non-human objects. First, a cascade of boosted classifiers, working with Haar-like features, are trained using a few hundred sample views of a particular object. Positive examples are scaled to roughly the same size (say, 20×20 pixels) and negative examples are arbitrary images of the same size. The trained classifier is applied to regions of interest, of the same size used during the training, in the input image. The output is '1' if the region is likely to show the object (*e.g.*, car, bike, tree), and '0' otherwise. One is required to move the search window across the image and to apply the classifier at every location. The classifier is able to 're-size' itself in order to find objects of interest at different sizes, which is more efficient than resizing the image. As a consequence, an object of the unknown size can be found by repeating the scan procedure for several times at different scales.

It is named as the 'cascade' classifier because the resultant classifier consists of several simpler classifiers that are applied subsequently to regions of interest until at some stage the candidate is found, or no candidate is found at all stages. The word 'boosted' means that the classifiers at every stage of the cascade are complex themselves and they are built out of basic classifiers using various boosting techniques (weighted voting). We implemented the Haar classifier available from *OPENCV*[7] for rapid object detection [21]. The classifier was able to successfully identify non-human objects such as a car, a bus, a motor bike, a bicycle, a building, a tree, a table, a chair, a cup, a bottle and a TV-monitor. Their average precision[8] scores ranged between 44.8 (table) and 77.8 (car).

Indoor/outdoor. Objects of interest are typically shown in the central part of the image, however they hardly play an important role for indoor/outdoor classification. Instead we look at areas close to edges to get clues for this purpose. The ECOH descriptor combines EOH and COH, effectively classifying indoor and outdoor images because it is robust to the effect of sky and grass colours in both classes. EOH helps to differentiate objects based on their edge shapes, *i.e.*, boundaries of objects, while COH captures colour information and explains objects with respect to their colour information [20].

20 video clips in the 'Indoor/Outdoor' category consisted of 12 outdoor and 8 indoor clips. All 12 outdoor clips were stationary scenes with little movement, showing trees, greenery, or buildings, while 3 out of 8 indoor scenes were stationary, presenting a still shot for chairs, tables and cups. The ECOH descriptor was able to correctly classify all 12 outdoor scenes and 6 out of 8 indoor scenes. Finally 140 videos in the full dataset consisted of 84 outdoor and 56 indoor clips, of which only 60% were correctly classified. Presence of multiple objects seems to have caused negative impact on EOH and COH features, hence resulting in many erroneous classifications.

2.3. Formalising Spatial Relations

Identification of spatial relations between multiple humans and objects is important when describing a visual scene. It specifies how a certain object is spatially located in relation to a reference object. The latter is usually a part of the foreground in a video stream. Prepositions (*e.g.*, '*on*', '*at*', '*inside*', '*above*') can present the spatial relations

[7] opencv.willowgarage.com/wiki/

[8] The average precision was defined by [22].

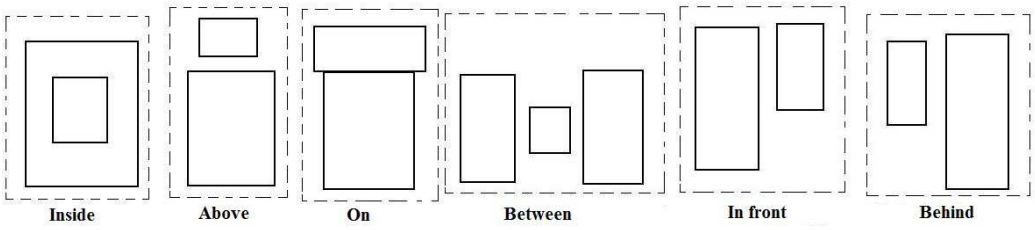


Figure 2: Some spatial relations commonly observed between multiple humans and objects.

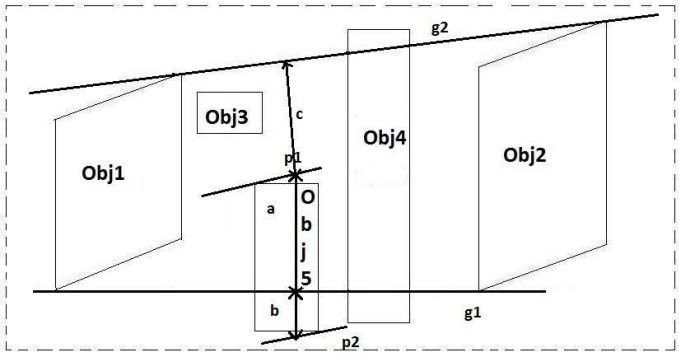


Figure 3: Procedure for calculating the '*between*' relation. Obj 1 and 2 are the two reference objects, while Obj 3, 4 and 5 are the target objects. Step 1: calculate the two tangents g_1 and g_2 between the reference objects using their closed-rectangle representation; Step 2: if both tangents cross the target or its rectangle representation (see Obj 4 in the figure), or the target is totally enclosed by the tangents and the references (Obj 3), the relationship '*between*' is true; Step 3; if only one tangent intersects the target (Obj 5), the applicability depends on its penetration depth in the area between the tangents, thus calculate $\max\{a/(a+b), a/(a+c)\}$; Step 4: otherwise '*between*' relation does not hold.

between objects, and their effective use helps to generate smooth and clear descriptions. For example, '*A man is sitting on the chair*' is more descriptive than '*A man is sitting*' and '*There is a chair*'. Indeed it is good to know whether '*A person is riding on a bike*' or '*A person is carrying a bike on his shoulder*'.

Spatial relations can be categorised into static (relations between unmoving objects), dynamic (direction and path of moving objects), and inter-static and dynamic (relations between moving and unmoving 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 video segments containing humans are considered candidates for dynamic and inter-static and dynamic relations. Videos having little motion information are candidates for static relations. Figure 2 shows some of the commonly observed relations. Spatial relations can be estimated using positions of humans and objects (or their bounding boxes, to be more precise). Figure 3 illustrates steps for calculating the three-place relationship '*between*' [23].

2.4. Impact of HLF Extraction on Natural Language Description

Although the outcomes for the HLF extraction tasks outlined above were roughly comparable to the recent computer vision technologies, at the very best we can only state that it is possible to identify a selected list of visual HLFs with varying precision. The framework for generating natural language description of video streams can be affected by shortcomings of image processing techniques in many aspects. They warrant careful consideration on a wide range of problems, which are summarised in the following three areas: (1) a limited number of HLFs can be processed; (2) some HLFs may fail to be identified; (3) some HLFs may be incorrectly identified. We now look at the individual problems below.

Firstly at this early stage of development, it is a practical decision that we focus on human related visual information observed in a video stream, primarily aiming at restricting the number of visual features to be processed. Apart

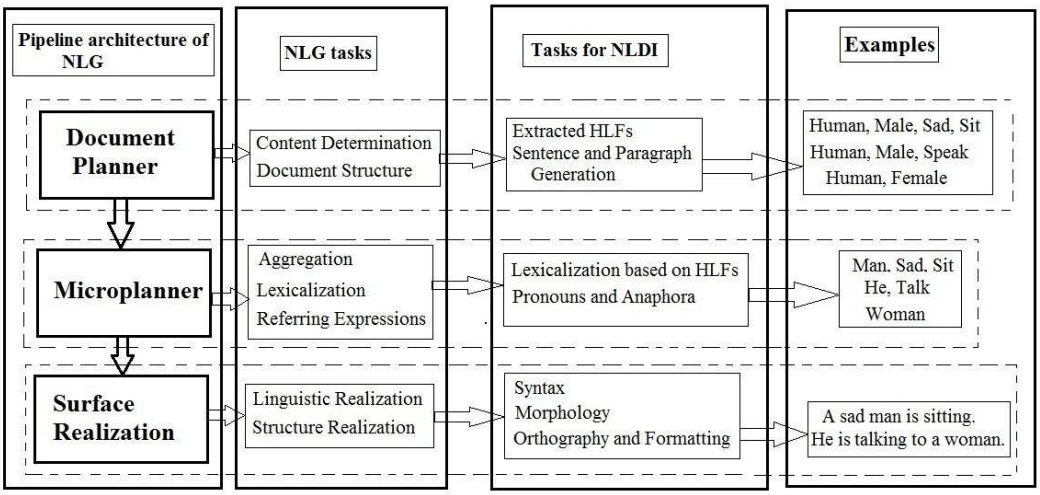


Figure 4: This figure illustrates the pipeline NLG architecture by Reiter and Dale [24] and the tasks for natural language description of a video stream. First and second columns present three modules and their subtasks. Third and fourth columns present corresponding tasks for description of visual information and the example scenario.

from human related features, only a limited number of feature extraction algorithms can be executed. This obviously limits the scope of visual information that can be described in language. For example, a scene setting is categorised into two broad categories — indoor and outdoor — in the current development, whereas it could have been more interesting to identify a specific scene setting such as a room, a park, a hotel, or a street.

In the experiments, extraction of a human face was considered as an indicator for the presence of a human. This assumption worked well where the majority of a frontal face was clearly shown in video. In situations such as occlusion, the side or the rear views of a human face, it was very difficult to identify. This resulted in the second case above (*i.e.*, some HLFs may be failed to be identified). It could also be possible to rely on detection of a human body, instead of a face, however it had its own limitation caused by various reasons such as poses and clothings. A body structure was important when finding the action, hence it was unfeasible to find the correct action without identification of a human body. Similarly, we were able to extract HLFs such as age, gender and emotion only when a human face was successfully identified. They were also difficult to calculate due to broad variation in appearances and structures of human faces (*e.g.*, a woman with a make-up, a man with a beard and a person wearing a mask), hence leading to the third case of problems (*i.e.*, some HLFs may be erroneously identified).

3. Bottom-Up Approach to Natural Language Generation

This work is concerned with an interpretation of video streams, producing a compact presentation of human actions, behavior and their interaction with their surroundings. Visual features, extracted from a video stream, are passed to a natural language generation (NLG) module which generates a human-centred description of the video. This section outlines the bottom up approach to creating language description. Figure 4 presents the pipeline NLG architecture proposed by [24]. It consists of three modules — a document planner, a microplanner, and a surface realiser:

document planner: A document planner determines the contents and the structure of a document to be created. Extraction of visual information, outlined in Section 2, may be considered as the first component of a document planner. It also decides the sentence and the paragraph structure, providing coherency. Visual features are incorporated into predefined sentence structures. These sentences are then put together to generate a full length paragraph with coherent description of the video.

microplanner: A microplanner is responsible for performing three subtasks. Firstly words and syntactic constructs are selected and annotations are marked up (lexicalisation). At this step HLFs are mapped onto their proper

human structure related	human — yes, no
	gender — male, female
	age — baby, child, young, old
	body parts — hand, head, body
human actions and emotions	action — stand, sit, walk, run, wave, clap
	emotion — happy, sad, serious, surprise, angry
	speak — yes, no
	hand gesture — yes, no
	head nodding —- yes, no
objects and scene settings	scene setting — indoor, outdoor
	objects — car, cup, table, chair, bicycle, TV-monitor
spatial relations among objects	in front of, behind, to the left, to the right, at, on, in, between, around

Table 4: Predicates for a single human scene. One visual HLF corresponds to one predicate, *e.g.*, presence of human (yes, no), scene setting (indoor, outdoor) *etc.* 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.

semantic tags such as humans, objects and events. Second it is decided how much information is communicated by each sentence (aggregation). Finally referring expressions — *i.e.*, what phrases should be used to identify entities — are determined.

surface realisation: Surface realisation is a purely linguistic level, which takes choices of words and syntactic structures made during sentence planning and constructs a sentence using them. Section 3.1 presents the approach to surface realisation which combines context free grammar (CFG) with templates for syntactically and semantically correct text generation.

First column in Figure 4 presents three main tasks of the NLG pipeline, while second column shows subtasks against each main task. Third column presents corresponding tasks for natural language description of visual images. Column four illustrates one plausible scenario: '*A man with a sad appearance and a woman talking to each other*'. Initially a document planner stores high-level visual features such as *human*, *male*, *sad*, *sit*, *speak*, *human* and *female* (with potentially duplicated information). It also decides the structure of sentences which may further build up to a paragraph. A microplanner selects proper lexicons for the extracted HLFs — for example, '*human + male*' is replaced with '*man*', '*speak*' with '*talk*', *etc.* Finally a surface realisation module generates syntactically and semantically correct sentences. For example for lexicons '*man*', '*sad*', '*sit*', '*talk*', '*woman*', created sentences are '*A sad man is sitting; he is talking to a woman*'.

3.1. Natural Language Generation

Extraction of visual features results in a list of predicates for sentence generation. Table 4 shows predicates for describing a scene with a single human; their combination may be used if multiple humans are present. HLFs acquired by image processing require abstraction and fine tuning for generating syntactically and semantically sound natural language expressions. Some predicates are derived by combining multiple HLFs, *e.g.*, '*boy*' may be inferred when a '*human*' is a '*male*' and a '*child*'. A part of speech (POS) tag is assigned to each HLF using the NLTK[9] (*Natural Language Toolkit*) POS tagger. 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 the human's activities. These objects are usually helpful for description of the background and the 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 a grammar. The lexicon is a vocabulary containing HLFs extracted from a video stream (Table 5). The grammar assures syntactical correctness of the sentence. Template rules are defined for selection of proper lexicons with a well defined grammar. Given a video frame, a sentence is generated for each of most important entities. A simple template can be

subject (S) performs action (A) on object (O) (*e.g.*) '*He* (S) *kicked* (A) *the ball* (O)'

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

Table 5: Lexicons and their part of speech (POS) tags.

subject + verb:	*A man is walking;*
	A woman is standing;
subject + verb + object:	*A person is smoking a cigarette;*
	A man is drinking tea;
subject + verb + complement:	*He looks tired;*
	A man is old;
subject + verb + object + complement:	*He left the door open;*
	A man is kicking the ball with his right leg;
present continuous tense:	*They are jogging;*
	A man is walking;

Table 6: A partial list of templates and their examples for sentence generation. To fill in the template, a POS tagger assigns labels for all HLFs, such as subject, verb, complement, object — direct and indirect object.

Table 6 presents a partial list of templates used for this work.

Template rules. Template rules are employed for selection of the appropriate lexicons. The following are some template rules used in this work:

Base returns a pre-defined string (*e.g.*, when no visual feature 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 the 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 condition 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 5 illustrates the template rules selection procedure. This example assumes human presence in the video. The **if**-**else** statement is used for fitting a proper gender in the template. The human can perform only one action at a time referred by **Select 1**. There can be multiple objects which are either part of the background or interacting with humans. Objects are selected by the **Select n** rule. These values can be directly attained from the HLFs extraction step. The **elaboration** rule is used for generating new words by joining multiple HLFs. For example, '*driving*' may be inferred by combining '*A person is inside the car*' and '*The car is moving*'.

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

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

[9] www.nltk.org

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** '*A person is inside the car*' and '*The car is moving*') then '*A person is driving the car*'

Figure 5: Template rules applied for creating a sentence '*A man is driving the car*'.

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*

Table 7: Grammar for lexicons shown in Table 5, with an example phrase for each rule.

where T is a set of terminals (lexicon) shown in Table 5, N is a set of non-terminals (usually POS tags), S is a start symbol (non-terminal). Finally R is rules / productions (Table 7) 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.

Implementation. For implementing the templates, *simpleNLG* is used [25]. It also performs some extra processing automatically: (1) the first letter of each sentence is capitalised, (2) '*-ing*' is added to the end of a verb as the progressive aspect of the verb is desired, (3) all words are put together in a grammatical form, (4) appropriate white spaces are inserted between words, and (5) a full stop is placed at the end of the sentence.

3.2. Creating Candidate Sentences

Given a list of lexicons, grammar and template rules, the sentence generation algorithm aims to produce a natural language expression without losing the original contents. It chooses a subject, a verb, objects, determiners, cardinality and adjectives using POS tags for extracted visual HLFs. The following three core structures are available:

subject;
subject + verb;
subject + verb + object.

Additionally a starting phrase is selected from the following three options:

determiner + ;
cardinal + ;
adjective + .

Suppose that visual features relating to a verb and an object are not found by the image processing, we may choose '*subject*' as a core structure. Suppose further that only a single person is identified, templates from '*determiner + subject*' are selected. Alternative templates can be chosen from '*cardinal + subject*' or '*adjective + subject*' in the case where there are two persons or a happy person. Multiple templates are prepared for each combination of a core structure and a starting phrase. For this particular example (*i.e.*, '*determiner + subject*'), there are two templates available: (i) '*a person is present*' and (ii) '*there is a person*'. Finally the language modelling score for each template is compared in order to decide the best candidate.

Another example. Suppose that the following visual features are identified: *human*, *male*, *happy*, *sit* and *chair*. With these HLFs, the core structure '*subject + verb + object*' is selected. For a starting phrase, a choice needs to be made between '*determiner +*' and '*adjective +*'. The former choice generates a sentence '*A man is happy and sitting on the chair*'. If the latter is selected for an adjective start, an expression '*A happy man is sitting on the chair*' is created. Their language modelling scores decide which one is the more likely candidate.

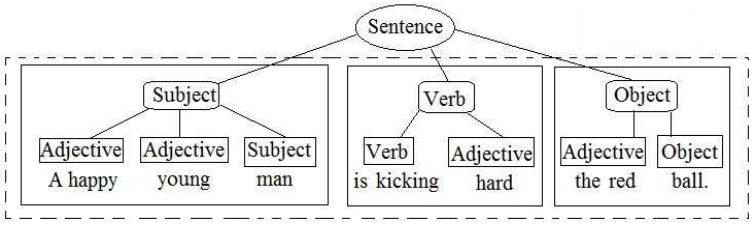
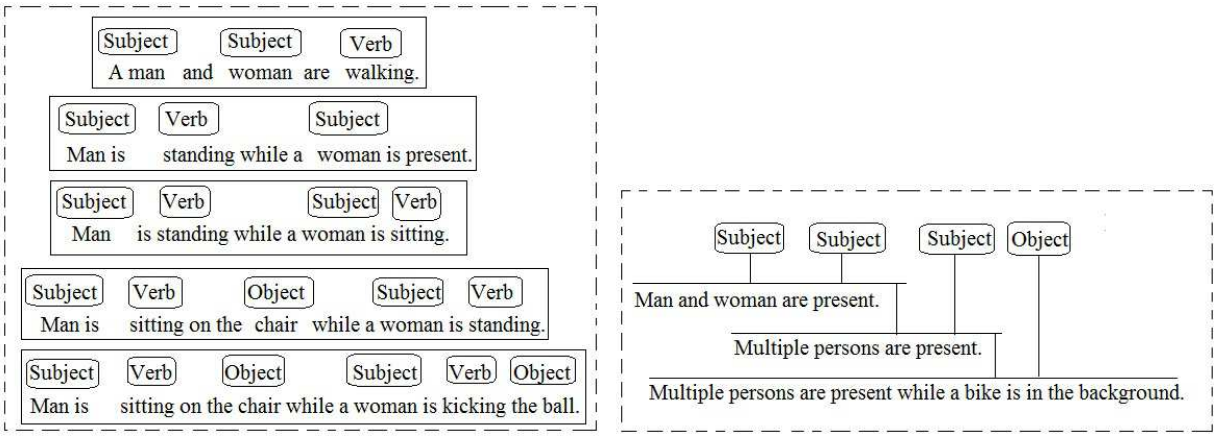


Figure 6: An example for creating a sentence when a single subject is present.



(a) two subjects

(b) multiple subjects

Figure 7: Examples for creating a sentence when a scene contains (a) two or (b) multiple, *i.e.*, more than two, subjects. Only one combination of human interaction is shown, although there can be several scenarios for human interactions.

Hierarchical sentence generation. Figure 6 illustrates an example when a single subject is present. The first block expresses a human subject with the age, gender and the emotion information. The second block contains a verb describing a human action, explaining the relation between the first and the third blocks. The spatial relation between the subject and other objects can be presented. The third block captures objects that may be either a part of the background or a target for subject's action. This three-block mechanism allows both a human and a non-human subject.

When expressing activities by multiple humans a CFG based presentation is defined. Ryoo and Aggarwal used a CFG for hierarchical presentation of human actions where complex actions were composed of simpler actions [26]. In contrast, we allow a scenario where there is no interaction between humans, *i.e.*, they perform individual actions without a particular relation — imagine a situation whereby three people are sitting around a desk while one person is passing behind them. The approach is hierarchical in the sense that we start with creating a single human grammar, then build up to express interactions between two or more than two humans as a combination of single human activities.

Figure 7(a) presents examples involving two subjects. There can be three scenarios; firstly two persons interact with each other to create a common single activity (*e.g.*, a 'hand shake' scene). The second scenario involves two related persons performing individual actions but they do not create a single activity (*e.g.*, both persons are walking together, sitting or standing). Finally two persons happen to be in the same scene at the same time, but there is no particular relation between them (*e.g.*, one person walks, passing behind the other person sitting on a chair). Figure 7(b) shows an example that involves an extension of a single human scenario to more than two subjects. Similarly to two-human scenarios, multiple subjects can create a single action, separate but related actions, or independent and unrelated actions altogether.

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 the HLF is moving — attach '*moving*' (verb) in E
— if one HLF is moving and the other is static — attach '*moving*' with the moving HLF in E, and static HLF is considered a part of the background
— **transfer** $E \rightarrow F$ **and clear** E

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

Figure 8: Procedure for generating natural language descriptions for individual frames.

Procedure for sentence generation. Figure 8 outlines the procedure for generating natural language descriptions for individual frames. First, subject(s) should be identified; there can be one, two or many (*i.e.*, 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. The age, gender and the emotion are selected as an adjective for each subject. The action and the pose (verb) are 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 the background. The most likely preposition (the spatial relation) is calculated and inserted between the subject, verb and objects.

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

3.3. Sample Scenario for Description Generation

When creating natural language description of a scene, visual features are identified first. Secondly, spatial relations between humans and other objects are calculated. Finally, the scene can be described hierarchically based on individual descriptions of humans and their spatial relations. We illustrate this process using an indoor scene (Figure 9) as an example where two humans are present.

Visual HLFs that can be extracted for one subject (a man on the left) are '*human*', '*male*', '*stand*' and '*speak*'; they can be '*human*', '*male*', '*sit*' and '*speak*' for the second (a man on the right). Additionally four '*chairs*' can be

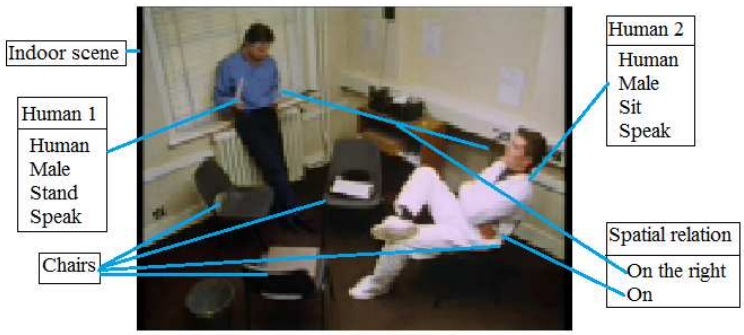


Figure 9: A sample scenario of an indoor scene with two humans. Boxes around the image list visual features (that can be extracted), from which natural language descriptions are created for an individual human. Finally the scene can be described hierarchically based on individual descriptions.

identified. It is possible to calculate their spatial relations; *e.g.*, 'the second subject is on the right of the first subject', and 'the second subject is sitting on the chair'. The scene setting is identified as '*indoor*'. Based on a set of visual features and the spatial relations, a number of sentences can be created. For the first subject, '*subject + verb*' is selected as the core structure. Since there is no information about a cardinal or adjective, '*determiner +*' will be the starting phrase:

A man is standing and speaking;

For the second subject, '*subject + verb + object*' is the appropriate core structure. Once again there is no information about a cardinal or adjective, thus '*determiner +*' is the starting phrase:

A man is sitting on the chair and speaking;

Further, a sentence can be created using the rest of the chairs as a subject. Because there is no movement information the core structure may be '*subject*'. The cardinal (*i.e.*, four) is known, hence the sentence starts with '*cardinal +*':

Four chairs are present;

The scene setting is presented using the fixed template:

This is an indoor scene;

Interaction between human subjects is explored in order to judge if sentences for individual humans can be joined together. To that end, we assign roles for a 'main sentence' and a 'sub sentence', aiming to merge the sub sentence into the main sentence. Selection of the role is made in the following steps:

1. A sentence having the larger number of HLFs is the main sentence;
2. If the numbers of HLFs are the same, then a sentence having the core structure of '*subject + verb + object*' will have higher priority than one with '*subject + verb*'; the latter will be selected over one with the '*subject*' only structure;
3. If both sentences possess the same core structure, then they are just glued together using function words.

As for the indoor scene with two humans in Figure 9, there exists more information for a man on the right; its core sentence structure is '*subject + verb + object*' with five HLFs. In comparison the structure is '*subject + verb*' for a man on the left, consisting of four HLFs. Using the above rules, the sentence for the second human subject ('*A man is sitting on the chair and speaking*') is selected as the main while the other ('*A man is standing and speaking*') is considered as the sub sentence. The combined description for the two human scene is

A man is sitting on the chair and speaking to a man standing on his left;

incorporating the spatial relation. It is worth noting that, according to a set of templates provided, a combined sentence such as '*A man is sitting on the chair while a woman is standing*' can be generated, however '*A woman is standing while a man is sitting on the chair*' is not possible due to the main/sub sentence configuration.

Figure 10: A video montage showing a woman walking in the outdoor scene. The scene can be described using natural language; the resulting expression depends on visual features successfully extracted by the image processing techniques.

4. Dealing with the Varying Number of Visual Features

In Section 2 we outlined implementation of visual feature extraction schemes and their shortcomings. It resulted in missing and erroneous features, the amount of which could vary depending on the quality of video data as well as the image processing methodology. This section focuses on approaches to natural language generation that aim to address the issue of potentially missing visual features. A number of templates are prepared, accommodating the different number of HLFs identified. The framework is scalable in that it is able to process video segments in different genres from the original seven categories presented in Section 2, producing syntactically correct and well structured sentences without a special arrangement.

4.1. Description Depends on Extracted Visual Features

The larger the number of visual features correctly identified, the more precise the video contents can be described. Unfortunately it is not feasible nor practical to produce a full list of HLFs. The current image processing technologies can handle a limited scope of visual features; *e.g.*, a couple of clearly displayed humans can be identified, but the success rate goes down sharply once they are occluded, unclear, side-viewed, or a crowd of humans. The primary reason for choices of seven categories made in Section 2 is to restrict the number of image processing techniques to be applied on a video stream, while being able to produce a semantically meaningful descriptions based on the limited number of extracted features.

Figure 10 is a video montage showing a woman walking in the outdoor scene. In the framework developed, a binary decision is made whether a scene is captured in indoor or outdoor, hence the scene setting is always present regardless of any other HLFs being identified. It, for example, results in the following baseline description:

This is an outdoor scene;

which consists of a single feature (*i.e.*, '*outdoor*'). If a human can be identified, then the description is revised:

There is a human in an outdoor scene;

with two visual features (*i.e.*, '*human*' and '*outdoor*'). Suppose further that a human gender can be identified (*e.g.*, '*female*'), the expression changes to:

There is a woman in an outdoor scene;

where a word '*woman*' was derived from two HLFs, '*human*' and '*female*'. If the human's action is recognised:

A woman is walking in an outdoor scene;

accumulating four HLFs. With further identification of visual features in the background the expression evolves:

A woman is walking while there is a motor bike in the background;
This is an outdoor scene; (6 features)

A woman is walking while there are two humans in the background;
This is an outdoor scene; (7 features)

A woman is walking while there is a man and a woman in the background;
This is an outdoor scene; (9 features)

Figure 11: A video montage showing an indoor dining scene with four people.

The last expression consists of nine HLFs with two '*women*' and a '*man*', each requires two HLFs, *i.e.*, identification of a human and a gender. Note that '*background*' is treated as one of HLFs.

Figure 11 presents another example showing an indoor dining scene with four people. With a single feature ('*indoor*'), the baseline statement is

This is an indoor scene;

Suppose that all four people are identified:

There are four persons in an indoor scene;

counting five HLFs (four '*humans*' and '*indoor*'). It requires correct identification of nine features in order to create the expression like this:

There are two men and two women in an indoor scene;

Finally, by extracting their action ('*sitting*') correctly, the expression becomes

Two men and two women are sitting in an indoor scene;

with 13 HLFs all together.

The number of visual features may vary, however it is possible to create grammatical expressions given some video content. We can see the 'usefulness' of natural language expressions improves with the number of HLFs incorporated. Nevertheless machine generated descriptions are inferior to manual annotations, due to the fundamental lack of a visual feature set that can be identified by the current level of image processing technologies. Human annotators can state many more features such as '*dining scene*' based on foods and plates on the table, a '*painting*' on the wall as well as people's clothing and facial expressions.

4.2. Humans can be Absent

The framework creates a description of a human (or humans) as a centre of focus. Suppose a human is absent, or failed to be extracted, the scene can be explained on the basis of non-human objects. The current development allows only a small number of objects such as a car, a table and a TV-monitor. For example, if a single chair is identified in a room, the following expression can be created:

There is a chair in an indoor scene;

If we also fail to extract a non-human object, any movement can be described. One example in this category may be a scene with an animal as a central object; the current development does not incorporate an algorithm for identification of any animal, however its movement can be detected.

This is an outdoor scene; There is some movement;

can be a simple description created in this situation.

(a) (b) (c)

Figure 12: Dealing with missing and erroneous visual features. Some important visual features, such as action, could not be identified in (a) and (b). With (c), a person on the left was erroneously identified as a female.

4.3. Discussion — Dealing with Missing Visual Features

The major challenge of the task is not only that the number in our inventory of image processing schemes is restricted, but that the schemes may not always be able to extract features successfully, thus resulting in missing, or sometimes erroneous, features. Figure 12 shows collection of image processing errors. Two humans, a table and their spatial relation were successfully extracted in Figure 12(a), however their action (*sitting*) was not identified, hence the following expression could be produced:

There is a table between two persons;

This example could have stated that it was a dining scene if we were able to detect either 'foods on the table' or 'eating action' (unfortunately such algorithms were not available). In Figure 12(b), three people were identified but their actions, genders, and the fourth person on the most right were not detected, resulting in the description:

Many persons are present;

The word '*many*' implies that more than two humans were identified.

Clearly the problem is that, with the current development of image processing technologies, detectable visual features are sparse and it is unavoidable in many cases that some important HLFs are not extracted. The approach, presented in this section, creates a natural language description depending on the number of (correctly or incorrectly) extracted features. It is able to address the above problem by providing a set of templates, a choice of which can accommodate a variable number of extracted HLFs. On the other hand, the developed framework does not have a mechanism for detecting or correcting incorrectly identified HLFs. Figure 12(c) shows one such example:

There is a smiling man and a woman;

where the latter is incorrectly identified as *female*, leading to the erroneous description of the image contents.

Although not implemented in the current work, a potential approach to recovering some missing features is the use of context information, available with various means. A visual context model can be built. Alternatively, we may consider application of the conventional natural language processing and information extraction/retrieval measures such as n-gram language modelling, probabilistic parsing, and a bag of words scheme [27, 28]. The idea can be effective for detection and correction of erroneously identified visual features. Suppose that a human's action could not be identified but the spatial relation indicated the human was on a chair. Then it would be likely that a human was *sitting* on the chair, thus recovering the missed or incorrectly identified action.

5. Evaluation

In recent work [14], we investigated 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. On the other hand, evaluation in this paper focuses on the quality of descriptions, affected by the number of successfully extracted visual features. To this end the task-based evaluation is conducted, and critical observations are made for various categories of videos.

(a) Action — two humans in the park (b) Close-up — a man talking to someone

(a) Action — two humans in the park

'detailed' description: A man is standing while a woman is sitting on a bench; Both of them look serious; Both of them are wearing formal clothes; A bus passes by in the background; It is an outdoor scene; (#HLFs: man(2), stand(1), woman(2), sit(1), serious(2), formal clothes(2), bus(1), outdoor(1))
'simple' description: A man is talking to a woman; (#HLFs: man(2), talk(1), woman(2))
'basic' description: A man is present; (#HLFs: man(2))

(b) Close-up — a man talking to someone

'detailed' description: A serious man in a formal suit is talking to someone; A police man is standing behind him; Two women wearing hats are standing behind him; (#HLFs: serious(1), man(2), formal suit(1), talk(1), someone(1), police(1), stand(1), women(2), ...)
'simple' description: A serious man is speaking; There is a person in the background; (#HLFs: serious(1), man(2), speak(1), human(1))
'basic' description: A man is present; (#HLFs: man(2))

Figure 13: Video frames from each of the original video categories. (a) is seen in '*20041101_160000_CCTV4_DAILY_NEWS_CHN*'. (b) is seen in '*MS206410*' from the 2007 BBC rushes videos summarisation task.

5.1. Task-based Evaluation

Evaluation was conducted by human subjects finding a video that corresponded to a machine generated natural language description. The purpose of the task was to measure the 'usefulness' of descriptions created from the different number of visual features identified for the same set of video clips. This evaluation was designed as follows: firstly we selected 12 shots, six each from the 'Action' and the 'Close-up' categories. We refer to them as $A_1, \ldots, A_6$ and $B_1, \ldots, B_6$, respectively. For each shot, a set of three descriptions were generated by machine where

'detailed' description: with eight or more (8+) visual HLFs;

'simple' description: with five (5) HLFs;

'basic' description: with only two (2) HLFs.

For this evaluation, visual HLFs were selected manually so that the specified number of predicates was presented for the natural language generation stage. We made effort here to spread the selection of HLFs; *e.g.*, two humans, two genders, and one action were selected for one 'simple' description, one human, one gender, one action, one age and one object were selected for another 'simple' description. Figure 13 shows examples, one each from the 'Action' and the 'Close-up' categories, comparing three types of descriptions prepared.

Nine human subjects took part in the experiments — we numbered them as subjects $1, 2, \ldots, 9$. Each subject was provided with textual descriptions, one at a time as a query, and 20 video segments. The same set of 20 video clips were repeatedly used, a half of which were selected from the 'Action' category and the rest were from the 'Close-up' category. This resulted in a pool of candidates, consisting of clearly distinctive videos (between categories) and videos with subtle differences (within the same category). They performed the task with the following schedule:

subject	queries
1,2,3	'detailed' description for clips A_1, A_2, B_1, B_2; 'simple' for A_3, A_4, B_3, B_4; 'basic' for A_5, A_6, B_5, B_6
4,5,6	'simple' description for clips A_1, A_2, B_1, B_2; 'basic' for A_3, A_4, B_3, B_4; 'detailed' for A_5, A_6, B_5, B_6
7,8,9	'basic' description for clips A_1, A_2, B_1, B_2; 'detailed' for A_3, A_4, B_3, B_4; 'simple' for A_5, A_6, B_5, B_6

This arrangement was needed to avoid a potential bias so that subjects should not see, *e.g.*, 'detailed' and 'simple' descriptions from the same video clip because they often found the task easier when they saw a description of the same video clip for the second time.

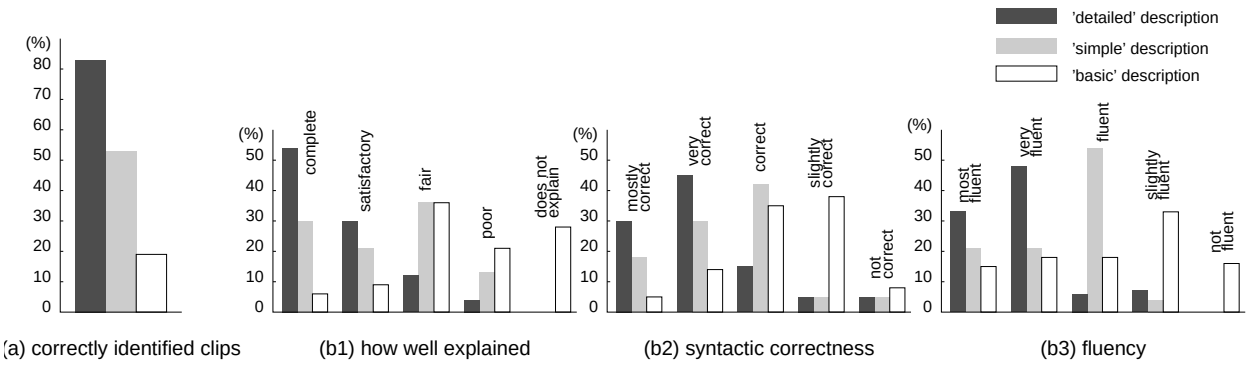


Figure 14: Graph (a) presents the task-based evaluation. Graphs (b1), (b2) and (b3) show the outcomes for the questionnaire.

Figure 14 presents the outcomes from this task-based evaluation. Firstly we counted the number of correctly identified video clips for 'detailed', 'simple' and basic' descriptions. Although the result was not very surprising, Figure 14(a) clearly shows that subjects had a better chance of identifying correct videos when 'detailed' descriptions were provided as a query. As the 'basic' descriptions in Figure 13(a) and (b) clearly show, it was a weakness of the algorithmic generation that it could lead to an identical expression for different videos when the number of visual features were small. 'Simple' descriptions did better although they were sometimes unable to provide clear clues that differentiate between multiple videos.

Secondly, upon completion of video clip search for each query, a questionnaire was set asking

1. how well the video stream was explained by the description, rating from 'explained complete', 'satisfactory', 'fair', 'poor', or 'does not explain';
2. syntactic correctness rating with five scales;
3. fluency rating with five scales.

According to Figure 14(b1) the majority considered the 'detailed' descriptions explained the video well, while 'basic' descriptions provided insufficient information to achieve the task. It is interesting to observe that many subjects felt the 'detailed' descriptions were also more fluent and syntactically correct than 'simple' or 'basic' descriptions — see Figures 14(b2) and (b3).

5.2. Critical Observation of Annotation Samples

Figure 15(a) presents a set of three annotations for 'Action' video shown in Figure 13(a), consisting of one machine output and two manual annotations, the latter being selected from 13 manual annotations created. The main interest in this category was to find humans and their activities. Successful identification of humans and their genders, actions (*e.g.*, '*sitting*', '*standing*') led to a well-phrased machine description. The bus and the car in the background were also identified. On the other hand, other visual HLFs such as their age and emotion were not recognised. The '*speaking*' action was also not recognised because their facial areas were too small. There are additional information found in the manual annotations. The woman was presented as '*young*' and sitting on the '*chair*'. Human clothing ('*formal clothes*') was noted and the location ('*park*') was reported. Finally, with the current implementation of the approach, machines could not identify interaction between multiple humans. This was a clear strength of annotations created by humans (*e.g.*, '*two persons are talking*').

Figure 15(b) shows annotations for the 'Close-up' video category whose frames can be seen in Figure 13(b). Because of the large size of human faces there was a chance that human's age, gender and emotion information could be extracted. With this particular example, the machine generated description was able to capture human gender and the emotion ('*serious*') while the age information was not recognised. Humans in the background were also successfully identified, although information related to their age, gender or emotions was not recognised. Manual annotations explained the video sequence with a further detail, such as clothing, hair colour and the 'windy' outdoor scene setting. They also explained the identity of a person in the background as a policeman and the clothing information (*e.g.*, '*wearing hats*').

(a) Action — two humans in the park
machine: A woman is sitting to the left of a standing man; There is a bus in the background; There is a car in the background;
manual 1: A young woman is sitting on a chair in a park and talking to man who is standing next to her;
manual 2: Two persons are talking; One is a man and other is a woman; The man is wearing formal clothes; The man is standing and the woman is sitting; A bus is travelling behind;

(b) Close-up — a man talking to someone
machine: A serious man is speaking; There are persons in the background;
manual 1: A man with brown hair is talking to someone; He is standing at some outdoor place; He is wearing formal clothes; He looks serious; It is windy;
manual 2: A man is talking to someone; He is wearing a formal suit; A policeman is standing behind him; Some people in the background are wearing hats;

Figure 15: A set of one machine and two manual annotations for the 'Action' and the 'Close-up' video categories. Sample frames for (a) and (b) are shown in the task-based evaluation in Figure 13.

We also tested the framework with the rest of categories (*i.e.*, 'News', 'Meeting', 'Grouping', 'Traffic' and 'Indoor/Outdoor' videos — samples are not shown due to the limited space). There were two issues that deserved some discussion. Firstly, because the framework was developed for 'description of humans' activities', these five categories were out of the target to a various extent. For example, the main theme in the 'News' category was 'presentation of news'; although news often involved humans and their activities, further knowledge would have been required in order to derive that they were humans 'presented in news'. Machines were able to identify HLFs such as humans and their activities, however it was apparent that manual annotations had clear advantage when the task involved even higher concepts such as news. 'Meeting' was another example; the machine annotation could describe little more than the existence of humans while the most of manual annotations stated that it was actually a scene from a meeting.

Secondly, we were always able to produce some description even for videos in which a human was not the major component, or not even present, *e.g.*, many videos in the 'Traffic' and 'Indoor/Outdoor' categories belonged to this case. For example in a 'Indoor/Outdoor' video, humans and many other detectable objects were typically absent and movement was also minimal. Not surprisingly manual annotations were rich, sometimes verbose, with various objects (*e.g.*, pavilion, trees) and even with higher concepts (*e.g.*, park). Although these HLFs and concepts were not targeted during the development, the approach was able to generate a default expression such as '*this is a static scene*', clearly demonstrating the positive effect of the framework that chose the expression depending on the number of identified visual features.

6. Scalability Study

This section explores the scalability of the natural language generation framework when the video genre is different from the original seven categories described in Section 2. The approach has been built up with these seven categories in consideration and the humans were the central focus of creating descriptions. In particular, image processing algorithms were prepared such that a restricted number of important visual features could be extracted for videos in these categories. Investigation here concerns the question: can the framework handle videos from different genres? To this end the dataset is extended to incorporate videos from the five new categories. The first four categories ('Costume', 'Crowd', 'Sports', and 'Violence') normally include humans and their activities as features for interest, while in the fifth category ('Animal') the focus is not on humans:

Costume: They are video clips from films and TV dramas, containing costumes, sets and properties that capture the ambiance of a particular era such as historical Victorian, ancient Roman civilisation, *etc.*

Crowd: A large number of humans with some activities are seen in a single screen, *e.g.*, people in a procession raising slogans and holding banners, people waiting for a train on the rush-hour platform.

Sports: Sports videos have special scene settings (*e.g.*, football pitch), human's clothing (*e.g.*, swimwear), and equipment (*e.g.*, cricket bat) information.

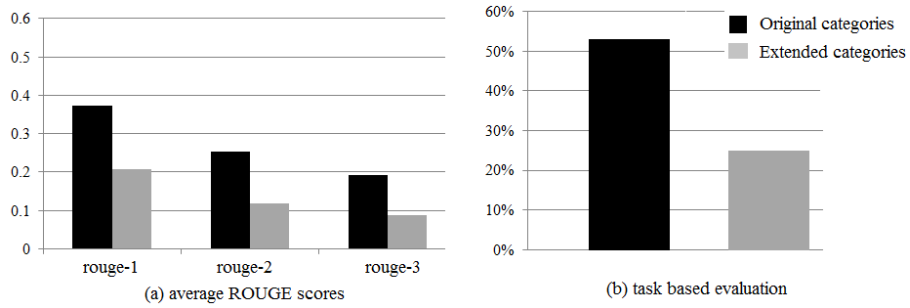


Figure 16: Graph (a) shows comparison of the average ROUGE scores between the original seven and the extended five video categories. Calculation was made by averaging similarity scores between a machine generated description and the individual manual annotations. ROUGE 1-3 shows *n*-gram overlap similarity between the reference and the model descriptions. Graph (b) presents correctly identified videos, comparing between the original seven and the extended five video categories.

Violence: They are characterised by objects such as guns and army tanks. Fire, smoke and damaged buildings are also frequent.

Animal: Animals are the centre of focus. Scene settings such as a park and a room are also frequent.

For each of the above five categories, four short video clips were selected from the HLF extraction task (2004) and the BBC rushes task (2008) in TREC Video. They were described manually by five annotators. Following the standard procedure visual features were extracted, upon which natural language descriptions were created. Figure 17 presents five frames, one each from five extra video categories. Each frame image is accompanied by a set of three descriptions, one machine output and two manual annotations, the latter were selected from five manual annotations created.

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 (Recall Oriented Understudy for Gisting Evaluation), widely used for evaluating automatic summarisation [32], to calculate the overlap between manual and machine generated annotations. In general, higher ROUGE score indicates closer match between them; a score of '1' means the perfect match while scores close to '0' suggest that matches occurred in a small portion of expressions and may be only accidental. Figure 16(a) shows comparison of the average ROUGE scores between the original seven and the extended five video categories.

Manual annotations were often subjective, and dependent on one's perception and understanding, that could be affected by the educational and professional background, personal interests and experiences. Nevertheless reasonable similarity scores were calculated between machine generated descriptions and manual annotations for the original seven video categories. In comparison, all measures in ROUGE indicated a significant decline when scores for the extended five categories were calculated. This is not very surprising, mainly because the image processing algorithms were targeted the original categories. We were not able to extract visual features that were essential when presenting video contents for the extended categories (*e.g.*, human clothing, animal type). Additionally, some of the extended categories required a level of interpretation (*e.g.*, type of costume, sports, violence) that was much higher than the typical visual HLFs that could be identified by the current image processing technology. Although grammatically correct, descriptions created for the extended categories were surely affected by the lack of development in these aspects.

6.2. Task-Based Evaluation

For the task-based evaluation, human subjects were instructed to find a video clip that corresponded to a natural language description. The evaluation was designed as follows: each subject was provided with one textual description and 20 video segments at one time. The procedure was different from the one presented earlier in Figure 14 in that, for this task, visual HLFs were extracted automatically (hence there existed potentially missing and erroneous features) then descriptions were created by the algorithm. Five human subjects conducted this task searching a corresponding

(a) Costume (b) Animal

(a) Costume — three humans wearing a mask and ancient Roman dress
machine: There are three persons; This is an indoor scene;
manual 1: Three persons wearing Greek garments are walking;
manual 2: Three men in Roman clothes are wearing masks and holding big and heavy sticks; It looks like old army parade scene;

(b) Animal — a whale
machine: This is an outdoor scene; There is some movement;
manual 1: A whale is dancing in the water; Some other whales are also dancing;
manual 2: A big fish is going up and down in the water; There are many other fishes shown;

Figure 17: A set of one machine and two manual annotations for the 'Costume' and the 'Animal' video categories. (a) is seen in '*MRS157475*' from the 2008 BBC rushes videos. (b) is seen in '*20041104_220001_CNN_AARONBROWN_ENG*' from the 2004 HLF extraction task.

video for each of five descriptions. They did not involve creation of the dataset, hence they saw these videos for the first time.

Figure 16(b) presents results for correctly identified videos, comparing between the original seven and the extended five video categories. By reading machine generated descriptions in the original categories, subjects were able to find the corresponding videos with 53% correct, which could be considered a significant improvement over selection by pure chance (5%). With descriptions created for the extended categories '25% correct' means that, although clearly better than pure chance, it was more difficult to identify the the corresponding video due to the lack of clear expressions for the video contents. For example, a description such as '*This is an outdoor scene; There is some movement*' can be applied for many candidates.

6.3. Annotation Samples and Discussion

Figure 17(a) was a frame from the 'Costume' video category, showing three humans wearing a mask and ancient Roman dress. Firstly it was erroneously recognised as an indoor scene by a machine. Despite of wearing a mask, three humans were successfully identified by the image processing step however the rest of visual features could not be extracted. On the other hand, human annotators paid full attention to sticks in hand, facial masks and clothing although one considered it was Greek instead of Roman. Unfortunately none of these features could have been extracted because image processing algorithms were not available for these HLFs. Among these, human dress could have been a particularly useful feature if we were able to process them. In general our experience suggests that human related features such as age, gender and emotion were not as accurately identified as we hoped due to the diversity in human appearance. Expression for higher-level scene settings such as '*village*' or '*army parade*' should have to be left for future development.

There are many videos showing shots of animals and their activities. Figure 17(b) is a frame from the 'Animal' video category, showing a whale jumping over the water. Manual annotations focused on the animal, its appearance and movement present in the segments (*e.g.*, '*A big fish is going up and down*'). Location of the scene (*e.g.*, '*in the park*', '*under the tree*') was also frequently stated. In comparison, a machine generated description mainly presented the scene setting (indoor/outdoor) and existence of any moves. Identification of the animal type was not conducted although it could have been feasible for a number of popular animals (*e.g.*, dog, cat). The algorithm always set a central role for humans if there were both humans and animals in the scene.

6.4. Comparison of the Work with Recent Studies

In Section 1 we reviewed that there have been a number of recent studies for natural language production from visual contents. We did not have a sufficient ground to make any quantitative comparison (*e.g.*, some study worked at

a video frame level while we processed frame sequences), nevertheless it is still worth presenting a brief qualitative discussion as a last part of the scalability study. Spatial and temporal relations are useful components when producing rich expressions. The work by Krishnamoorthy *et al.* [8] did not incorporate spatial features. Suppose we were to apply their approach for the 'Action' video shown in Figure 13, it might create an expression '*a man is talking to a woman*' but their spatial relation could not be known. Other work focused on a specific genre of videos. For example, work by Brendel *et al.* [29] and also one by Morariu and Davis [30] were concerned with the basketball players scenario. Using their approaches, it was not likely that any fluent description could be produced for the dataset in this paper. A training step would be required to learn a storyline in order to utilise the approach by Gupta *et al.* [31].

Among those, we were more successful with the approaches by Das *et al.* and by Barbu *et al.* Firstly, Das *et al.* [12] extracted important keywords based on low-level visual features. They found related concepts for the keywords from textual vocabularies using a latent topic model, where concepts were categorised as nouns or verbs. For the 'Action' video of Figure 13, their code produced:

A man is talking to a woman, an outdoor scene; One man is standing and one women is sitting;

They argued the benefits of top-down and bottom-up approaches, but their methodology was limited by the number of extracted keywords and related concepts. Another approach by Barbu *et al.* in 2012 [13] was conceptually similar to the work presented in this paper, in that they generated descriptions in a bottom-up fashion. Initially objects were extracted using image processing techniques, thus creating nouns; actions were treated as verbs and spatial relations between individual nouns were expressed using prepositions. For the same 'Action' video, their code created the following expression:

A person is standing at right of bench and a woman is sitting at left;

On the other hand they did not handle commonly observed cases such as scene settings and presence of multiple objects.

Despite the similarity in natural language generation flow with Barbu *et al.* and other studies, our work is different in that we focus on a scheme to handle missing visual features when generating descriptions. This may be more apparent by looking at some counter examples. Das *et al.* created their own dataset named *YouCook* from *YouTube* videos. For a sample video presented in Figure 6 of Das *et al.* [12], descriptions produced were:

by Das *et al.*: *A person is cooking with a bowl and stovetop;*
by our framework: *A person is facing camera and speaking;*

This dataset was related to the kitchen scenes and Das *et al.* were able to generate logical descriptions explaining the cooking scene with cookwares such as a bowl and a stovetop. On the other hand our approach did not identify any kitchen-wares, as it was not developed for that specific genre in mind, but it was able to produce the compact description within its remit, *e.g.*, person's action and the scene setting.

Language description could also be created for a sample video presented in Figure 5 of Barbu *et al.* [13]:

by Barbu *et al.*: *The upright person to the right of the motorcycle went away leftward;*
by our framework: *A person is sitting on a motorbike; A man is running on the right side of that person; This is an outdoor scene;*

Barbu *et al.* focussed on the temporal information of the foreground object. On the other hand our approach found the human action, spatial relation between two humans and the scene setting. Our approach might not create a full description of many scenes, nevertheless some description could be produced for any video, a scene setting at the very minimum. Recall the work by Morariu and Davis [30] where they built their approach for one-on-one basketball scenes from varying camera positions. We did not have a basketball recogniser but it was able to produce an expression '*Two men are walking in an outdoor scene*' (see Figure 1 from [30]), indicating that the framework was flexible to accommodate a range of variations regardless of video topic or genre.

7. Conclusions

This study addressed creation of natural language descriptions for visual contents present in video streams. We applied conventional image processing techniques to extract visual features, which were then converted into natural

language expressions using CFG based templates. Since feature extraction processes were erroneous at various levels, we explored approaches to create a coherent expression assuming that any visual features could fail to be identified.

When a list of image processing techniques was carefully selected, automatic evaluation resulted in good similarity between manual annotations and machine generated descriptions. The task-based evaluation indicated that produced natural language descriptions were useful for correct identification of the corresponding video stream. On the other hand, further experiments with an extended data set revealed the weakness of the scheme — when we do not have a sufficient list of image processing techniques, thus their visual features were not counted, we could only produce expressions that were, although grammatical, ambiguous for a video information retrieval purpose.

It is clear that, although processing time can become an issue, the scheme is able to produce useful descriptions of the contents by creating a large list of image processing techniques for identification of important visual features in various genre of videos. In particular we found that clothing information could be an interesting feature when humans were the central focus. The outcome of the study also indicated that a higher level of processing (than a conventional visual HLF extraction) would be needed for some video genres in order to create a useful expression of the video contents. This may include spatial and temporal relations between objects, and inference on higher concepts, examples of which were presented by video genre such as Costume and Violence.

This paper focused on the framework for handling potentially missing visual features. Another important issue relating to this work is existence of erroneously identified features (*e.g.*, identification of '*male*' instead of '*female*'). To address this problem, temporal information may be considered. For example a person may be identified as '*male*' initially, but as '*female*' later on — this requires a decision of which visual HLF is more likely. Use of context information may also be a viable idea. For example in a traffic scene, we are more likely to see a type of vehicle than a type of office equipment. We leave these as topics for future exploration.

Acknowledgment. The authors would like to thank Guy Brown for proofreading the manuscript.

References

[1] A. Torralba, K. P. Murphy, W. T. Freeman, M. A. Rubin, Context-based vision system for place and object recognition, in: Proceedings of the ICCV, 2008.

[2] R. M. Bolle, B. L. Yeo, M. M. Yeung, Video query: research directions, IBM Journal of Research and Development 42 (2) (2010) 233–252.

[3] P. Over, A. F. Smeaton, P. Kelly, The TRECVID 2007 BBC rushes summarization evaluation pilot, in: TRECVID BBC Rushes Summarization Workshop, 2007.

[4] P. Baiget, C. Fernández, X. Roca, J. Gonzàlez, Automatic learning of conceptual knowledge in image sequences for human behavior interpretation, Pattern Recognition and Image Analysis (2007) 507–514.

[5] M. W. Lee, A. Hakeem, N. Haering, S. C. Zhu, SAVE: a framework for semantic annotation of visual events, in: Proceedings of the CVPR Workshop, 2008.

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

[7] Y. Yang, C. L. Teo, H. Daumé III, C. Fermüller, Y. Aloimonos, Corpus-guided sentence gereration of natural images, in: Proceedings of the EMNLP, 2011.

[8] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, S. Guadarrama, Generating natural-language video descriptions using text-mined knowledge, in: Procedings of the AAAI Conference, 2013.

[9] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko, YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in: Procedings of the ICCV, 2013.

[10] F. Metze, D. Ding, E. Younessian, A. Hauptmann, Beyond audio and video retrieval: topic-oriented multimedia summarization, International Journal of Multimedia Information Retrieval 2 (2) (2013) 131–144.

[11] H. Yu, J. M. Siskind, Grounded language learning from video described with sentences, in: Proceedings of the Annual ACL Meeting, 2013.

[12] P. Das, C. Xu, R. F. Doell, J. J. Corso, A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching, in: Procedings of the CVPR, 2013.

[13] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, Z. Zhang, Video in sentences out, in: Procedings of the UAI, 2012.

[14] M. U. G. Khan, Y. Gotoh, Generating natural language tags for video information management (in review).

[15] P. Kuchi, P. Gabbur, P. S. Bhat, S. S. David, S. Smieee, Human face detection and tracking using skin color modeling and connected component operators, IETE Journal of Research 48 (3-4) (2002) 289–293.

[16] I. Maglogiannis, D. Vouyioukas, C. Aggelopoulos, Face detection and recognition of natural human emotion using Markov random fields, Personal and Ubiquitous Computing 13 (1) (2009) 95–101.

[17] H. S. Chen, H. T. Chen, Y. W. Chen, S. Y. Lee, Human action recognition using star skeleton, in: ACM International Workshop on Video Surveillance and Sensor Networks, 2006.

[18] A. Sundaresan, A. R. Chowdhury, R. Chellappa, A hidden Markov model based framework for recognition of humans from gait sequences, in: Proceedings of the ICIP, 2003.

[19] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features (2001) 511–518.
[20] W. Kim, J. Park, C. Kim, A novel method for efficient indoor-outdoor image classification, Journal of Signal Processing Systems (2010) 1–8.
[21] R. Lienhart, J. Maydt, An extended set of Haar-like features for rapid object detection, in: Proceedings of the ICIP, 2002.
[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, International Journal of Computer Vision 88 (2) (2010) 303–338.
[23] J. R. J. Schirra, G. Bosch, C. K. Sung, G. Zimmermann, From image sequences to natural language: a first step toward automatic perception and description of motions, Applied Artificial Intelligence an International Journal 1 (4) (1987) 287–305.
[24] E. Reiter, R. Dale, Building applied natural language generation systems, Natural Language Engineering 3 (1) (1997) 57–87.
[25] A. Gatt, E. Reiter, SimpleNLG: a realisation engine for practical applications, in: European Workshop on Natural Language Generation, 2009.
[26] M. S. Ryoo, J. K. Aggarwal, Semantic representation and recognition of continued and recursive human activities, International Journal of Computer Vision 82 (1) (2009) 1–24.
[27] J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, in: Proceedings of the SIGIR, 1998.
[28] G. Salton, C. C., Term-weighting approaches in automatic text retrieval, Information Processing and Management 24 (1988) 513–523.
[29] W. Brendel, A. Fern, S. Todorovic, Probabilistic event logic for interval-based event recognition, in: Procedings of the CVPR, 2011.
[30] V. I. Morariu, L. S. Davis, Multi-agent event recognition in structured scenarios, in: Procedings of the CVPR, 2011.
[31] A. Gupta, A. P. Srinivasan, J. Shi, L. S. Davis, Understanding videos, constructing plots learning a visually grounded storyline model, in: Procedings of the CVPR, 2009.
[32] C. Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Proceedings of the ACL Workshop, 2004.