Describing Video Contents in Natural Language

Muhammad Usman Ghani Khan
University of Sheffield
United Kingdom
ughani@dcs.shef.ac.uk

Yoshihiko Gotoh
University of Sheffield
United Kingdom
y.gotoh@dcs.shef.ac.uk

Abstract

This contribution addresses generation of natural language descriptions for human actions, behaviour and their relations with other objects observed in video streams. The work starts with implementation of conventional image processing techniques to extract high level features from video. These features are converted into natural language descriptions using context free grammar. Although feature extraction processes are erroneous at various levels, we explore approaches to putting them together to produce a coherent description. 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 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 a need for formalising video semantics to help users gain useful and refined information relevant to their demands and requirements. Human language is a natural way of communication. Useful entities extracted from videos and their inter-relations can be presented by natural language in a syntactically and semantically correct formulation.

While literature relating to object recognition (Galleguillos and Belongie, 2010), human action recognition (Torralba et al., 2008), and emotion detection (Zheng et al., 2010) are moving towards maturity, automatic description of visual scenes is still in its infancy. Most studies in video retrieval have been based on keywords (Bolle et al., 1998). An interesting extension to a keyword based scheme is natural language textual description of video streams. They are more human friendly. They can clarify context between keywords by capturing their relations. Descriptions can guide generation of video summaries by converting a video to natural language. They can provide basis for creating a multimedia repository for video analysis, retrieval and summarisation tasks.

Kojima et al. (2002) presented a method for describing human activities in videos based on a concept hierarchy of actions. They described head, hands and body movements using natural language. For a traffic control application, Nagel (2004) investigated automatic visual surveillance systems where human behaviour was presented by scenarios, consisting of predefined sequences of events. The scenario was evaluated and automatically translated into a text by analysing the visual contents over time, and deciding on the most suitable event. Lee et al. (2008) introduced a framework for semantic annotation of visual events in three steps; image parsing, event inference and language generation. Instead of humans and their specific activities, they focused on object detection, their inter-relations and events that were present in videos. Baiget et al. (2007) performed human identification and scene modelling manually and focused on human behaviour description for crosswalk scenes. Yao et al. (2010) introduced their work on video to text description which is dependent on the significant amount of annotated data, a requirement that is avoided in this paper. Yang et al. (2011) presented a

Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012, pages 27–35,
Avignon, France, April 23 2012. ©2012 Association for Computational Linguistics

framework for static images to textual descriptions where they contained to image with up to two objects. In contrast, this paper presents a work on video streams, handling not only objects but also other features such as actions, age, gender and emotions.

The study presented in this paper is concerned with production of natural language description for visual scenes in a time series using a bottom-up approach. Initially high level features (HLFs) are identified in video frames. They may be 'keywords', such as a particular object and its position/moves, used for a semantic indexing task in video retrieval. Spatial relations between HLFs are important when explaining the semantics of visual scene. Extracted HLFs are then presented by syntactically and semantically correct expressions using a template based approach. Image processing techniques are far from perfect; there can be many missing, misidentified and erroneously extracted HLFs. We present scenarios to overcome these shortcomings and to generate coherent natural descriptions. The approach is evaluated using video segments drafted manually from the TREC video dataset. ROUGE scores is calculated between human annotated and machine generated descriptions. A task based evaluation is performed by human subjects, providing qualitative evaluation of generated descriptions.

2 Dataset Creation

The dataset was manually created from a subset of rushes and HLF extraction task videos in 2007/2008 TREC video evaluations (Over et al., 2007). It consists of 140 segments, with each segment containing one camera shot, spanning 10 to 30 seconds in length. There are 20 video segments for each of the seven categories:

Action: Human can be seen performing some action (*e.g.*, sit, walk)

Closeup: Facial expressions/emotions can be seen (*e.g.*, happy, sad)

News: Anchor/reporter may be seen; particular scene settings (*e.g.*, weather board in the background)

Meeting: Multiple humans are seen interacting; presence of objects such as chairs and a table

Grouping: Multiple humans are seen but not in meeting scenarios; chairs and table may not be present

Traffic: Vehicles (*e.g.*, car, bus, truck) / traffic signals are seen

Indoor/Outdoor: Scene settings are more obvious than human activities (*e.g.*, park scene, office)

13 human subjects individually annotated these videos in one to seven short sentences. They are referred to as **hand annotations** in the rest of this paper.

3 Processing High Level Features

Identification of human face or body can prove the presence of human in a video. The method by Kuchi et al. (2002) 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 (Maglogiannis et al., 2009). 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[1].

To recognise human actions the approach based on a star skeleton and a hidden Markov model (HMM) is implemented (Chen et al., 2006). 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 (Sundaresan et al., 2003). Further Haar features are extracted and classifiers are trained to identify non-human objects (Viola and Jones, 2001). They include car, bus, motorbike, 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) (Kim et al., 2010).

3.1 Performance of HLF Extraction

In the experiments, video frames were extracted using *ffmpeg*[2], sampled at 1 fps (frame per second), resulting in 2520 frames in total. Most of

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

[2] 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 numerous formats. The default conversion rate is 25 fps. See http://www.ffmpeg.org/

	(ground truth)	
	exist	not exist
exist	1795	29
not exist	95	601

(a) human detection

	(ground truth)	
	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 ground truth, and rows indicate the automatic recognition results. The human detection task is biased towards existence of human, while in the gender identification presence of male and female are roughly balanced.

	(ground truth)					
	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 ground truth, and rows indicate the automatic recognition results. Some actions (*e.g.*, 'standing') were more commonly seen than others (*e.g.*, 'waving').

	(ground truth)				
	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 ground truth, and rows indicate the automatic recognition results.

HLFs required one frame to evaluate. Human activities were shown in 45 videos and they were sampled at 4 fps, yielding 3600 frames. Upon 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 matrix 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 matrix 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 ups, variety of hair styles and wearing hats, veils and scarfs.

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

There were 15 videos where human or any other moving HLF (*e.g.*, car, bus) were absent. Out of these 15 videos, 12 were related to outdoor environments where trees, greenery, or buildings were present. Three videos showed indoor settings with objects such as chairs, tables and cups. All frames from outdoor scenes were correctly identified; for indoor scenes 80% of frames were correct. Presence of multiple objects seems to have caused negative impact on EOH and COH features, hence resulted in some erroneous classifications. The recognition performances for non-human objects were also evaluated with the dataset. We found their average precision[3] scores ranging between 44.8 (table) and 77.8 (car).

3.2 Formalising Spatial Relations

To develop a grammar robust for describing human related scenes, there is a need for formalising spatial relations among multiple HLFs. Their effective use leads to smooth description of visual scenes. 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.

[3] defined by Everingham et al. (2010).

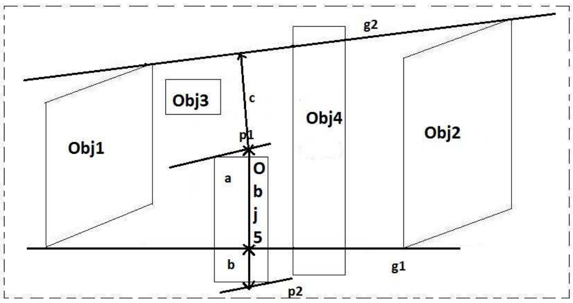

Figure 1: Procedure for calculating the 'between' relation. Obj 1 and 2 are the two reference objects, while Obj 3, 4 and 5 are the target objects.

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 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).

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'. Figure 1 illustrates steps for calculating the three-place relationship 'between'. Schirra et al. (1987) explained the algorithm:

- Calculate the two tangents g_1 and g_2 between the reference objects using their closed-rectangle representation;

- If (1) both tangents cross the target or its rectangle representation (see Obj 4 in the figure), or (2) the target is totally enclosed by the tangents and the references (Obj 3), the relationship 'between' is true.

- If only one tangent intersects the subject (Obj 5), the applicability depends on its penetration depth in the area between the tangents, thus calculate: max(a/(a+b), a/(a+c))

- Otherwise 'between' relation does not hold.

3.3 Predicates for Sentence Generation

<u>Human structure related</u>
human (yes, no)
gender (male, female)
age (baby, child, young, old)
body parts (hand, head, body)
grouping (one, two, many)

<u>Human actions and emotions</u>
action (stand, sit, walk, run, wave, clap)
emotion (happy, sad, serious, surprise, angry)

<u>Objects and scene settings</u>
scene setting (indoor, outdoor)
objects (car, cup, table, chair, bicycle, TV-monitor)

<u>Spatial relations among objects</u>
in front of, behind, to the left, to the right, beside, at, on, in, between

Figure 2: Predicates for single human scenes.

Figure 2 presents a list of predicates to be used 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 2 are for describing single human scenes; combination of these predicates may be used if multiple humans are present.

4 Natural Language Generation

HLFs acquired by image processing require abstraction and fine tuning for generating syntactically and semantically sound natural language expressions. Firstly, a part of speech (POS) tag is assigned to each HLF using NLTK[4] POS tagger. Further humans and objects need to be assigned proper semantic roles. In this study, a human is 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 based on context free grammar (CFG) is implemented for sentence generation. A template is a pre-defined structure with slots for user specified parameters. Each template requires three parts for proper functioning: lexicons, template rules and grammar. Lexicon is a vocabulary containing HLFs extracted from a video stream (Figure 3). Grammar assures syntactical correctness of the sentence. Template rules are defined for selection of proper lexicons

[4] `www.nltk.org/`

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

Figure 3: Lexicons and their POS tags.

with well defined grammar.

4.1 Template Rules

Template rules are employed for the selection of appropriate lexicons for sentence generation. Followings are some template rules used in this work:

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

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

Select 1 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 a second rule

Alternative is used for selecting the most specific template when multiple templates can be used

Elaboration evaluates the value of a template slot

Figure 4 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*'.

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

4.2 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:

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

where T is set of terminals (lexicon) shown in Figure 3, N is a set of non-terminals (usually POS tags), S is a start symbol (one of non-terminals). Finally R is rules / productions 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.

For implementing the templates, *simpleNLG* is used (Gatt and Reiter, 2009). 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.

4.3 Hierarchical Sentence Generation

In this work we define a CFG based presentation for expressing activities by multiple humans. Ryoo and Aggarwal (2009) used CFG for hierarchical presentation of human actions where complex actions were composed of simpler actions. 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.

Figure 5 shows an example for sentence generation related to a single human. This mechanism is built with three blocks when only one subject[5] is present. The first block expresses a

[5] Non human subject is also allowed in the mechanism.

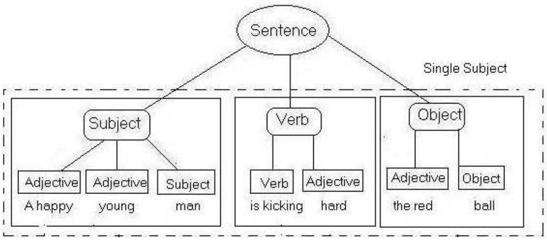

Figure 5: A scenario with a single human.

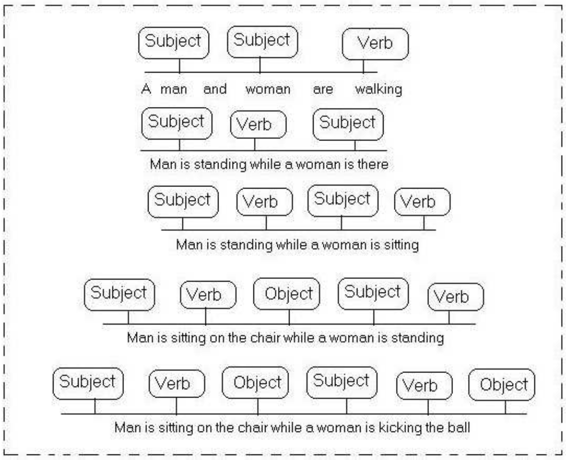
Subject
Subject
Verb
A man and woman are walking
Subject
Verb
Subject
Man is standing while a woman is there
Subject
Verb
Subject
Verb
Man is standing while a woman is sitting
Subject
Verb
Object
Subject
Verb
Man is sitting on the chair while a woman is standing
Subject
Verb
Object
Subject
Verb
Object
Man is sitting on the chair while a woman is kicking the ball

Figure 6: A scenario with two humans.

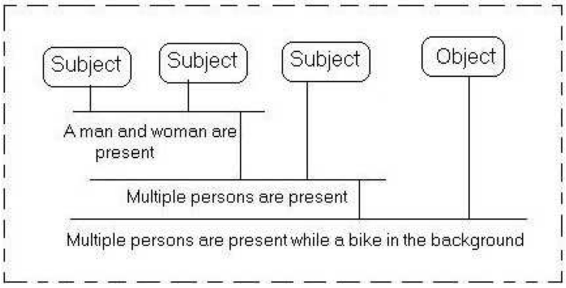
Subject
Subject
Subject
Object
A man and woman are present
Multiple persons are present
Multiple persons are present while a bike in the background

Figure 7: A scenario with multiple humans.

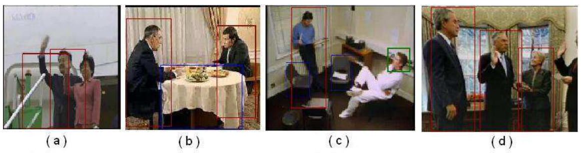

Figure 8: Template selection: (a) subject + subject + verb: '*man and woman are waving hands*'; (b) subject + subject + object: '*two persons around the table*'; (c) subject + verb, noun phrase / subject, noun phrase / subject: '*a man is standing; a person is present; there are two chairs*'; (d) subject + subject + subject + verb: '*multiple persons are present*'.

human subject with age, gender and emotion information. The second block contains a verb describing a human action, to explain the relation between the first and the third blocks. Spatial relation between the subject and other objects can also be presented. The third block captures other objects which may be either a part of background or a target for subject's action.

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 6 presents examples involving two subjects. There can be three scenarios; firstly two persons interact with each other to generate some common single activity (*e.g.*, 'hand shake' scene). The second scenario involves two related humans performing individual actions but they do not create a single action (*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 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 actions, or different actions altogether.

4.4 Application Scenarios

This section overviews different scenarios for application of the sentence generation framework. Figure 8 presents examples for template selection procedure. Although syntactically and semantically correct sentences can be generated in all scenes, immaturity of image processing would cause some errors and missing information.

Missing HLFs. For example, action ('*sitting*') was not identified in Figure 8(b). Further, detec-

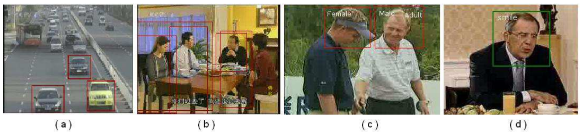

Figure 9: Image processing can be erroneous: (a) only three cars are identified although there are many vehicles prominent, (b) five persons (in red rectangles) are detected although four are present; (c) one male is identified correctly, other male is identified as 'female'; (d) detected emotion is 'smiling' though he shows a serious face.

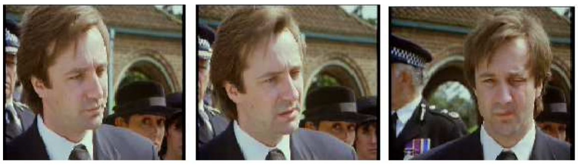

Figure 10: Closeup of a man talking to someone in the outdoor scene — seen in ‘*MS206410*’ from the 2007 rushes summarisation task. **Machine annotation:** A serious man is speaking; There are humans in the background. **Hand annotation 1:** A man is talking to someone; He is wearing a formal suit; A police man is standing behind him; Some people in the background are wearing hats. **Hand annotation 2:** 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.

tion of food on the table might have led to more semantic description of the scene (*e.g.*, ‘*dinning scene*’). In 8(d), fourth human and actions by two humans (‘*raising hands*’) were not extracted. Recognition of the road and many more vehicles in Figure 9(a) could have produced more semantic expression (*e.g.*, ‘*heavy traffic scene*’).

Non human subjects. Suppose a human is absent, or failed to be extracted, the scene is explained on the basis of objects. They are treated as subjects for which sentences are generated. Figure 9(a) presents such a scenario; description generated was ‘*multiple cars are moving*’.

Errors in HLF extraction. In Figure 9(c), one person was found correctly but the other was erroneously identified as female. Description generated was ‘*a smiling adult man is present with a woman*’. Detected emotion was ‘smile’ in 9(d) though real emotion was ‘serious’. Description generated was ‘*a man is smiling*’.

5 Experiments

5.1 Machine Generated Annotation Samples

Figures 10 to 12 present machine generated annotation and two hand annotations for randomly selected videos related to three categories from dataset.

Face closeup (Figure 10). Main interest was to find human gender and emotion information. Machine generated description was able to capture human emotion and background information. Hand annotations explained the sequence more, *e.g.*, dressing, identity of a person as policeman, hair colour and windy outdoor scene settings.

Traffic scene (Figure 11). Humans were absent in most of traffic video. Object detector was able to identify most prominent objects (*e.g.*, car, bus) for description. Hand annotations produced further details such as colours of car and other objects (*e.g.*, flyover, bridge). This sequence was also described as a highway.

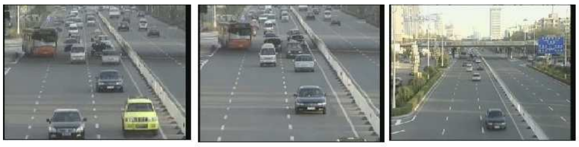

Figure 11: A traffic scene with many vehicles — seen in ‘*20041101_110000_CCTV4_NEWS3_CHN*’ from the HLF extraction task. **Machine annotation:** Many cars are present; Cars are moving; A bus is present. **Hand annotation 1:** There is a red bus, one yellow and many other cars on the highway; This is a scene of daytime traffic; There is a blue road sign on the big tower; There is also a bridge on the road. **Hand annotation 2:** There are many cars; There is a fly-over; Some buses are running on the fly-over; There is vehicle parapet; This is a traffic scene on a highway.

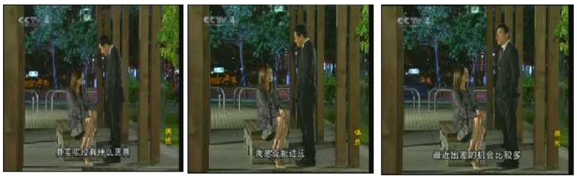

Figure 12: An action scene of two humans — seen in ‘*20041101_160000_CCTV4_DAILY_NEWS_CHN*’ from the HLF extraction task. **Machine annotation:** A woman is sitting while a man is standing; There is a bus in the background; There is a car in the background. **Hand annotation 1:** Two persons are talking; One is a man and other is woman; The man is wearing formal clothes; The man is standing and woman is sitting; A bus is travellings behind. **Hand annotation 2:** Young woman is sitting on a chair in a park and talking to man who is standing next to her.

Action scene (Figure 12). Main interest was to find humans and their activities. Successful recognition of man, woman and their actions (*e.g.*, ‘sitting’, ‘standing’) led to well phrased description. The bus and the car at the background were also identified. In hand annotations dressing was noted and location was reported as a park.

5.2 Evaluation with 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 (Lin, 2004), to calculate the overlap between machine generated and hand annotations. Table 4 shows the results where higher ROUGE score indicates closer match between them.

In overall scores were not very high, demonstrating the fact that humans have different observations and interests while watching the same video. Descriptions were often subjective, de-

	Action	Closeup	In/Outdoor	Grouping	Meeting	News	Traffic
ROUGE-1	0.4369	0.5385	0.2544	0.3067	0.3330	0.4321	0.3121
ROUGE-2	0.3087	0.3109	0.1877	0.2619	0.2462	0.3218	0.1268
ROUGE-3	0.2994	0.2106	0.1302	0.1229	0.2400	0.2219	0.1250
ROUGE-L	0.4369	0.4110	0.2544	0.3067	0.3330	0.3321	0.3121
ROUGE-W	0.4147	0.4385	0.2877	0.3619	0.3265	0.3318	0.3147
ROUGE-S	0.3563	0.4193	0.2302	0.2229	0.2648	0.3233	0.3236
ROUGE-SU	0.3686	0.4413	0.2544	0.3067	0.2754	0.3419	0.3407

Table 4: ROUGE scores between machine generated descriptions (reference) and 13 hand annotations (model). ROUGE 1-3 shows n-gram overlap similarity between reference and model descriptions. ROUGE-L is based on longest common subsequence (LCS). ROUGE-W is for weighted LCS. ROUGE-S skips bigram co-occurrence without gap length. ROUGE-SU shows results for skip bigram co-occurrence with unigrams.

pendent on one's perception and understanding, that might have been affected by their educational and professional background, personal interests and experiences. Nevertheless ROUGE scores were not hopelessly low for machine generated descriptions; Closeup, Action and News videos had higher scores because of presence of humans with well defined actions and emotions. Indoor/Outdoor videos show the poorest results due to the limited capability of image processing techniques.

5.3 Task Based Evaluation by Human

Similar to human in the loop evaluation (Nwogu et al., 2011), a task based evaluation was performed to make qualitative evaluation of the generated descriptions. Given a machine generated description, human subjects were instructed to find a corresponding video stream out of 10 candidate videos having the same theme (*e.g.*, a description of a Closeup against 10 Closeup videos). Once a choice was made, each subject was provided with the correct video stream and a questionnaire. The first question was how well the description explained the actual video, rating from 'explained completely', 'satisfactorily', 'fairly', 'poorly', or 'does not explain'. The second question was concerned with the ranking of usefulness for including various visual contents (*e.g.*, human, objects, their moves, their relations, background) in the description.

Seven human subjects conducted this evaluation searching a corresponding video for each of ten machine generated descriptions. They did not involve creation of the dataset, hence they saw these videos for the first time. On average, they were able to identify correct videos for 53%[6] of descriptions. They rated 68%, 48%, and 40% of descriptions explained the actual video 'fairly', 'satisfactorily', and 'completely'. Because multiple videos might have very similar text descriptions, it was worth testing meaningfulness of descriptions for choosing the corresponding video. Finally, usefulness of visual contents had mix results. For about 84% of descriptions, subjects were able to identify videos based on information related to humans, their actions, emotions and interactions with other objects.

6 Conclusion

This paper explored the bottom up approach to describing video contents in natural language. The 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 works include detection of groups, extension of behavioural models, more complex interactions among humans and other objects.

Acknowledgements

Muhammad Usman Ghani Khan thanks University of Engineering & Technology, Lahore, Pakistan for funding his work under the Faculty Development Program.

[6] It is interesting to note the correct identification rate went up to 70% for three subjects who also conducted creation of the dataset.

References

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

R.M. Bolle, B.L. Yeo, and M.M. Yeung. 1998. Video query: Research directions. *IBM Journal of Research and Development*, 42(2):233–252.

H.S. Chen, H.T. Chen, Y.W. Chen, and S.Y. Lee. 2006. Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 171–178. ACM.

M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

C. Galleguillos and S. Belongie. 2010. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722.

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

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

A. Kojima, T. Tamura, and K. Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184.

P. Kuchi, P. Gabbur, P. SUBBANNA BHAT, et al. 2002. Human face detection and tracking using skin color modeling and connected component operators. *IETE journal of research*, 48(3-4):289–293.

M.W. Lee, A. Hakeem, N. Haering, and S.C. Zhu. 2008. Save: A framework for semantic annotation of visual events. In *Computer Vision and Pattern Recognition Workshops. CVPRW'08*, pages 1–8. IEEE.

C.Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *WAS*.

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

H.H. Nagel. 2004. Steps toward a cognitive vision system. *AI Magazine*, 25(2):31.

I. Nwogu, Y. Zhou, and C. Brown. 2011. Disco: Describing images using scene contexts and objects. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

P. Over, W. Kraaij, and A.F. Smeaton. 2007. Trecvid 2007: an introduction. In *TREC Video retrieval evaluation online proceedings*.

M.S. Ryoo and J.K. Aggarwal. 2009. Semantic representation and recognition of continued and recursive human activities. *International journal of computer vision*, 82(1):1–24.

J.R.J. Schirra, G. Bosch, CK Sung, and G. Zimmermann. 1987. From image sequences to natural language: a first step toward automatic perception and description of motions. *Applied Artificial Intelligence an International Journal*, 1(4):287–305.

A. Sundaresan, A. RoyChowdhury, and R. Chellappa. 2003. A hidden markov model based framework for recognition of humans from gait sequences. In *International Conference on Image Processing, ICIP 2003*, volume 2. IEEE.

A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. 2008. Context-based vision system for place and object recognition. In *Ninth IEEE International Conference on Computer Vision*, pages 273–280. IEEE.

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

Y. Yang, C.L. Teo, H. Daumé III, C. Fermüller, and Y. Aloimonos. 2011. Corpus-guided sentence geration of natural images. In *EMNLP*.

B.Z. Yao, X. Yang, L. Lin, M.W. Lee, and S.C. Zhu. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508.

W. Zheng, H. Tang, Z. Lin, and T. Huang. 2010. Emotion recognition from arbitrary view facial images. *Computer Vision–ECCV 2010*, pages 490–503.