Information Sciences 303 (2015) 61–82

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A framework for creating natural language descriptions of video streams

Muhammad Usman Ghani Khan [a,*,1], Nouf Al Harbi [b], Yoshihiko Gotoh [b]

[a] Department of Computer Science, University of Engineering & Technology, Lahore, Pakistan
[b] Department of Computer Science, University of Sheffield, United Kingdom

ARTICLE INFO

Article history:
Received 21 June 2013
Received in revised form 8 December 2014
Accepted 11 December 2014
Available online 12 January 2015

Keywords:
Video retrieval
Video annotation
Natural language generation

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.

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 [28] but only a small number of works exist towards automatic description of visual scenes. Most studies in video retrieval have been based on keywords [3]. 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.

* Corresponding author.
E-mail addresses: usmanghanikhan@gmail.com (M.U.G. Khan), nmalharbi1@sheffield.ac.uk (N. Al Harbi), y.gotoh@dcs.shef.ac.uk (Y. Gotoh).
[1] The work was conducted while the first author was in the University of Sheffield.

http://dx.doi.org/10.1016/j.ins.2014.12.034
0020-0255/

1.1. 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. 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[2] data [21], 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.

1.2. 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 [1]. Lee et al. introduced a framework for semantic annotation of visual events in three steps; image parsing, event inference and language generation [15]. 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 [31]; 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 [30]. 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 [13]. 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 [19]. 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 and Siskind generated sentences for video sequences which were comprised of nouns, verbs, prepositions, adjectives and adverbs [32]. 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. [6] and by Barbu et al. [2], 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 [21]. It consisted of 140 segments of videos; each segment contained a single camera shot, spanning between 10 and 30 s 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'.

[2] trecvid.nist.gov.