Natural Language Descriptions for Video Streams

Muhammad Usman Ghani Khan† Nouf Al Harbi‡ Yoshihiko Gotoh‡
† Al-Khawarzimi Institute of Computer Science, UET Lahore, Pakistan
‡ Department of Computer Science, University of Sheffield, United Kingdom
(emails: usmanghanikhan@gmail.com, n.alharbi@dcs.shef.ac.uk, y.gotoh@dcs.shef.ac.uk)

Overview

Digital images and videos collection has increased exponentially in the recent years as more and more data is available in the form of personal photo albums, handheld camera videos, feature films and multilingual broadcast news videos, presenting visual data ranging from unstructured to highly structured. Today video data accounts for 80 percent of all network traffic. There is a need for qualitative filtering to find relevant information according to user requirements. Such a distillation process requires comprehensive information processing including categorisation and summarisation of multimedia resources. One approach to addressing this issue is to convert them into a more accessible form such as human language.

Most previous studies were related to semantic indexing of video using keywords [Naphade and Huang, 2001, Chang et al., 2007]. However it is often difficult with keywords alone to represent relations between various events present in video. Recent TREC video evaluation series have chosen the instance search based on visual contents as a pilot task, however works are still at a laboratory level [Over et al., 2011]. In practice meta data description is fundamental for the conventional video retrieval tasks. The problem is that, given the sheer volume of multimedia produced publicly and privately, only a small fraction is annotated with meta data, and the large majority cannot be retrieved using natural language queries.

An interesting extension to a keyword based scheme is natural language textual description in a syntactically and semantically correct formulation. They can clarify context between keywords by capturing their relations. Our recent work has addressed generation of textual descriptions for human actions, behaviour and their relations with other objects observed in video streams [Khan et al., 2011a,b, 2012]. Visual high level features (HLFs) extracted from video may be ‘keywords’, such as a particular object and its position/moves, used for the semantic indexing task in video retrieval. They are converted into natural language descriptions using a context free grammar (CFG). This formulation leads to a variety of useful applications. Video search engines can be implemented that make use of longer and more semantic queries. Natural language based video summarisation may play a significant role in reducing the storage space without losing too much information of the contents and helps in efficient and quick video retrieval according to user needs.

Several studies have emerged in the past few years which focused on mapping visual information to natural language. Kulkarni et al. [2011] introduced a template based approach to generating a natural language description for images. This was achieved by detecting objects, modifiers and spatial relationships using image processing techniques. The outputs were filtered using statistics obtained by parsing a large amount of text data, and finally a sentence was generated by an n-gram model or a template based approach. Mitchell et al. [2012] proposed a novel system to generate a humanlike description of images. By utilising a syntactic model of co-occurrence statistics, reliable syntactic trees were created by removing erroneous detection of visual contents. A study by Tan et al. [2011] proposed generation of natural descriptions for complex video contents using audio-visual concepts classifiers. For three scenarios (assembling a shelter, baseball batting and making a cake), visual and audio semantic concepts were identified and converted to a textual description using a template based approach.
Summary of Our Recent and Current Work

Our recent work was concerned with generation of natural language descriptions for humans observed in video streams. The work started with the implementation of conventional image processing techniques to extract HLFs from individual video frames. Although feature extraction processes were erroneous at various levels, approaches were explored to put them together for producing coherent descriptions. The following topics have been addressed:

Corpus generation and analysis. [Khan et al., 2012]
Video clips were classified into several categories and hand annotated with key phrases, a short title, and a full description. The resource was used for evaluation at the later stage. Its comprehensive analysis also presented insights into human interest, helping to develop strategies for the automatic scheme.

Framework for natural language description of individual video frames. [Khan et al., 2011a]
Identified visual HLFs were converted to natural language descriptions using CFG based templates. The framework put human actions, behaviour and their relations with other objects into context.

Scalability study. [Khan and Gotoh, 2012]
Approaches were explored to accommodate potentially erroneous and missing HLFs. The same approach could deal with videos from various genres for which HLF processing tools might not be available.

Framework for coherent natural language description of a video stream. [Khan et al., 2011b]
Descriptions of individual frames were crude, repeated and in some cases missing useful information due to sparseness of HLFs. This framework established approaches to creating a coherent and compact natural language description from a sequence of multiple video frames.

Currently we are working along the line of the following research questions:

Application of NLP techniques for image processing. Using natural language processing (NLP) can we recover missing or potentially erroneous features generated by the current image processing techniques? Can textual context improve identification of visual context?

Identification of spatial and temporal information across multiple HLFs. Current image processing techniques are able to identify a number of HLFs (typically objects and their moves) based on visual contents, however their spatial and temporal relations are often implicit. Can we describe such relations in addition to occurrence of HLFs?

Statistical modelling for description generation. The majority of recent studies relied on template based techniques for natural language generation. Can we develop a statistical model for description of visual contents? With the increasing amount of multimedia data available, can we create a data set for that purpose?

References

In EACL workshop, Avignon, 2012.
In ACM Multimedia, Scottsdale, Arizona, 2011.