TUZ at TRECVID 2015: Video Hyperlinking Task

Ersin Esen, Savaş Özkan and İlkay Atıl
TUBITAK UZAY Image Processing Group
METU Balgat 06531
Ankara, Turkey
Email: ersin.esen@tubitak.gov.tr

***Abstract*—In this paper, we present our video hyperlinking systems for the TRECVID 2015 Video Hyperlinking Task [1]. We used the provided BBC Dataset video keyframes and subtitles to develop two different systems and submit two separate runs. Our first run (tv15lnk_TUZ_L_1_F_M_M_MERGE1) uses subtitles to discover possible semantic links between video segments. Our second run (tv15lnk_TUZ_L_1_F_M_M_MERGE2) follows a different approach and uses only visual sentences extracted from keyframes to discover visual links between video segments. When we compare our two different approaches w.r.t. their MAP scores, subtitle based linking performs better. This is probably due to the fact that speech text contains more robust semantic data then visual sentences. We were planning to use both the first and the second systems to get a better result but our third system was not ready in time for submission. Overall results of the TRECVID 2015 Video Hyperlinking Task shows that our subtitle based first system is placed at the third quartile (%50-%75) among all participants whereas our visual sentence based second system is placed at the fourth quartile /%75-%100). We are planning to further our work by merging our two systems and discovering more interlinks between BBC Dataset videos.**

I. Introduction

In the age of big data, we capture video faster than we can analyze it. Each minute, hours of video content are uploaded to video sharing websites for sharing. As the data size grows, it gets harder to search for a specific content we are interested in. User provided video tags are only useful for labeling an entire video. Information about the sections of a video and traversing between related videos is still an unsolved problem.

Competitions like TRECVID 2015 Video Hyperlinking task [1] aims to direct researcher groups who work on video evaluation research field to develop a system which can create links between semantically related sections of videos in a vast video database. It basically aims to obtain a search engine for video segments much like a text-based search engine for web pages. Benefits of such a system will be a much better content discovery system than the current text-based query systems.

II. Video Hyperlinking System

We developed two different systems for the video hyperlinking task. The first system uses video subtitles to create a link between different video segments while the second system uses video keyframes to extract visual sentences to link different video segments.

A. Subtitle Based Hyperlinking (RUN1)

The idea of subtitle based hyperlinking is to use the semantic information contained in dialogues, narratives etc. about the scene or a certain topic. It also provides a fast way to process a scene since text data is easier to process compared to visual data due to its low data size (the entire subtitle corpus is 1.2GB where keyframes takes 304GB of space).

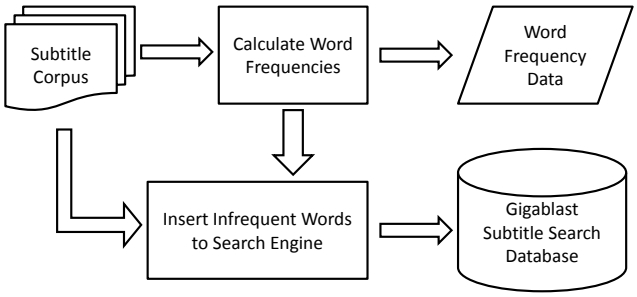


Fig. 1: Creation of a Subtitle Search Database

1) Creating Subtitle Search Database: Creation steps of our subtitle search database are shown at Figure 1. We first gather all subtitles into a subtitle corpus and calculate the word frequencies. This is necessary to differentiate between common grammatical words and useful words for hyperlinking. There are 186,782 unique words totalling to 12,918,000 occurances. We disregard words which has a greater frequency than 10,000 as common words. After we filter common words, we insert all subtitle text into a text search engine [2] to be able to perform fast searches. The text search engine keeps the subtitles and their corresponding record labels. When a subtitle is returned as a result, we use this record label to fetch the exact start and end times of a video segment which the retrieved subtitle belongs to.

2) Searching the Subtitle Database: The steps of our subtitle based hyperlinking search are shown at Figure 2. For a given query segment, we first filter subtitle words with a word frequency threshold. Then for each remaining word, we search the subtitle database using our text search engine. All search results are then gathered and sorted with respect to their relevance scores. We also fetch the video start and end times of video segments which belong to the retrieved subtitle results. We return the top N sorted results (1000 for our task) to the user as the hyperlinked video segments.

B. Visual Sentence Based Hyperlinking (RUN2)

The idea of visual sentence based hyperlinking is to use the visual information present at each keyframe to find the links between different video segments. Differently than subtitle based hyperlinking which can link semantically related video

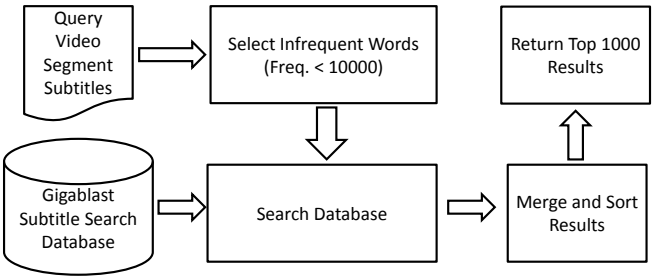


Fig. 2: Searching on Subtitle Database

segments, visual sentence based hyperlinking requires a visual similarity between two video segments in order to link them.

1) Creating Visual Sentence Search Database: The steps of creation of our visual sentence search database are shown at Figure 4. We developed a *visual sentence* approach which are generated by concatenating visual words from a 100K codebook. The 100K codebook is generated by applying k-means to Root-SIFT [3] features that are computed on the Flickr 100K image dataset [4][1]. Visual words (codebook indexes) are concatenated to form a visual sentence based on their proximity on a keyframe. A visual sentence has three components; the first visual word, the second visual word and the angle between the two words quantized to 8. Figure 3 shows an example visual sentence creation and match-up between query visual sentence and a visual sentence from the database.

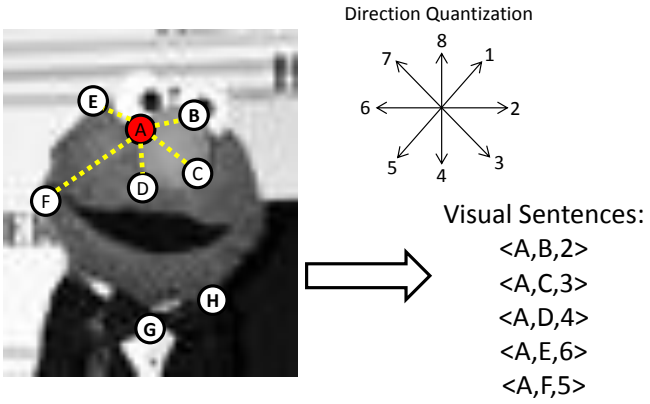
Given the BBC dataset, we extract RootSIFT features by using a single frame for each second of video. For each RootSIFT feature found in a frame, we generate five visual sentences by using five nearest neighbours. All of the visual sentences are then inserted into the gigablast search engine [2] to be searched afterwards.

2) Searching the Visual Sentence Database: The steps of our visual sentence based hyperlinking search are shown at Figure 5. When a query video segment is given, we extract RootSIFT from all frames of the query video segment and generate visual sentences. The newly obtained visual sentences are then used to query on the visual sentence search engine. Obtained results are then gathered and sorted according to their relevance scores given by the gigablast search engine and then top N results (1000 for our task) are returned to the user as the hyperlinked video segments.

III. Discussion

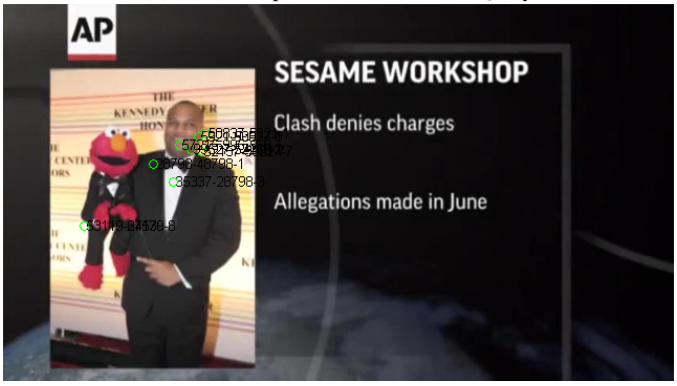
In this paper, we presented our two different video hyperlinking systems, subtitle based and visual sentence based, developed for the TRECVID 2015 Video Hyperlinking task. Overall results of the task shows that the subtitle based approach performs better than the visual sentence based approach. Our subtitle based system run is placed at the third quartile (%50-%75) among all participants whereas our visual sentence based system run is placed at the last quartile (%75-%100). This is mainly because subtitles contain more robust

[1]http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/flickr100k.html



(a) Creation of a Visual Sentence

(b) An Example Visual Sentence Query



(c) An Example Visual Sentence Match From Database

Fig. 3: Creation of a Visual Sentence and an Example Match-Up

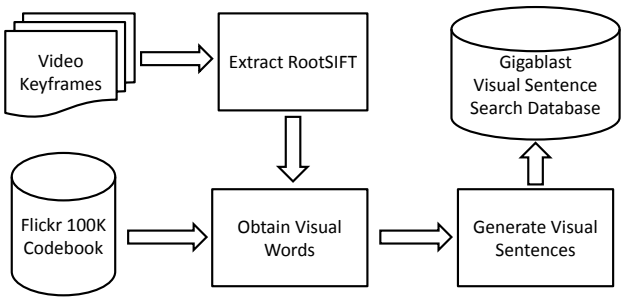


Fig. 4: Creation of our Visual Sentence Search Database

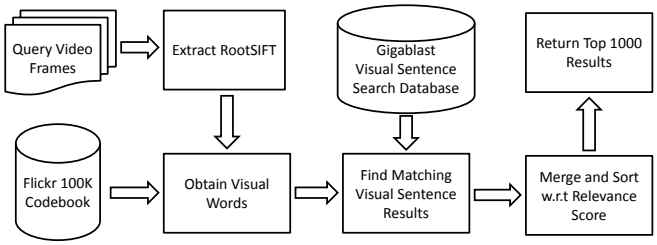


Fig. 5: Searching our Visual Sentence Database

semantic information about a video segment, exceeding the visual sentences semantic generalization ability. We originally developed the visual sentences for video copy detection on a large database and observed good performance, therefore we decided to use it for the video hyperlinking task. Apparently, visual sentences are not usable in their current form. We still think that they can be used to capture the semantic essense of a video segment with some changes. We used our two systems separately but the original idea was to use them together. However, we were not able to submit our third run due to time constraints. In the future, we are planning to use both subtitles and visual sentences to obtain good performance for the video hyperlinking task.

ACKNOWLEDGMENT

The authors would like to thank to the Scientific and Technological Research Council of TURKEY (TUBITAK), Space Technologies Research Institute for their support to our team's TRECVID 2015 participation.

REFERENCES

[1] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quenot, and R. Ordelman, "Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[2] M. Wells. (2000) Gigablast open source search engine. [Online]. Available: http://www.gigablast.com

[3] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2911–2918.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.