Linking visual and textual data on video

Pinar Duygulu, Howard Wactlar, Norman Papernick, Dorbin Ng

Informedia Project,
School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA.

1. Introduction

The Informedia Digital Video Library Project at Carnegie Mellon University [1] combines speech, image and natural language understanding to automatically transcribe, segment and index video for intelligent search and image retrieval. Since 1995, thousands hours of video (over two terabytes of data) have been collected, with automatically generated metadata and indices for retrieving videos from the library. The distinguishing feature of the project is the integration of speech, language and image understanding technologies for efficient creation and exploration of the library.

In video, the visual and textual information occur together. Integration of visual features (such as color, texture, shape, and motion features extracted from the visual data) with the semantics of the content (which comes from the text) is a challenging goal.

Although, visual features and text can be insufficient alone, they are complementary for each other [2]. Integrating these two types of data improves the performance for several applications [2,3,4,5,6]. Some methods are proposed to link images with text using the annotated image collections [5,6,7,8]. In this study, similar approaches will be applied to video to link the visual features with text.

In Section 2, our previous study that links image regions to words is described. Section 3 discusses the problems in video data that require finding the association of frames and text. The experimental results are presented in Section 4. Section 5 summarizes the results and discusses possible improvements to the model as a means of future directions.

2. Translating images to words

There are a wide variety of annotated image collections (e.g. Corel data set, most of the museum collections, the news on the web, etc.). In these collections, although the annotation words are associated with the image, the correspondence between the words and the image regions are unknown. The problem of finding such correspondences can be considered as the translation of image regions to words, similar to the translation of text from one language to another [4,6]. In that sense, there is an analogy between learning a *lexicon* for machine translation and learning a correspondence model for associating words with image regions. Learning a lexicon from data is a standard problem in machine translation literature [9,10]. Typically, lexicons are learned from a type of data set known as an *aligned bitext*. Assuming an unknown one-to-one correspondence between words, coming up with a joint probability distribution linking words in two languages is a missing data problem [9] and can be dealt by application of the *EM algorithm* [11].

Data sets consisting of annotated images are similar to aligned bitexts. There is a set of images, consisting of a number of regions and a set of associated words. We use k-means to vector-quantize the set of features representing an image region. Each region then gets a single label (blob token). The problem is then to construct a probability table that links the blob tokens with word tokens. This is solved using EM which iterates between the two steps: (i) use an estimate of the probability table to predict correspondences; (ii) then use the correspondences to refine the estimate of the probability table.

Once learned, the correspondences are used to predict words corresponding to particular image regions (*region naming*), or words associated with whole images (*auto-annotation*) [5]. Region naming is a model of object recognition, and auto-annotation may help to organize and access large collections of images.

3. Associating video frames with text

Similar correspondence problems occur in video. There are sets of frames and transcripts extracted from speech, but the correspondences between them are not clear. In most of the news examples, the anchor talks about an event or person, but the images related with the event or person occur later. Similarly, during a story the text and the associated video frames may not match directly. Some examples are shown in Figure 1.

If there is no direct association between text and video frames, a query based on text may produce incorrect results. The goal is now to find the correspondences between the video frames and associated text. When the correspondences are found, it can be used for annotating the video frames with more reliable words.

4. Experimental results

In the experiments, TREC 2001 data is used, because it is a truthed and extensively analyzed data set. For each shot a single image (keyframe) is chosen. Different than a still image with some annotated keywords, in video, text is not associated with a single frame. Therefore, the text for the surrounding frames is also taken as the associated words for a frame. For the experiments, the window size is arbitrarily taken as 5 (i.e. for each frame, text associated with the 5 preceding and 5 following frames are taken together with the text for that frame). Each image is divided into 7 x 7 grids, causing 49 segments for each image. Some simple color, texture, shape and position features are extracted from each segment. Blob tokens are obtained by applying k-means. The probability table is initialized to the co-occurrences of blobs and words. Then, EM algorithm is applied to construct the final translation probability table. To annotate images, the word posterior probabilities for each blob in the image are summed into a single word posterior. Then, the first ten words with the highest probability are chosen as the predicted words for each image. In these experiments all the words are taken as the vocabulary words. Therefore, most of the words are noise. In Figure 2, we choose some of the words from the top ten words that are used to annotate the images. As it can be seen, the predicted words are the correct words that describe what is in the image. Figure 3 shows the query results for “statue of liberty” using the current Informedia system. In some cases, neither “statue” nor “liberty” matches with the frames where the statue of liberty appears, and in some of the frames where the statue doesn’t appear, the text mentions the name. With the proposed approach, all the frames with the statue of liberty are annotated with “statue” and “liberty” in the top 3 words. Also, for the frames that do not include the statue, neither “statue” nor “liberty” is predicted. Therefore, when a query is performed on “statue of liberty”, with the proposed approach the results will give only the correct matches.

5. Discussion and future work

In this study, integration of visual and textual data is proposed to solve the correspondence problem between video frames and associated text. The preliminary results show that using this approach it is possible to have better annotations for the video frames that can be used to improve the performance of the text based queries.

Currently the system used grids as the blobs. Using a segmenter will give better results. Instead of current simple features, a better set of features, that also include some detectors (e.g. face detector) and some motion information is likely to improve the performance. Since all the words are taken as the vocabulary words, the textual data is noisy. Some lexical analysis needs to be done to understand which type of words more precisely correspond to visual features and should be taken as the vocabulary words. Furthermore, instead of taking one image from a shot (or a segment), all the frames can be taken to understand the association between the frames and text.

In the current system a relatively simple data set (TREC 2001) is used. The next step is to apply a similar approach to the news data where there are terabytes of video. Using a large set of data allows us to perform statistical analysis. It is an interesting problem to come with a distance metric for choosing the text that best describes the frame. The window size depends on the data. For example, it may be several frames for a documentary whereas a few frames for a news story that share the same text. Also, a weighting strategy can be applied, where the closer frames have higher weights while the farther ones have lower weights.

References

[1] The Informedia Digital Video Library Project, http://www.informedia.cs.cmu.edu

[2] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures", *Proc. International Conference on Computer Vision*, pp. II:408-415, 2001.

[3] K. Barnard, P. Duygulu, and D. Forsyth, "Clustering Art", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, pp. II:434-441, 2001.

[4] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary", *Proc. The Seventh European Conference on Computer Vision*, Copenhagen, Denmark, pp. IV:97-112, 2002.

[5] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching Words and Pictures", *Journal of Machine Learning Research*, vol. 3, pp. 1107-1135, 2003.

[6] P. Duygulu, "Translating Images to Words: A Novel Approach for Object Recognition", PhD Thesis, Middle East Technical University, Turkey, 2003.

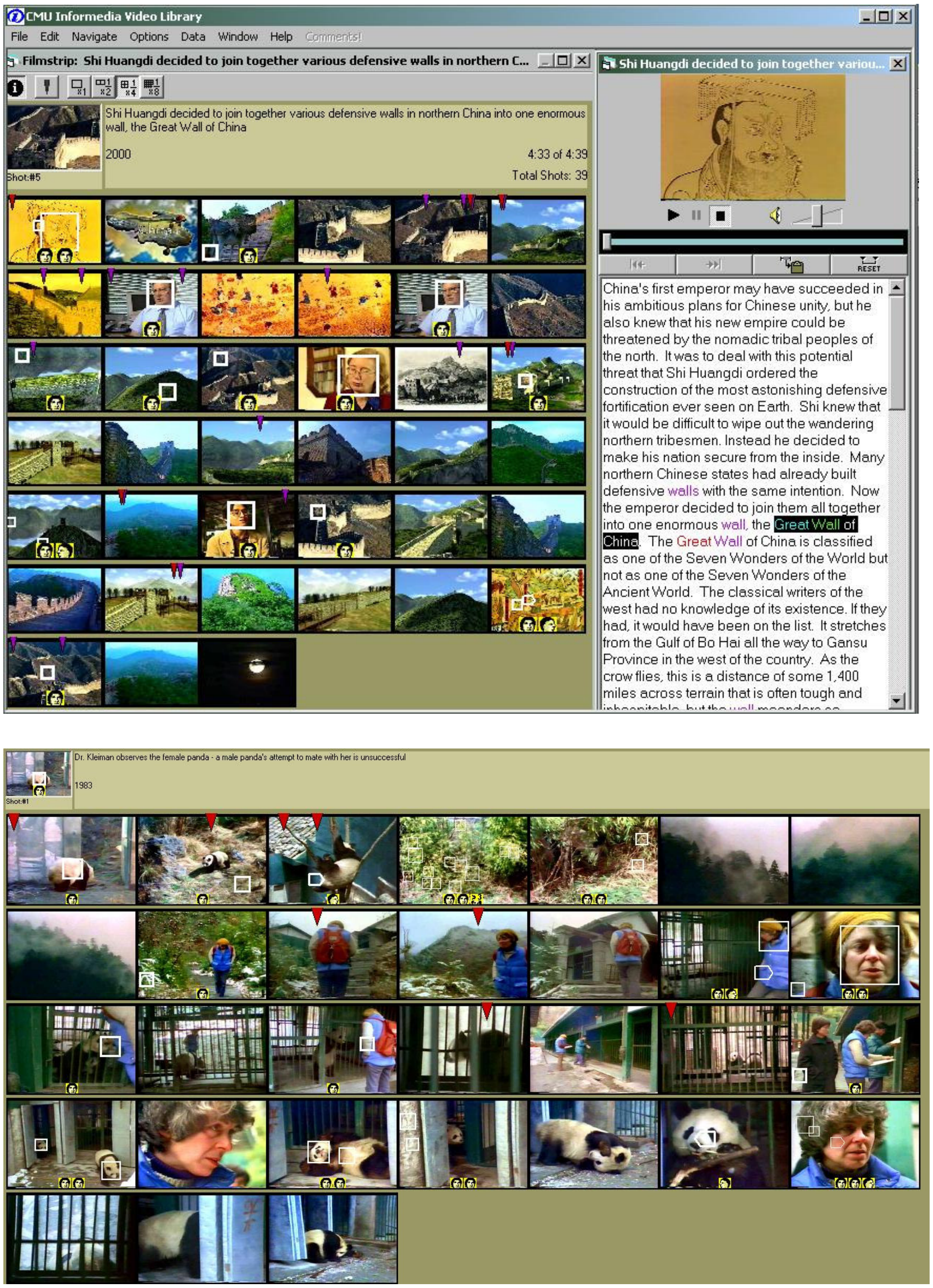
[7] O. Maron and A. L. Ratan, "Multiple-Instance Learning for Natural Scene Classification," *Proc. The Fifteenth International Conference on Machine Learning*, 1998.

[8] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," *Proc. First International Workshop on Multimedia Intelligent Storage and Retrieval Management (in conjunction with ACM Multimedia Conference 1999)*, Florida, 1999.

[9] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of machine translation: parameter estimation," *Computational Linguistics*, vol. 19, pp. 263-311, 1993.

[10] I.D. Melamed, , "Empirical Methods for Exploiting Parallel Texts", MIT Press, 2001.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1-38, 1977.


CMU Informedia Video Library
File Edit Navigate Options Data Window Help Comments!
Filmstrip: Shi Huangdi decided to join together various defensive walls in northern C...
Shi Huangdi decided to join together variou...
RESET

Shi Huangdi decided to join together various defensive walls in northern China into one enormous wall, the Great Wall of China

2000

4:33 of 4:39

Total Shots: 39

Shot:#5

China's first emperor may have succeeded in his ambitious plans for Chinese unity, but he also knew that his new empire could be threatened by the nomadic tribal peoples of the north. It was to deal with this potential threat that Shi Huangdi ordered the construction of the most astonishing defensive fortification ever seen on Earth. Shi knew that it would be difficult to wipe out the wandering northern tribesmen. Instead he decided to make his nation secure from the inside. Many northern Chinese states had already built defensive walls with the same intention. Now the emperor decided to join them all together into one enormous wall, the Great Wall of China. The Great Wall of China is classified as one of the Seven Wonders of the World but not as one of the Seven Wonders of the Ancient World. The classical writers of the west had no knowledge of its existence. If they had, it would have been on the list. It stretches from the Gulf of Bo Hai all the way to Gansu Province in the west of the country. As the crow flies, this is a distance of some 1,400 miles across terrain that is often tough and

Dr. Kleiman observes the female panda - a male panda's attempt to mate with her is unsuccessful

1983

Shot:#1

Figure 1: The results of Informedia system on some query words are shown. The red arrows represent the frames where the query word occurs in the text associated with the frame. There are correspondence problems between the frames and text. **Top:** query on the words “great wall”, **bottom :** query on the word “panda”: In some of the frames great wall / panda occurs both visually and in the text associating with the frame; however most of the nice great wall / panda images are missed, and frames that doesn’t include any wall / panda is matched.

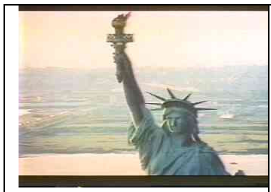
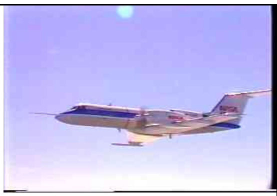
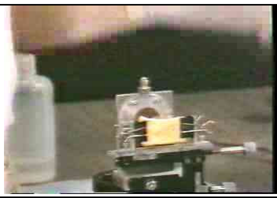
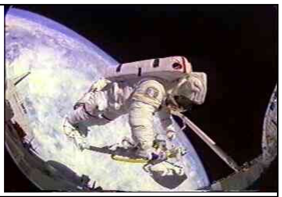
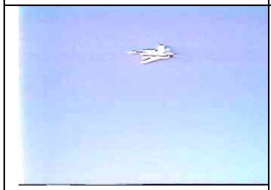
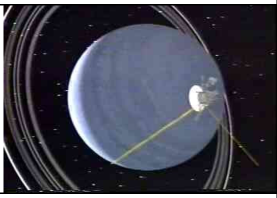
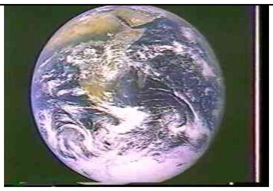
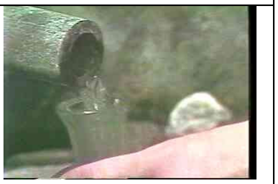
			
liberty(1) statue(2)	plane (2)	robot (5)	space(6) astronaut(7)
			
plane (2)	space(5)telescope(10)	space(1) world(6)	water(1) research(3)

Figure 2: The images are annotated with the top 10 words with the highest prediction probability given the regions of the image. For each image some of the annotation words and their rank is shown. For example, for the first image liberty was the highest probability word, and statue was the second. As the results show, the words that describe the images are predicted.

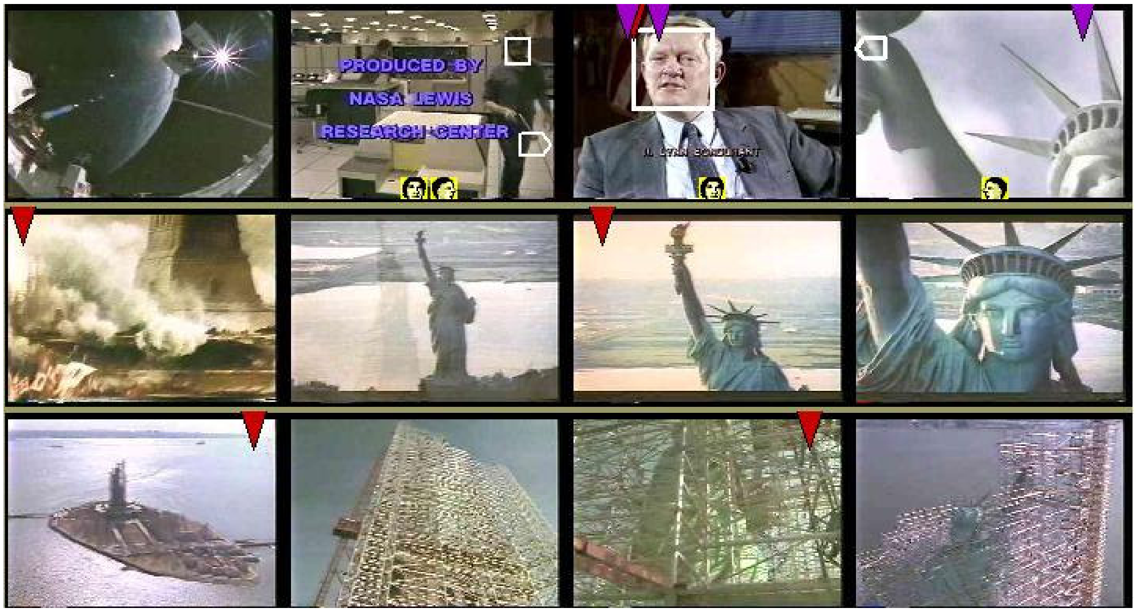


Figure 3: Query results for "statue of liberty" using the current Informedia system. Although, statue of liberty occurs in the second and fourth images on the second row, none of the words are matched. However, the third image on the first row, and the first image on the second row is matched, although they are not the images of statue of liberty. With the proposed approach, the results are corrected by predicting both "statue" and "liberty" in the top 3 words for all the images where the statue appears, and by not predicting any of the words for the others where the statue doesn't appear.