Associating video frames with text

Pınar Duygulu and Howard D. Wactlar
Informedia Project
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
pinar@cs.cmu.edu, wactlar@cmu.edu

ABSTRACT

In this study, integration of visual and textual data is proposed to solve the correspondence problem between video frames and associated text in order to annotate video frames with more reliable labels and descriptions. Visual features extracted from video frames are linked to text that is obtained from the audio transcripts using joint statistics. The results show that using this approach it is possible to have better annotations for video frames that can be later used to improve the performance of text based queries. The proposed approach will be integrated into the Informedia Digital Video Library Project at Carnegie Mellon University.

1. INTRODUCTION

Video is a rich source of data where visual and textual information occur together. A system that combines these two types of data is more powerful than a system addressing only one of them, since text and visual features may be ambiguous if treated in isolation. In order to make better use of the available data, visual features (such as color, texture, shape, and motion features) and semantics derived from text can be integrated. The text provides high-level semantics for a video sequence that cannot be obtained using current computer vision techniques, and visual properties supply rich information that can be difficult to express in a textual format.

When text and visual properties are integrated, several applications become possible including better search and browsing. There are many collections where images/videos are annotated with some descriptive text (e.g., Corel data set, some museum collections, news photographs on the web with captions, etc.). Manual annotation of these collections is subjective and requires a huge amount of effort [14]. Some approaches are proposed to make this process easier and automatic [2, 4, 5, 12, 13].

Figure 1: An example query result from the Informedia system [1], where the query word is `president`. A story is segmented into shots where each shot is represented with a keyframe. Transcripts that are extracted from the audio narrative are aligned with the shots. The arrows represent the shots where the query word occurs in the transcript aligned with the shot. There are correspondence problems between the frames and text. The word `president` is mentioned when the anchor and/or reporter were talking; but the president appears when his name is not mentioned. If a single frame is required as the result of the query, the system will produce the anchorperson/reporter frames which are not desirable in many cases. It is better to produce a frame where the president appears.

In these manually annotated image collections, although it is known that the annotation words are associated with the image, the correspondence between the words and the image regions are unknown. Some methods are proposed to find the correspondences by modeling the joint statistics of words and image regions [2, 10, 11, 15, 17].

Similar correspondence problems occur in video data. There are sets of video frames and transcripts extracted from the audio speech narrative, but the semantic correspondences between them are not fixed because they may not be co-occurring in time. If there is no direct association between

text and video frames, a query based on text may produce incorrect visual results.

For example, in most news videos (see Figure 1) the anchorperson talks about an event, place or person, but the images relating to the event, place, or person appear later in the video. Therefore, a query based only on text related to a person, place, or event, and showing the frames at the matching narrative, will yield incorrect frames of the anchorperson as the result.

The goal of this study is to determine the correspondences between the video frames and associated text in order to annotate the video frames with more reliable labels and descriptions. This enables a textual query to return more accurate semantically corresponding images, and enables an image-based query or response to provide more meaningful descriptors.

In Section 2, our approach to solve the correspondence problem between image regions and text will be explained. The strategy to apply a similar approach on video data will be described in Section 3. In Sections 4 and 5, the procedure and the experimental results will be presented for two different data sets : TREC 2001 and Chinese cultural data sets respectively. Section 6 will discuss the results and present future directions.

2. MULTIMEDIA TRANSLATION

The problem of finding the correspondences can be considered as the translation of visual features to words, similar to the translation of text from one language to another. 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 [7, 16]. Typically, lexicons are learned from a type of data set known as an aligned corpora. 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 [7] and can be dealt by application of the EM (Expectation Maximization) algorithm [9].

Data sets consisting of annotated images are similar to aligned corpora. There is a set of images, each consisting of a number of regions and a set of associated words. Each image is segmented into regions and from each region a set of features, including color, texture, shape, position and size, are extracted. In order to exploit the analogy with machine translation, we vector-quantize the set of features representing an image region using k-means. 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. The probability table is initialized to the co-occurrences of blobs and words. The final translation probability table is constructed using the EM algorithm 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 probability table is used to predict words corresponding to particular image regions (region naming), or words associated with whole images (auto-annotation) (see [10, 11] for the details).

3. CORRESPONDENCES ON VIDEO

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. The current library consists of terabytes of data (broadcast news captured over the last years, documentaries produced for public television and government agencies, classroom lectures, and other video genres) with automatically extracted metadata and indices.

Broadcast news is very challenging data, but due to its nature it is mostly based on people and requires a focus on person detection/recognition. In this study, we use subsets of Chinese culture and TREC 2001 data sets which are relatively simpler.

The data consists of video frames and the associated transcript extracted from the audio. The frames and transcripts are associated on the shot-basis. Each shot is represented with a single keyframe. Transcripts are aligned with the shots by determining when each word was spoken through an automatic alignment process using Sphinx-III speech recognizer.

Each keyframe is segmented into regions, and features are extracted from each region. In this study, fixed sized grids are used as the regions because of their simplicity to apply on a large volume of data. A feature vector of size 46 is formed to represent each region. Position is represented by the coordinates (x, y) of the region's center of gravity in relation to the image dimensions. Color is redundantly represented using the mean and variance of the HSV and RGB color spaces. Texture is represented by using the mean and variance of 16 filter responses. We used four difference of Gaussian filters with different sigmas and twelve oriented filters, aligned in 30 degree increments.

The vocabulary consists of only nouns which are extracted by applying Brill's tagger [6] to the transcript. Due to the errorful transcripts obtained from the speech recognition, many noisy words remain in the vocabulary.

4. TREC 2001 DATA

The TREC 2001 data is chosen, because it is a truthed and extensively analyzed data set. 2232 images are used from this data set. All the nouns (1938 words) are taken as the vocabulary words.

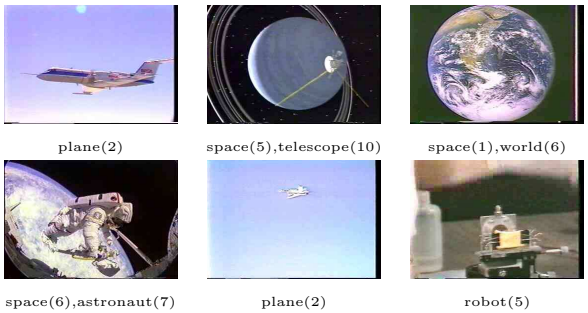


Figure 2: Example annotation results for TREC 2001 data. The images are annotated with the top 10 words with the highest prediction probability given the regions of the image. For each image some of the annotation words and their ranks are shown.

Different from a still image with some annotated keywords, in video, text is not usually associated with a single frame. The words for the surrounding frames can also be associated with the current frame. For the experiments on TREC data, the text for the surrounding frames is also considered by setting the window size arbitrarily to five (i.e., for each frame, text associated with the five preceding and five following frames are taken together with the text for that frame).

Each image is divided into 7 x 7 blocks, resulting in 49 regions for each image. Then, features are extracted from each of these blocks and feature space is vector quantized using k-means where k is set to 500. An initial translation probability table is constructed by taking the co-occurrences of 500 blob tokens and 1938 word tokens. The final translation probability table is obtained by applying EM. In order to annotate the images, the word posterior probabilities for the image blobs are summed into a single word posterior, and the highest probability words are taken as the annotation words.

In Figure 2, some of the annotation words predicted in the top ten words with the highest probability for an image are shown. As 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 doesnt 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 three words as shown in Figure 4. 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 where the Statue appears.

Figure 3: Query results for `statue of liberty` using the current Informedia system. Each shot is represented with a single keyframe and the transcript extracted from the audio is aligned with the shots. The arrows indicate the shots where the words `statue` or `liberty` occur in the corresponding audio/transcript. For example, the first and the second purple arrows indicate the word `statue` for the third and fourth images on the first row respectively, the first red arrow indicates the word `liberty` for the first image on the second row, and so on. Although, the Statue of Liberty appears 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 three words for all the images where the Statue appears, and by not predicting any of the words for the others where the statue doesnt appear.

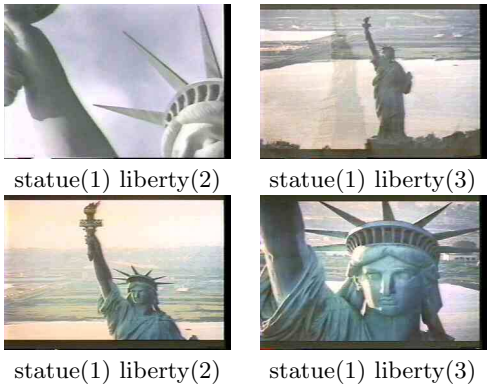


Figure 4: For the shots that Statue of Liberty appears, rank of the `statue` and `liberty` words in the predictions.

5. CHINESE CULTURE DATA

The other data set used for the experiments consists of Chinese cultural documentaries [8, 18]. Figure 5 and Figure 6 show some examples from this data set. Similar correspondence problems occur in this data set. For example, in the story about the Great Wall, some frames show the views of the Wall, while the others show the reporter or the other related people/objects; therefore the word `wall` may not match with the view of the Wall.

For the experiments, the movies that includes interviews with people are removed due to their dependence on face detection/recognition, and the remaining 3785 shots are used. Each shot is represented with a single keyframe which is divided into 5 x 5 blocks. Then, features are extracted from each of these blocks and feature space is vector quantized using k-means where k is set to 1000. Therefore there are 1000 blob tokens.

In order to construct the vocabulary, nouns extracted from Brill's tagger are used. Figure 7 shows that the frequency of the words lies in a large range. The number of words in the vocabulary is 2597, but only about 20% of the words are in a meaningful range. We eliminate the words that occur less than 5 times, or more than 250 times. This pruning process reduces the vocabulary to 626 words. Shots that are not associated with any word after this pruning process are also eliminated, resulting in 2785 remaining shots.

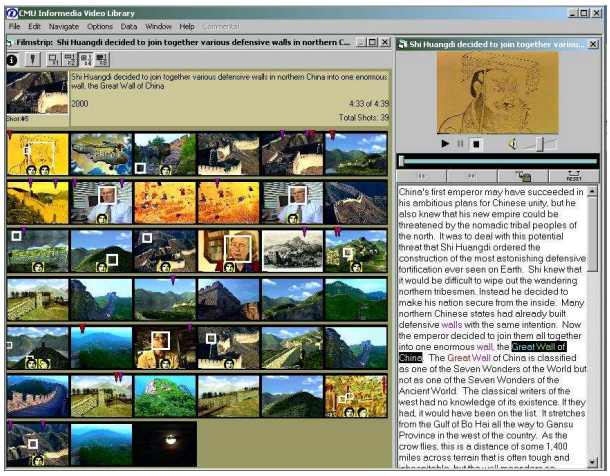


Figure 5: Results of a query for `great wall` on Chinese culture data set using the current Informedia system. The arrows indicate the shots where the query words occur in the corresponding audio.

For the experiments on the TREC data, the text for the surrounding shots is also taken by setting a fixed window size. In the Chinese data set, the window size is varied. First, we give the results using only the text aligned with a single shot, then we compare the results for different window sizes.

Figure 8 shows some frames that predict (a)`panda`, (b)`wall`, (c)`emperor` in the first three words with the highest probability. As can be seen, the right words are predicted for the frames that are related with the word. The word `panda` is

Figure 6: Results of a query for `panda` on Chinese culture data set using the current Informedia system. The arrows indicate the shots where the query word occurs in the corresponding audio.

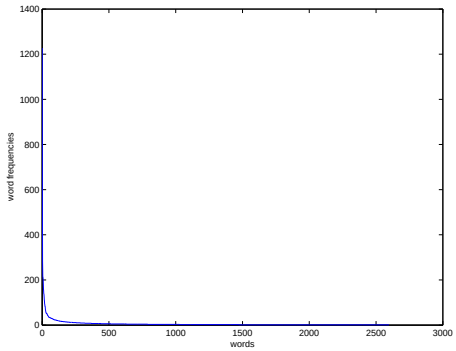
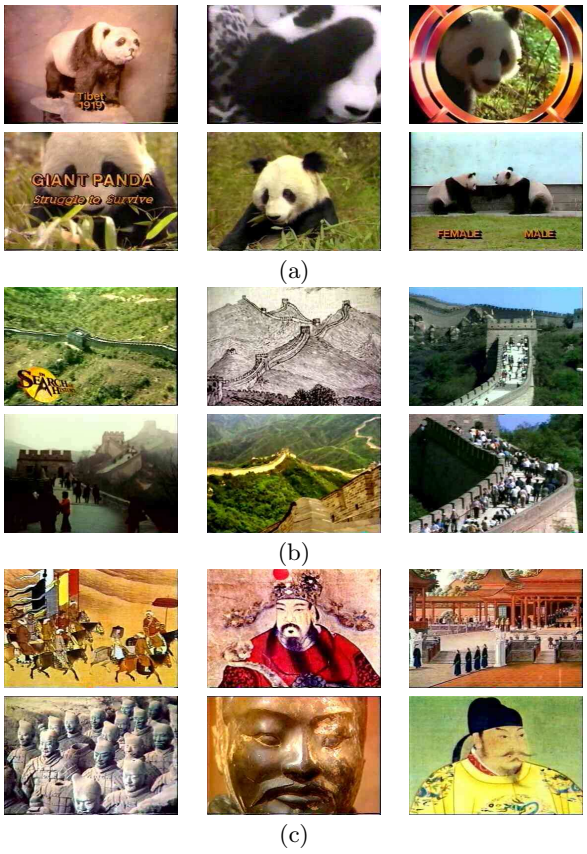


Figure 7: Word frequencies for the Chinese culture data set. Originally there are 2597 words in the vocabulary. However, almost 80% of the vocabulary occurs less then five times. We eliminate the words that occur less than 5 times or more than 250 times to have a new vocabulary. The number of words remained in the pruned vocabulary is 626.

correctly predicted for different views of panda. Similarly, the word `wall` is correctly predicted when the Great Wall appears in the image, although the images are not very similar. The images for `emperor` are in a larger range, since it is hard to associate this word with a specific object or scene.

In order to evaluate the results on a larger scale 189 images are visually investigated for the word `panda`. The results are shown in Table 1. 127 images are associated with the word `panda` (using the aligned transcript), but only in 42 of these images a panda appears. The system predicts `panda` as the first word with the highest probability for 121 images, and in 51 of them a panda appears in the image. In 25 of these images the word `panda` is predicted for the images where a panda appears, even though the word was not associated with the frame originally. The results show that we missed 11 images where the `panda` word was associated and a panda appears in the image, but cannot be predicted as the first word. For the images where the system cannot predict `panda` as the first word, it was usually the second or the third word.



(a) (b) (c)

Figure 8: Sample images that predict (a)panda, (b)wall, and (c)emperor in the first three words with the highest probability.

The predicted words can be used for two different purposes: to have better annotations for the frames on the training set and to annotate the frames on the test set that do not have associated text. As explained before, we eliminated 1000 images since the words associated with those images were the less frequent words. These images are used as the test set. For `panda`, we randomly choose two sets of frames where the story is about a panda. Among these, the frames that are in the test set are analyzed. We look at the images where panda is predicted as the first word, and check whether the panda appears in the image. In the first set 11 images over 25 images, and in the second set 2 images over 25 images, predict `panda` as the first word when the panda appears in the image.

The scene shown in Figure 6 is further used to analyze the results for predicting the word `panda` both for the training and test sets. Figure 9 shows the rank of the word `panda` as the predicted word for the corresponding frames. Red (dark gray) numbers correspond to the images that are in the test set (the frames that are initially eliminated), and green (light gray) numbers correspond to the images that are in the training set. The ranking in the training set is also important, since most of the correct associations are missing and our goal is finding more reliable annotations for the frames. As the figure shows, on most of the training images if a panda appears in the image, we predict the word

Table 1: Correspondence results for `panda`. 189 frames are inspected. *Associated* means that the word is in the transcript that is aligned with the frame. *Predicted* means that the system predict `panda` for the frame as the first word. *Correct* means a panda appears, and *incorrect* means a panda doesn't appear in the frame.

number of associations	127
number of correct associations	42
number of incorrect associations	85
number of predictions	121
number of correct predictions	51
number of incorrect predictions	70
associated but not predicted	64
association is correct but not predicted	11
association is incorrect not predicted	53
not associated but predicted	61
not associated but correctly predicted	25
not associated and incorrectly predicted	36
associated and predicted and correct	26
associated and predicted but incorrect	31

`panda` in the first four words with the highest probability. The predictions for the test images are also satisfactory: some of the images where a panda appears are annotated with the word `panda`, and for the images that don't show any panda the annotation rank is very low. However, both for the training and test images, we see a problem when the woman appears: since the woman frames highly co-occur with the word `panda`, the system learns that the woman frames are also associated with `panda` and we incorrectly predict `panda` with a high score.

Figure 9: The predictions for the panda scene. The numbers show the rank of the word `panda` among the predicted words. Red (dark gray) numbers correspond to the images that are not in the training set, green (light gray) numbers correspond to the images in the training set.

We investigate the effect of window size, by choosing either only the words from a single shot, or also from the surrounding shots (window size is set to 1,2 or 3). The average number of words associated with a single shot is 3.7275 when only the words that are associated with a single shot are used; 10.1670 when window size is set to 1; 16.4470 when window size is 2; and 22.6725 when window size is 3. The maximum number of words are 25, 47, 63 and 70 respectively.

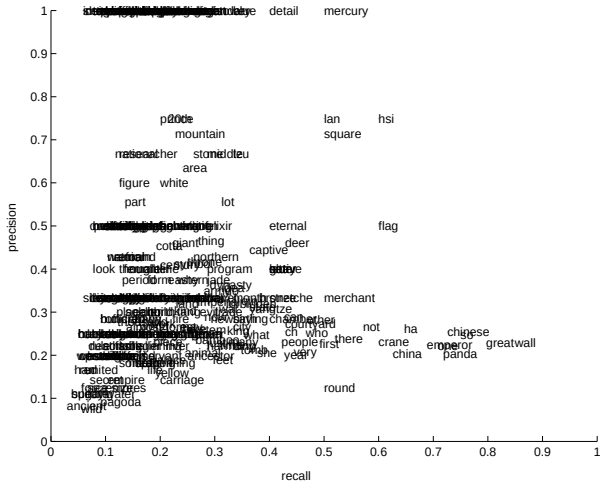


Figure 10: Recall vs precision graph : single frame. *Recall* is defined as the number of correct predictions over the number of times that the word occurs in the data, and *precision* is defined as the number of correct predictions over all predictions.

Figures 10, 11, 12 and 13 show the recall and precision values as a function of words when the associated words are chosen from a single frame, or from a window size 1, 2, and 3 respectively. **Recall** is defined as the number of correct predictions over the number of times that the word occurs in the data, and **precision** is defined as the number of correct predictions over all predictions. Correct predictions are found by comparing the first n words that are predicted with the highest probability (where n is the number of actual words associated with the frame), with the actual ones. It is hard to understand the figures clearly, but it is easy to see that the system predicts more words as the window size increases. Also, the figures indicate that for most of the words the recall and/or precision values increase as the number of words associated with a frame increases. Table 2 shows the recall and precision values for some selected words.

Table 2: Comparison of recall and precision values for some selected words as a function of window size.

	panda rec - prec	wall rec - prec	emperor rec - prec
single frame	0.718 - 0.204	0.849 - 0.228	0.689 - 0.224
wsize = 1	0.818 - 0.243	0.918 - 0.292	0.874 - 0.286
wsize = 2	0.886 - 0.256	0.952 - 0.315	0.945 - 0.326
wsize = 3	0.915 - 0.260	0.972 - 0.319	0.969 - 0.363

Table 3: Prediction measures when different window sizes are used to obtain the words associating with a shot.

	prediction measure
single frame	0.1851
wsize = 1	0.2469
wsize = 2	0.2783
wsize = 3	0.2975

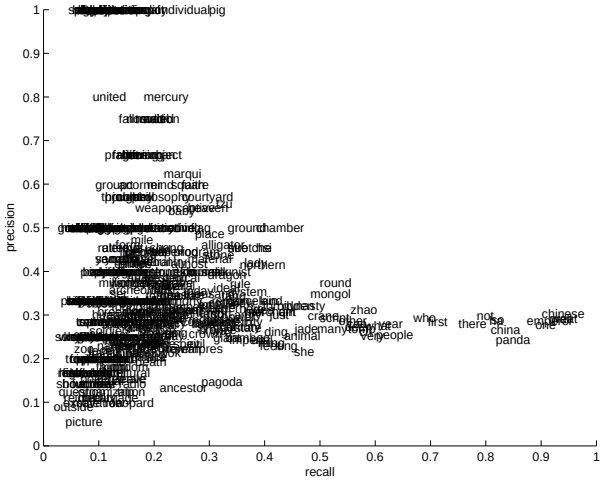


Figure 11: Recall vs precision : window size = 1

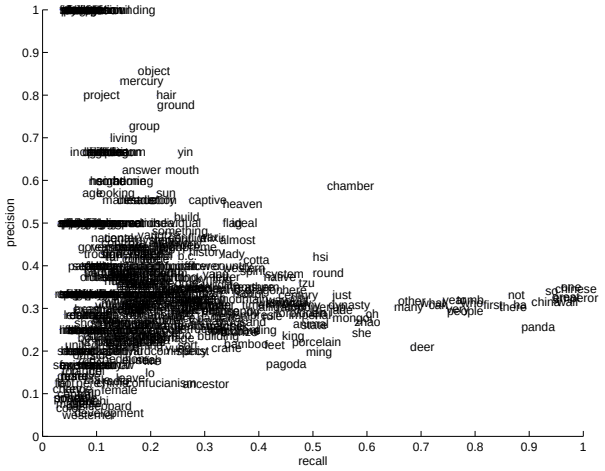


Figure 12: Recall vs precision : window size = 2

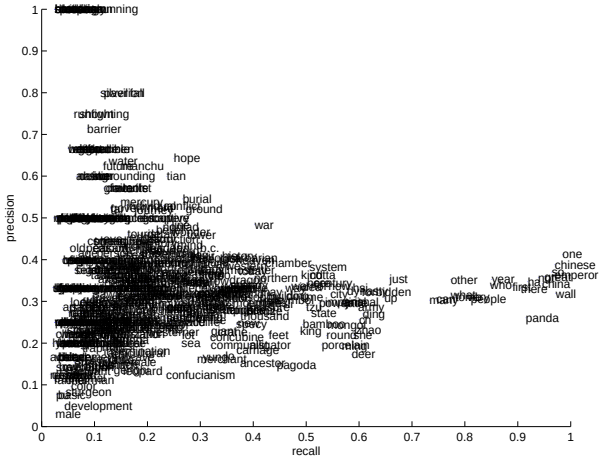


Figure 13: Recall vs precision : window size = 3

The results are also compared using the **prediction measure**, which is calculated by averaging the number of correct predictions over the number of words in an image for all the images. As Table 3 shows, prediction measure is better as the window size increases.

6. 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.

In the current system relatively simpler and smaller data sets are used. Our goal is to apply a similar approach to broadcast news which is a harder dataset since there are terabytes of video and it requires focusing on people.

Currently the system uses simple features extracted from the still images. A better set of features that also include some detectors (e.g., face detector) and some motion information is likely to improve the performance. In the experiments, fixed size grids are used. Using a segmenter, better results can be obtained [3]. Also, in video the objects are mostly the ones that moves. Therefore, it is better to use the temporal information to segment the moving objects.

Using a large number of data may allow 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. In this study, individual words are taken as the vocabulary words. Taking the noun phrases or the compound words (e.g. "statue of liberty") may improve the performance. Also, 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.

7. REFERENCES

[1] Informedia Project. http://informedia.cs.cmu.edu.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. A. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[4] K. Barnard and D. A. Forsyth. Exploiting image semantics for picture libraries. In *The First ACM/IEEE-CS Joint Conference on Digital Libraries*, page 469, 2001.

[5] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *Int. Conf. on Computer Vision*, pages 408–15, 2001.

[6] E. Brill. A simple rule-based part of speech tagger. In *In Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[7] P. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[8] C. Chen. First Emperor of China, Voyager CD-ROM. http://www.voyagerco.com/cdrom/, 1994.

[9] A. P. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 1(39):1–38, 1977.

[10] P. Duygulu, K. Barnard, N. Freitas, and D. A. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision (ECCV)*, volume 4, pages 97–112, 2002.

[11] P. Duygulu-Sahin. *Translating Images to words: A novel approach for object recognition.* PhD thesis, Middle East Technical University, Turkey, 2003.

[12] C. Jelmini and S. Marchand-Maillet. Deva: an extensible ontology-based annotation model for visual document collections. In *Proceedings of SPIE Photonics West, Electronic Imaging 2002, Internet Imaging IV, Santa Clara, CA, USA*, 2003.

[13] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):14, 2003.

[14] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1:259–285, 2000.

[15] O. Maron. *Learning from Ambiguity.* PhD thesis, MIT, 1998.

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

[17] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[18] H. Wactlar and C. Chen. Enhanced perspectives for historical and cultural documentaries using informedia technologies. *Joint Conference on Digital Libraries (JCDL'02), Portland, Oregon*, July 13-17, 1994.