Translating images to words for recognizing objects in large image and video collections

Pınar Duygulu and Muhammet Başţan

Department of Computer Engineering, Bilkent University, Ankara, Turkey
(duygulu,bastan)@cs.bilkent.edu.tr

Abstract. We present a new approach to the object recognition problem, motivated by the recent availability of large annotated image and video collections. This approach considers object recognition as the translation of visual elements to words, similar to the translation of text from one language to another. The visual elements represented in feature space are categorized into a finite set of blobs. The correspondences between the blobs and the words are learned, using a method adapted from Statistical Machine Translation. Once learned, these correspondences can be used to predict words corresponding to particular image regions (region naming), to predict words associated with the entire images (auto-annotation), or to associate the speech transcript text with the correct video frames (video alignment). We present our results on the Corel data set which consists of annotated images and on the TRECVID 2004 data set which consists of video frames associated with speech transcript text and manual annotations.

1 Introduction

Object recognition is one of the major problems in computer vision and there has been many effort to solve this problem (see [13] for a detailed review of recent approaches). However, recognition on the large scale is still a challenge. We consider the object recognition problem as translating the visual elements to semantic labels. This view of object recognition allows us to recognize large number of objects in the large image and video collections.

Classical object recognition systems require supervised data where regions corresponding to objects are manually labeled. However, creation of such data is labor intensive and error prone. Recently, many annotated image and video collections have become available. Examples include stock photographs annotated with keywords, museum image collections with metadata, captioned news photographs on the web, and news videos associated with captions or speech recognition transcripts (Fig.1). These annotated data sets, provide labels not on the region level but on the image level. Although, that is only loosely labeled data, it is available in large quantities. By making use of this data, the object recognition problem can be transformed into finding the correspondences between the image structes and annotation words.

Fig. 1. Examples of annotated images **Top:** Corel data set. **Bottom:** TRECVID news videos data set

Recent studies show that, with careful use of these large annotated data sets, it is possible to predict words for the images by integrating the visual and textual data [22, 30, 19, 24, 27]. More recently, probabilistic models are proposed to capture the joint statistics between images and words, including the hierarchical aspect model [5, 4], relevance based models [16, 18, 12], mixture of multi-modal latent Dirichlet allocation model [3], and a method based on Hidden Markov Model [15].

Predicting words for the images, which is referred as **auto-annotation**, is helpful since considerable amount of work for manually annotating the images can be eliminated. However, that is not a solution to the recognition problem, since the correspondences between image structures and words are unknown. For example, an image with the keyword `tiger` is likely to contain a tiger object, but we don't know which part of the image corresponds to tiger (Fig.2).

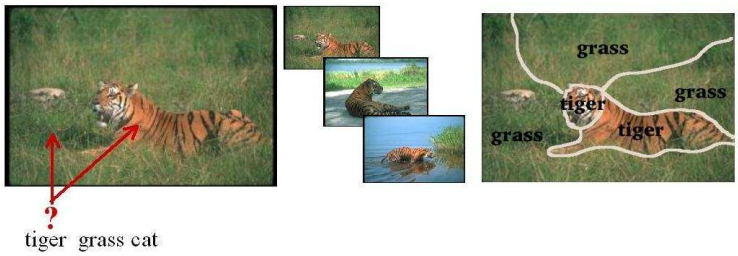


Fig. 2. The correspondence problem between image regions and words. The keywords `tiger`, `cat` and `grass` are associated with the image, but the word-to-region correspondences are unknown. If there are other images, the correct correspondences can be learned and used to automatically label each region in the image with correct words or to auto-annotate a given image

The correspondence problem is very similar to the correspondence problem faced in statistical machine translation literature (Fig.3). There is one form of data (image structures or English words) and we want to transform it into another form of data (keywords or French words). Learning a lexicon (a device that can predict one representation given the other representation) from large data sets (referred as aligned bitex) is a standard problem in the statistical machine translation literature [8, 23, 17, 21]. Aligned bitexts consist of many small blocks of text in both languages, corresponding to each other at paragraph or sentence level, but not at the word level. Using the aligned bitexts the problem of lexicon learning is transformed into the problem of finding the correspondences between words of different languages, which can then be tackled by machine learning methods.

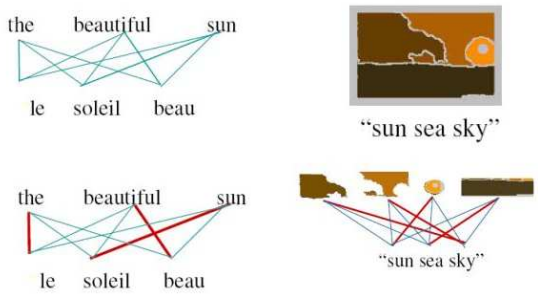


Fig. 3. The analogy with the statistical machine translation. We want to transform one form of data (image structures or English words) to another form of data (keywords or French words)

Due to the similarity of the problems, correspondence problem between image structures and keywords can be attacked as a problem of translating visual features into words, as first proposed in [10]. Given a set of training images, the problem is to create a probability table that associates words and visual elements. This translation table can then be used to find the corresponding words for the given test images (**auto-annotation**) or to label the image components with words as a novel approach to recognition (**region labeling**).

A similar correspondence problem occurs 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 Fig.4) 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.

american home twenty two anticipated ... (1) so today it was an energized president **CLINTON** who formally presented his one point seven three trillion dollar budget to the congress and told them there'd be money left over first of the white house a.b.c's sam donaldson (2) ready this (3) morning here at the whitehouse and why not (4) next year's projected budget deficit zero where they've presidental shelf and tell *this* (5) *budget marks the hand of an era and ended decades of deficits that have shackled our economy paralyzed our politics and held our people back* ..... (6) [empty] (7) [empty] (8) administration officials say this balanced budget are the results of the president's sound policies he's critics say it's merely a matter of benefiting from a strong economy that other forces are driving for the matter why it couldn't come at a better time just another upward push for mr **CLINTON**'s new sudden sky high job approval rating peter thanks very ...

Fig. 4. Keyframes and corresponding speech transcripts for a sample sequence of shots for a story related to Clinton. Italic text shows Clinton's speech, and capitalized letters show when Clinton's name appears in the transcript. Note that, Clinton's name is mentioned when an anchorperson or reporter is speaking, but not when he is in the picture

The goal is to determine the correspondences between the video frames and speech transcript text in order to associate the video frames with more reliable labels and descriptions, which we refer as **video alignment**. This enables a textual query to return more accurate semantically corresponding images. We will show that, a modified version of the translation model can be used to solve the correspondence problem faced in video data.

The other models proposed to attack the correspondence problem include the simple co-occurrence model [25], Correlation Latent Dirichlet Allocation (LDA) model [6] and an extension of translation approach using MRFs [9].

2 Translation Approach

Brown *et al.* [8] propose a set of models for statistical machine translation. These models aim to maximize the conditional probability density $p(\mathbf{f} \mid \mathbf{e})$, which is called as the likelihood of translation $(\mathbf{f}, \mathbf{e})$, where $\mathbf{f}$ is a set of French words, and $\mathbf{e}$ is a set of English words.

In machine translation, a lexicon links a set of discrete objects (words in one language) onto another set of discrete objects (words in the other language). In our case, the data consist of visual elements associated with words. The **words** are in the dicrete form. In order to exploit the analogy with machine translation, the visual data, represented as a set of feature vectors also need to be broken up into discrete items. For this purpose, the features are grouped by vector quantization techniques such as k-means and the labels of the classes, which we call as **blobs**, are used as the discrete items for the visual data. Then, an aligned

bitext, consisting of the blobs and the words for each image is obtained and used to construct a probability table linking blobs with words.

In our case, the goal is to maximize $p(\mathbf{w} \mid \mathbf{b})$, where $\mathbf{b}$ is a set of blobs and $\mathbf{w}$ is a set of words. Each word is aligned with the blobs in the image. The alignments (referred as $\mathbf{a}$) provide a correspondence between each word and all the blobs. The model requires the sum over all possible assignments for each pair of aligned sentences, so that $p(\mathbf{w} \mid \mathbf{b})$ can be written in terms of the conditional probability density $p(\mathbf{w}, \mathbf{a} \mid \mathbf{b})$ as

$$p(\mathbf{w} \mid \mathbf{b}) = \sum_{\mathbf{a}} p(\mathbf{w}, \mathbf{a} \mid \mathbf{b}) \tag{1}$$

The simplest model (Model-1), assumes that all connections for each French position are equally likely. This model is adapted to translate blobs into words, since there is no order relation among the blobs or words in the data [29]. In Model-1 it is assumed that each word is aligned exactly with a single blob. If the image has l blobs and m words, the alignment is determined by specifying the values of a_j such that if the j^{th} word is connected to the i^{th} blob, then $a_j = i$, and if it is not connected to any blob $a_j = 0$. Assuming a uniform alignment probability (each alignment is equally probable), given a blob the joint likelihood of a word and an alignment is then can be written as:

$$p(\mathbf{w}, \mathbf{a} \mid \mathbf{b}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(w_j \mid b_{a_j}) \tag{2}$$

where $t(w_j \mid b_{a_j})$ is the translation probability of the word w_j given the blob b_{a_j}, and ϵ is a fixed small number.

The alignment is determined by specifying the values of a_j for j from 1 to m each of which can take a value from 0 to l. Then, $p(\mathbf{w} \mid \mathbf{b})$ can be written as:

$$p(\mathbf{w} \mid \mathbf{b}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(w_j \mid b_{a_j}) \tag{3}$$

Our goal is to maximize $p(\mathbf{w} \mid \mathbf{b})$ subject to the constraint that for each b

$$\sum_{w} t(w \mid b) = 1 \tag{4}$$

This maximization problem can be solved with the EM (Expectation Maximization) formulation [8, 10]. In this study, we use the Giza++ tool [1, 26] -which is a part of the Statistical Machine Translation toolkit developed during summer 1999 at CLSP at Johns Hopkins University- to learn the probabilities. Note that, we use the direct translation model throughtout the study.

The learned association probabilities are kept in a translation probability table, and then used to predict words for the test data.

3 Associating visual elements with words

In this study, we attack two types of correspondence problems between visual elements and words. The first problem is between the image regions and words in annoatated image collections. The second problem is between the frames of a video sequence and the corresponding speech transcript text.

In the annotated image and video collections, the images are usually annotated with a few keywords which describe the images. However, correspondences between image regions and words are unknown. In order to solve this correspondence problem, first we segment the images into regions and represent each region with a set of visual features. A vector quantization technique, such as k-means, is used to transform the visual features into labels which are called as blobs. The words are in the form of keywords, therefore no further processing is required. The blobs and words are associated with certain probabilities using the translation approach. The translation table can then be used for two purposes: region naming and auto-annotation.

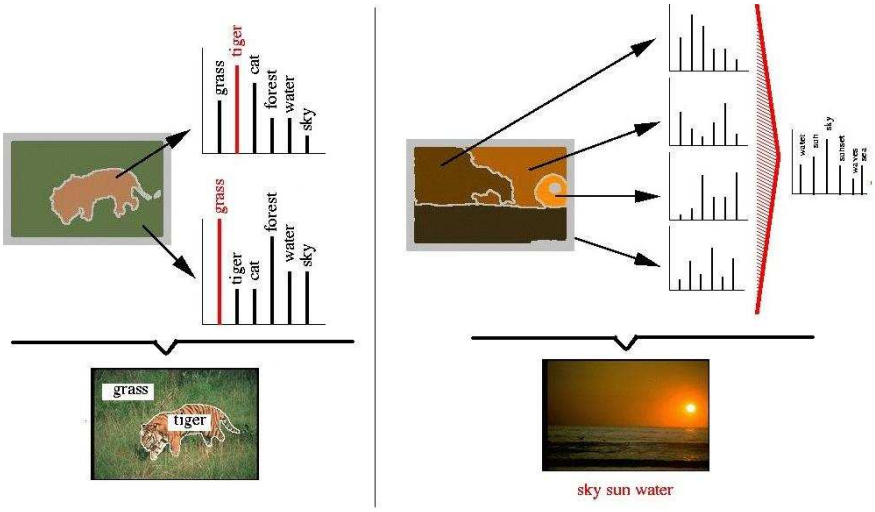


Fig. 5. Left: Region naming. **Right:** Auto-annotation. For region naming, the word with the highest probability is used to label the region. For auto-annotation the word posterior probabilities of the image regions are marginalized to obtain the probabilities for the entire image and then the top N words with the highest probabilities are used to annotate the image

Region naming refers to predicting the labels for the regions, which is clearly recognition. For region naming, given a blob b corresponding to the region, the word w with the highest probability ($p(w \mid b)$) is chosen and used to label the region (Fig.5).

In order to automatically annotate the images, the word posterior probabilities for the entire image are obtained by marginalizing the word posterior probabilities of all the blobs in the image as:

$$p(w|I_b) = 1/|I_b| \sum_{b \in I_b} p(w|b) \tag{5}$$

where b is a blob, I_b is the set of all blobs of the image and w is a word. Then, the word posterior probabilities are normalized. The first N words with the highest posterior probabilities are used as the annotation words (Fig.5).

The other correspondence problem that we attack is the video alignment problem. Specifically, we will concentrate on the video alignment problem in the news videos. In these videos, the speech transcript text is temporally aligned with the video frames and each shot is associated with a portion of the transcript that falls within its boundary. Most of the retrieval systems use the speech recognition text aligned with the shots to search for persons, places or events. However, the frames of the resulting shots may not visually correspond to the query (Fig.4). For example, in [31], it is shown that for person queries the name appears in a close proximity to the shot including the face of the person in the corresponding keyframe, but it can be a few seconds before or after.

We modify the translation approach to solve the correspondence problem between video frames and speech transcript text. For this purpose, we select the keyframes as the representative images for the shots and process the speech transcript text -which is in free text form- to obtain the descriptive words aligned with a given shot. The correspondence problem appears, since the words related to the visual content of the shot may be aligned not with the current shot but also with the neighboring shots. One solution is to use also the words aligned with the preceding and the following shots during the learning phase [11]. However, this strategy may use incorrect annotation words, since the speech transcript text a few shots before or after may correspond to other stories that are not related with the current shot.

News videos consist of story segments each corresponding to different topics (Fig.6). Using this characteristics of news videos, we use a story based approach. Each story is taken as the basic unit, and the correspondence problem is turned into finding the associations between the keyframes and the speech transcript words of the story segments. To make the analogy with the correspondence problem between image regions and annotation keywords, the story corresponds to image, the keyframes correspond to regions and speech transcript text corresponds to annotation keywords. The features extracted from the entire images of the keyframes are vector quantized to represent each image with a single label which is again referred as blob. Then, the translation tables are constructed similar to the one constructed for annotated images. The associations can then be used either to align the keyframes with the correct words or for predicting words for the entire story.

Story 1: (1-3) he says the u.s. may use force in a matter of weeks to try to compelling rock to allow u.n. weapons inspectors unrestricted access to suspected weapons sites russian news agencies reports iraqi president saddam was saying he's ready to allow inspectors to monitor eight new sites must the ground joining the sides of the latest u.s. defense secretary william cohen says that he's not an appropriate solution
Story 2: (4-5) darkness has led air transportation officials in the philippines to temporarily call of the helicopter searched for a missing passenger plane bound teams are continuing to look for the cebu pacific and d.c. nine it was carrying one hundred four people when it disappeared on its way from manila tuned and other parts of the southern philippines the pilot last contacted the airport tower minutes before that plane was supposed to land he made no mention of any trouble with the plane
Story 3: (6-7) the sarbes extending to unbeaten streak to five games we'll fight to win over the panthers final singer with the bow ahead goal detroit rallied with three goals in the final period
Story 4: (8) this is orelon sidney with your headline news weather update a low pressure storm moving out of the james bay region will mean a chance of snow flurries for the upper peninsula if michigan cold temperatures are due in the forecast for the north as the cold front moves into the mississippi and ohio river valleys

Fig. 6. Keyframes and speech transcripts for some stories from TRECVID2004 news videos. Numbers in paranthesis correspond to the keyframes of the stories

4 Data Sets and Input Representation

In this study, we use the annotated images from Corel stock phtograph data set and the news videos from TRECVID2004 corpus.

The Corel data set consists of images annotated with 3-5 keywords. We segment the images using the Normalized Cuts algorithm [28] and represent the 8 largest regions in each image with 30 features including the region size, position, color, texture and shape features. Regions are then clustered into blobs using k-means.

The TRECVID 2004 corpus [2] provided by NIST consists of over 150 hours of CNN and ABC broadcast news videos. The shot boundaries, and the keyframes extracted from each shot are provided by NIST. The keyframes are represented by a set of features including global color histogram, and mean and standard deviation of color, edge and texture features extracted from 5x7 grids. Videos are manually annotated with a collaborative effort of the TRECVID participants with a few keywords [20]. The automatic speech recognition (ASR) transcripts provided by LIMSI are aligned with the shots on the time basis [14]. The speech transcripts are in the free text form and requires preprocesing. First, we use Brill's part of speech tagger [7] to extract nouns which are expected to correspond to object names. Then, we apply a stemmer and remove the stop words and also the least frequent words appearing less than 300 times to obtain the descriptive words.

5 Measuring the performance

The trivial way to measure the performance of region naming is to check the labels of each region visually. However, considering the huge size of the data sets, this is not a practical solution. One alternative is to label the regions of a small set of images manually and then compare the predictions with the manual labels. Then, the performance can be measured in terms of recall and precision where recall is defined as the number of correct predictions of the word over the number of times that the word is a label word, and precision is defined as the number of correct predictions of the word over the number of times that the word is predicted.

Another solution, applicable to large number of images, is to predict the words for the entire images and use the annotation performance as a proxy. If the image has N annotation keywords, the system will also predict N words. A word prediction measure (WP) [3] can then be defined as:

$$WP = c/N \tag{6}$$

where c is the number of words predicted correctly. Thus, if there are three keywords, `sky`, `water`, and `sun`, then N=3, and we allow the model to predict 3 words for that image. The range of this score is clearly from 0 to 1.

Recall and precision can also be used to measure the annotation performance. In this case, the word is defined to be predicted correctly, if it is predicted as one of the best N words (where N is the number of words in the manual annotation) and it matches with one of the annotation keywords. Then, recall is defined as the number of times that the word is predicted correctly over the number of times that the word is used as an annotation keyword throughout the entire data set, and precision is defined as the number of times that the word is predicted correctly over the total number of times that is predicted.

The performance of video alignment can be measured similarly. We predict N words with the highest probability for a given story and compare them with the actual speech transcript words.

6 Results on Corel data set

For the experiments, we used 160 CD's, each consisting of 100 images on a relatively specific topic. The words occuring less than 20 times are excluded, resulting in vocabularies in the order of 155 words. As the visual features, color is represented by the average and standard deviation of (R,G,B) and (L,a,b) over the region; texture is represented using the average of 12 oriented energy filters aligned in 30 degree increments; and shape is represented by the ratio of the area to the perimeter squared, the moment of inertia and the region of the area to that of its convex hull. The features are quantized into 500 blobs using k-means.

Fig.7 shows some examples of region labeling. The label words are the words predicted with the highest probability for the corresponding blobs. We are generally successful in predicting words like `sky` and `buildings`. Rare words such as `plane` and `fish` are also predicted correctly in these examples.

In order to test the performance of region labeling, 450 images are manually labeled with a set of 117 words. Table 1 shows the region labeling performances in the form of recall and precision for a set of words.

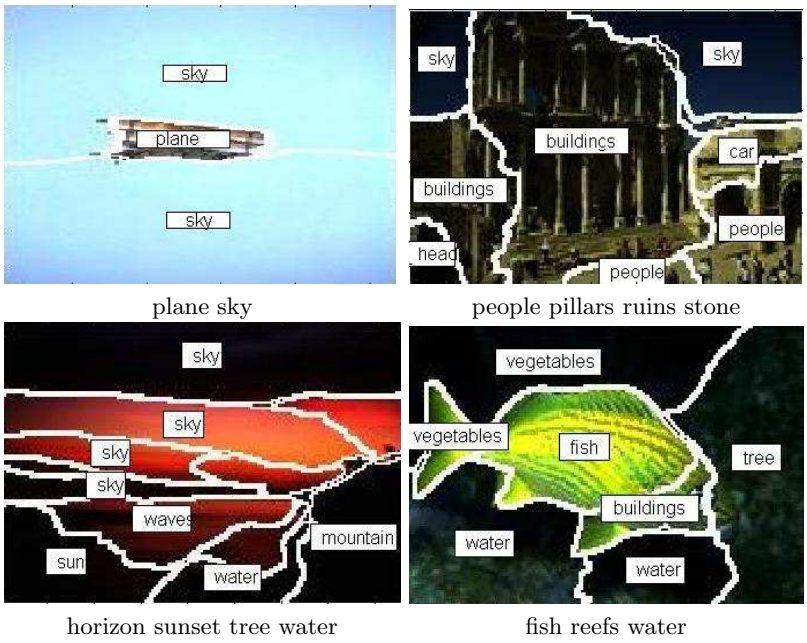


Fig. 7. Sample images and the word prediction results for the Corel data set. Manual annotations are shown for comparison

Table 1. Region labeling performance for some words on the Corel data set

word	recall	precision	word	recall	precision	word	recall	precision
sea	0.67	0.50	sky	0.31	0.34	windows	0.33	0.25
snake	0.20	0.33	water	0.40	0.20	buildings	0.16	0.17
tree	0.28	0.15	pillars	0.17	0.11	clouds	0.19	0.06
people	0.32	0.04	grass	0.09	0.19	flowers	0.08	0.16
car	0.10	0.12	coral	0.05	0.20	lion	0.05	0.17

Fig.8 shows some auto-annotation examples. Most of the words are predicted correctly and most of the incorrect matches are due to the missing manual annotations. For instance, although tree appears at the topleft image, the word `tree` it is not in the manual annotations.

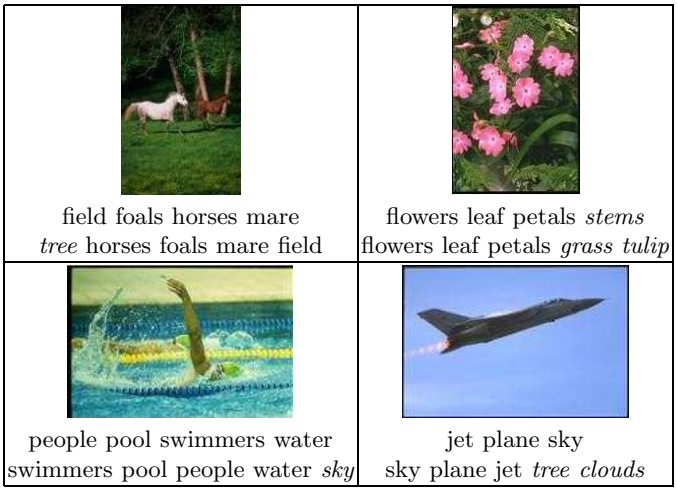


Fig. 8. Auto-annotation examples for Corel data set. The manual annotations are shown at the top, and the top 5 predicted words are shown at the bottom. Italic words correspond to incorrect matches

Table 2. Word prediction measures for each of the ten experimental data sets

set	training	standard test	novel test
001	0.2708	0.2171	0.2236
002	0.2799	0.2262	0.2173
003	0.2763	0.2288	0.2095
004	0.2592	0.1925	0.2172
005	0.2853	0.2370	0.2059
006	0.2776	0.2198	0.2163
007	0.2632	0.2036	0.2217
008	0.2799	0.2363	0.2102
009	0.2659	0.2223	0.2114
010	0.2815	0.2297	0.1991

In order to measure the performance of auto-annotation, we create ten experimental data set each consisting of 80 CDs which are randomly chosen. Each experimental data set is further split up into training and standard test sets, containing 75% and 25% of the images respectively. The images from the remaining CD's are used to form a more difficult novel test set.

Table 2 shows the word prediction results for each of the ten data sets on training, standard test and novel test sets. The average number of annotation words per image is three. The prediction performances show that on the average we are predicting one of the three words correctly.

7 Results on TRECVID data set

In the TRECVID 2004 corpus, there are 229 videos in the training set and 128 videos in the test set. On the average, there are around 300 keyframes for each shot. 114 videos from the training set are manually annotated by the TRECVID participants. We only use the annotations for the keyframes, and therefore eliminate the videos where the annotations are provided for the frames which are not keyframes, resulting in 92 videos. The original annotations consisting of 614 words have many spelling and format errors. After correcting the errors and removing the least frequent words we pruned the vocabulary downto 76 words.

		
StudioSetting Graphics FemaleNewsPerson MaleNewsSubject Person	People Basketball	WaterBody Boat
FemaleNewsPerson StudioSetting People MaleFace Graphics Person SceneText	People Graphics Basketball FemaleNewsPerson SceneText MaleNewsSubject StudioSetting	Sky Graphics WaterBody Building Boat Person MaleNewsPerson
		
Sky Building Road Car Graphics	Tree Snow People	Forest MaleNewsSubject FemaleFace Person Graphics
Road ManMadeObject People Sky Building Car ManMadeScene	Graphics People Person MaleFace MaleNewsSubject Tree Snow	People Person Graphics MaleFace Greenery SceneText FemaleFace

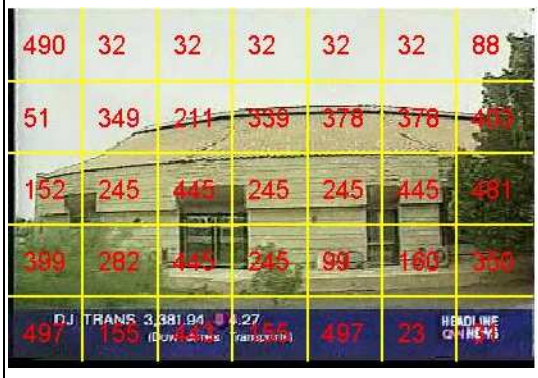
Fig. 9. Auto-annotation examples for the TRECVID data set. The manual annotations are shown at the top, and the predicted words (top 7 words with the highest probability) are shown at the bottom

We use the manually annotated data set for learning the correspondences between image regions (which are in the form of fixed sized grids) and the keywords for region labeling and for auto-annotation similar to the Corel data set. The grids are represented by the mean and standard deviation of HSV values. The features are clustered into 500 blobs. On the test data, we obtain word prediction performance as 0.27, and average recall and precision values for the words that are predicted at least once as 0.15 and 0.21 respectively.

Fig.9 shows the auto-annotation results for some sample shots. The results show that when the annotations are not available the predicted words can be used for a better retrieval. Fig.10 shows some region labeling results. Note that words like `female-news-person`, `female-face`, `studio-setting`, `sky` and `building` are correctly predicted.

studio-setting female-news-person

300,225: female-news-person
468,359,213: female-face
202,429,320,43,46,79: studio-setting
167,272,346,443: graphics
81,299: scene-text
104,404: person
223,475,317: male-face
437: people
61: flag
319: basketball



tree greenery sky building graphics

445,245: building
32: sky
403: man-made-object
350: greenery
152: tree
23,31,443: graphics
378: water-body
99: road, 349: snow
497,490: scene-text
51,88,339: person
282,481: male-news-subject
155: female-news-person
160: people
399: male-face
211: female-face

Fig. 10. Region labeling results for the TRECVID dataset. Manual annotations are shown for comparison

For video alignment, 114 videos are used for training and 39 videos are used for testing. The story boundaries provided by NIST are used. Speech transcript text (ASR) is processed by applying tagging, stemming and stop word elimination steps and only the nouns having frequencies more than 300 are used in the final vocabulary. We remove the stories associated with less than 4 words, and use the remaining 2165 stories consisting of 30801 keyframes for training and 1050 stories consisting of 10326 keyframes for testing. The number of words corresponding to the stories vary between 4 and 105, and the average number of words per story is 15. Each keyframe is represented by a blob obtained by vector quantization of HSV color histogram values extracted from the entire image and also with another blob corresponding to number of faces in the keyframe. Color feature is represented with 1000 blobs and face count is represented with 4 blobs.

The translation probabilities are used for predicting words for the individual shots (Fig.11) and for predicting words for the stories (Fig.12). The results show that especially for the stories related to weather, sports or economy, which frequently appear in the broadcast news, the system can predict the correct words. Note that, the system can predict words which are better than the original speech transcript words. This characteristic is important for a better retrieval. The prediction performance obtained by comparing the predicted words for a given story with the original ASR words is 0.15 and the average recall and precision values are 0.13 and 0.16 respectively.

An important aspect of predicting words for the video segments is to retrieve the related shots when speech transcript is not available or include unrelated words. In such cases it would not be possible to retrieve such shots with a text based retrieval system if the predicted words were not available. Fig.13 shows that the proposed system is able to detect the associations between the `sport` word and different types of sport scenes, and therefore can be used in retrieving sport shots even when the ASR is not available. Similarly, the system is successful in capturing the relationships between the visual features and words for scenes such as `snow`, `night` or `office` as in Fig.14 or objects such as `plane`, `house`, `water` or `car` as in Fig.15. Note that, these examples include objects and scenes which can be described by color information.

One of the main goals of solving the video alignment problem is to associate the words with the correct shots. Fig.16 shows an example to the solution of video alignment problem. Originally the word `clinton` was aligned with the anchorperson shot. After correcting the association problem, the shot which predicts `clinton` inside the story corresponds to the shot where Clinton appears. We should mention here that, this is not a solution to face recognition. In this example, the goal is to find the shot which has the highest probability to be associated with the `clinton` word inside the story segment. The third shot has the highest probability to be associated with `clinton` since it includes faces and also the black suits which can be described by color information. The second shot is probably eliminated since there were no faces detected, and the first shot is eliminated since the anchorperson shots having the studio setting at the background are associated with many words.

temperature weather forecast	point nasdaq stock	sport time game	jenning people evening

Fig. 11. Top three words predicted for some shots using the ASR outputs

ASR : center headline thunderstorm morning line move state area pressure chance shower lake head monday west end weekend percent temperature gulf coast tuesday
PREDICTED : weather thunderstorm rain temperature system shower west coast snow pressure

ASR : check peace york morning charge dollar share nasdaq market issue percent consumer month
PREDICTED : market stock york nasdaq street check point yesterday record share

ASR : night game sery story
PREDICTED : game headline sport goal team product business record time shot

Fig. 12. For sample stories corresponding ASR outputs and top 10 words predicted

Fig. 13. Shots having no attached ASR ouput but including `sport` keyword in their top 2 predicted words

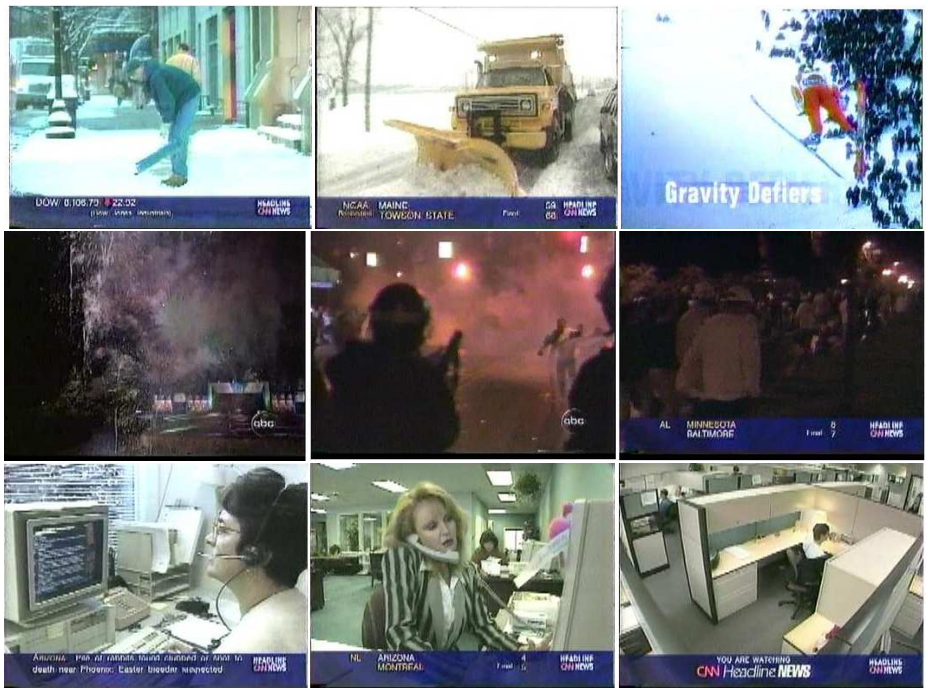


Fig. 14. Shots having no related ASR ouput but including `snow`, `night` and `office` keywords in their top 7 predicted words respectively

Fig. 15. Example shots predicting `plane`, `house`, `water` and `car` as their top 7th, 1st, 3rd and 7th words respectively

ASR outputs : (1) home washington president clinton (2) office president state department (3) deal

Fig. 16. For a story about Clinton with three shots, the keyframes and the ASR outputs associated with each of the shots on the time basis are shown. Note that, `clinton` is associated with the first shot where the anchorperson appears. When we search over the predicted words, the shot corresponding to `clinton` word with the highest probability is the third shot where Clinton actually appears

8 Conclusion and Future Work

We associate visual features with words using a translation approach. The proposed method allows novel applications on image and video databases including region naming as a way of recognizing objects, auto-annotation for better access to image databases and video alignment which is a crucial process for effective retrieval of video data.

In video data, motion information also plays an important role. Usually, moving objects have more importance than still objects. The regions corresponding to these objects can be extracted using the motion information rather than using any segmentation algorithm. Also, besides associating the visual features such as color, texture and shape with nouns for naming the objects, the motion information can be associated with verbs for naming the actions.

Translation approach can also be used as a novel method for face recognition. The correspondence problem that appears between the face of a person and his/her name can be attacked similarly for naming the people. The example about Clinton story promises that such an approach is possible for naming large number of faces.

9 Acknowledgements

This research is partially supported by TÜBİTAK Career grant number 104E065 and grant number 104E077.

References

1. Giza++. http://www.fjoch.com/GIZA++.html.
2. Trec vieo retrieval evaluation. http://www-nlpir.nist.gov/projects/trecvid.
3. 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.
4. K. Barnard, P. Duygulu, and D. A. Forsyth. Clustering art. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 434–439, 2001.
5. K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 408–415, 2001.
6. D. Blei and M. I. Jordan. Modeling annotated data. In *26th Annual International ACM SIGIR Conference*, pages 127–134, Toronto, Canada, July 28-August 1 2003.
7. E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, Trento, Italy, 1992.
8. 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.
9. P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Eight European Conference on Computer Vision (ECCV)*, Prague, Czech Republic, May 11-14 2004.
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, Copenhagen Denmark, May 27 - June 2 2002.
11. P. Duygulu and H. Wactlar. Associating video frames with text. In *Multimedia Information Retrieval Workshop in conjuction with the 26th annual ACM SIGIR conference on Information Retrieval*, Toronto, Canada, August 1 2003.
12. S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *the Proceedings of the International Conference on Pattern Recognition (CVPR 2004)*, volume 2, pages 1002–1009, 2004.
13. D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice-Hall, 2002.
14. J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.
15. A. Ghoshal, P. Ircing, and S. Khudanpur. Hidden markov models for automatic annotation and content based retrieval of images and video. In *The 28th International ACM SIGIR Conference*, Salvador, Brazil, August 15-19 2005.
16. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *26th Annual International ACM SIGIR Conference*, pages 119–126, Toronto, Canada, July 28-August 1 2003.
17. D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*. Prentice-Hall, 2000.
18. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *the Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems*, volume 16, pages 553–560, 2003.

19. J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, September 2003.
20. C.-Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum:establishing ground-truth labels on large multimedia datasets. In *NIST TREC-2003 Video Retrieval Evaluation Conference*, Gaithersburg, MD, November 2003.
21. C. D. Manning and H. S. utze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge Massachusetts, 1999.
22. O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *The Fifteenth International Conference on Machine Learning*, pages 341–349, 1998.
23. I. D. Melamed. *Empirical Methods for Exploiting Parallel Texts*. MIT Press, Cambridge Massachusetts, 2001.
24. F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, Berkeley, CA, USA, November 2003.
25. 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.
26. F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 1(29):19–51, 2003.
27. J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the 10th ACM SIGKDD Conference*, Seatle, WA, August 22-25 2004.
28. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.
29. P. Virga and P. Duygulu. Systematic evaluation of machine translation methods for image and video annotation. In *The Fourth International Conference on Image and Video Retrieval (CIVR 2005)*, Singapore, July 20-22 2005.
30. L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In *Proc. INTERACT : Conference on Human-Computer Interaction*, pages 326–333, Tokyo Japan, July 9-13 2001.
31. J. Yang, M.-Y. Chen, and A. Hauptmann. Finding person x: Correlating names with visual appearances. In *International Conference on Image and Video Retrieval (CIVR'04)*, Dublin City University Ireland, July 21-23 2004.