

---

# Linking image and text for semantic labeling of images and videos

Pınar Duygulu, Muhammet Baştan, Derya Ozkan

Bilkent University, Department of Computer Engineering, Ankara, 06800, Turkey  
{duygulu,bastan,deryao}@cs.bilkent.edu.tr

## 1 Introduction

There is an increasing demand to efficiently and effectively access large volumes of image and video collections. This demand leads to many systems to be introduced for indexing, searching and browsing of multimedia data. However, as observed in user studies [1, 2, 3], most of the systems do not satisfy the user requirements. The main bottleneck in building large scale and realistic systems is the disconnection between the low-level representation of the data and the high level semantics, which is usually referred as the “semantic gap” problem [4].

Early work on image retrieval systems were based on text input, in which the images are annotated by text and then text based methods are used for retrieval [5]. However, two major difficulties are encountered with text based approaches: First, manual annotation, which is a necessary step for these approaches, is labor-intensive and becomes impractical when the collection is large. Second, keyword annotations are subjective; the same image/video may be annotated differently by different annotators.

In order to overcome these difficulties, content-based image retrieval (CBIR) was proposed in the early 1990’s. In CBIR systems, instead of text-based annotations, images are indexed, searched or browsed by their visual features, such as color, texture or shape (see [6, 4, 7, 8, 9] for recent surveys in this area). However, such systems are usually based only on low level features and therefore cannot capture semantics.

In recent studies, the semantics derived from the text is incorporated with the visual appearance. The Blobworld system [10] use a simple conjunction of keywords and image features to search for the images. In [11], Cascia *et al.* unify the textual and visual statistics in a single indexing vector for retrieval of web documents. Similarly, Zhao and Grosky [12] use both textual keywords and image features to discover the latent semantic structure of web documents. In the work of Chen *et al.* [13], image and text features are used together to iteratively narrow the search space for browsing and retrieval of

web documents. Benitez and Chang [14, 15] combine the textual and visual descriptors in the annotated image collections for clustering and further for sense disambiguation. Srihari *et al.* [16, 17, 18] use the textual captions for the interpretation of the corresponding photographs, especially for face identification applications. The integration of textual and visual information is also commonly used in video indexing and retrieval [9, 19].

Although these methods provide better access to image and video collections, the “semantic gap” problem still exists. Many systems are proposed to bridge the semantic gap in the form of recognition of objects and scenes. However, most of these systems require supervised input for labeling and therefore cannot be adapted to large scale problems.

There are a variety of collections where images or videos are associated with some form of text. Examples include stock photographs annotated with keywords, museum images associated with metadata or descriptions, news photographs with captions, and news videos with associated speech transcript (Figure 1). These datasets do not provide one-to-one associations between semantics and appearance (e.g. labels for regions), but provide loosely labeled data (e.g. labels at the image level).



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

In recent studies, approaches that can use such loosely labeled data to learn links between multimedia and semantic data are introduced as a promising direction in reducing the manual effort for labeling.

The range of applications which make use of these links is large: improved search and browsing capabilities, automatic annotation of images, region naming as a possible direction to recognize large number of objects, face naming as a way of recognizing large number of people in different conditions, semantic alignment of video sequences with speech transcript text, etc.

In this chapter, first we concentrate on the image annotation and region naming problems for the semantic labeling of images, and discuss the recent studies in this direction. Then, we describe two other domains for associating the semantics with appearances. Namely, (i) association of visual elements and text in videos for recognition of scenes and objects, and (ii) association of names and faces for improving the performance of person queries.

## 2 Semantic labeling of images

In some studies, semi-automatic strategies are used for annotation of images. Wenyin *et al.* [20] use the query keywords which receive positive feedback from the user as possible annotations to the retrieved images. Similarly, Izquierdo and Dorado [21, 22] propose a semi-automatic image annotation strategy by first manually labeling a set of images and then extracting candidate keywords for annotating a new image from its most similar images after a frequent pattern mining process.

For automatic indexing of pictures, Li and Wang [23] propose a method which models image concepts by 2-D multi resolution Hidden Markov Models and then labels an image with the concepts best fit the content. In [24], Monay *et al.* extend Latent Semantic Analysis (LSA) models proposed for text and combine keywords with visual terms in a single vector representation. Singular Value Decomposition (SVD) is used to reduce the dimensionality and annotation is then performed by propagating the annotations of the most similar images in the corpus to the un-annotated image in the projected space.

In some other studies image annotation is viewed as a classification problem where the goal is to classify the entire image or the parts of the image into one of the categories corresponding to annotation keywords. In [25], Maron and Ratan propose multiple-instance learning as a way of classifying the images and use labeled images as bags of examples. Classifiers are built for each concept separately and an image is taken as positive if it contains a concept (e.g. waterfall) somewhere in the image and negative if it doesn't. Using a similar multiple instance learning formulation Argillander *et al.* [26] propose a Maximum Entropy based approach and build multiple binary classifiers to annotate the images. Carneiro and Vasconcelos [27] formulate the problem as M-ary classification where each of the semantic concepts of interest defines an image class and the classes directly compete at the annotation time. Feng *et al.* [28] develop a co-training framework to bootstrap the process of annotating large WWW image collections by exploiting both the visual contents and their associated HTML text and build text based and visual based classifiers using probabilistic SVM.

Recently, probabilistic approaches which models the joint distribution of words and image components are proposed. The first model proposed by Mori *et al.* [29] use a fixed size grid representation and obtain visual clusters by vec-

tor quantization of features extracted from these grids. The joint distribution of visual clusters and words are then learned using the co-occurrence statistics.

In [30] it is shown that learning the associations between visual elements and words can be attacked as a problem of translating visual elements into words. The visual elements, called as **blobs**, are constructed by vector quantization of the features extracted from the regions which are obtained using the Normalized Cuts segmentation algorithm [31]. Given a set of training images, the problem is turned into creating a probability table that translates blobs to words. The probability table is learned using a method adapted from the Statistical Machine Translation literature [32]. Different models are experimented in [33].

Pan *et al.* [34] use a similar blob representation and construct word-by-document and blob-by-document matrices. They discover the correlations between blobs and words based on the co-occurrence counts and also on the cosine similarity of the occurrence patterns - the documents including those items. For improvement, words and blobs are weighted inversely proportional to their occurrences and Singular Value Decomposition (SVD) is applied over the matrices to suppress the noise.

In [35], Jeon *et al.* adapt the relevance based language models [36, 37] to anotation problem and introduce cross-media relevance model. The images are represented in terms of both words and blobs. Given an image the probability of a word/blob is found as the ratio of the occurrence of the word/blob in the image to the total count of the word/blob in the training set. Then, they use the training set of annotated images to estimate the joint probability of observing a word with a set of blobs in the same image.

In the following studies of the same authors, the use of discrete blob representation is replaced with the direct modeling of continuous features, and two new models are proposed: Continuous Relevance Model [38, 39] and Multiple Bernoilli Relevance Model [40].

The model proposed by Barnard *et al.* [41, 42, 43] is a generative hierarchical aspect model inspired from Hofmann's model proposed for text [44, 45]. The model combines the aspect model with a soft clustering model. Images and corresponding words are generated by nodes arranged in a tree structure. Image regions represented by continuous features are modeled using a Gaussian distribution, and words are modeled using a multinomial distribution.

Blei and Jordan [46] propose Corr-LDA (correspondence latent Dirichlet allocation) model which finds conditional relationships between latent variable representations of sets of image regions and sets of words. The model first generates the region descriptions and then the caption words.

In [47], Monay and Gatica-Perez, modify the probabilistic latent space models to give higher importance to the semantic concepts and propose linked pair of PLSA models. They first constrain the definition of the latent space by focusing on textual features, and then learn visual variations conditioned on the space learned from text.

Carbonetto *et al.* [48] consider the spatial context and estimate the probability of an image blob being aligned to a particular word depending on the word assignments of its neighbouring blobs using Markov Random Fields.

Other approaches proposed for image annotation include Maximum Entropy based model [49], a method based on Hidden Markov Model [50], a graph based approach [51], and active learning method [52].

These methods learn the links in images. In the following, we present a method to generalize the problem to learn the links between textual and visual elements in videos.

### 3 Linking visual elements to words in news videos

Being an important source, broadcast news videos are acknowledged by NIST as a challenging data set and used for TRECVID Video Retrieval and Evaluation track since 2003 [53]. For retrieving the relevant information from the news videos, it is common to use speech transcript or closed caption text and perform text-based queries. However, there are cases where text is not available or errorful. Also, text is aligned with the shots only temporally and therefore the retrieved shots may not be related to the visual content (we refer this problem as “video alignment problem”). For example, when we retrieve the shots where a keyword is spoken in the transcript we may come up with visually non-relevant shots where an anchor/reporter is introducing or wrapping up a story (Figure 2). An alternative is to use the annotation words, but due to the huge amount of human effort required for manual annotation it is not practical.

Some of the studies discussed above are applied to video data, and the limited amount of manually annotated keyframes are used to annotate the other keyframes [33, 50]. However, since the vocabulary is limited and annotations are errorful, the usability of such data is narrow. On the other hand, speech transcript text is available for all the videos and provides an unrestricted vocabulary. However, since the frames are aligned with the speech transcript text temporally the semantic relationships are lost.

Now, we discuss how the methods proposed for linking images with words can be generalized to solve the video alignment problem and to recognize the objects and scenes in news videos and present an extended version of machine translation method proposed in [30] adapted for this problem.

#### 3.1 Translation approach to solve video alignment problem

An annotated image consists of two parts: set of regions and set of words. As discussed above, the goal was to learn the links between the elements of these two sets. A similar relationship occurs in video data. There are a set of video frames and a set of words extracted from the speech recognition text. While, the temporal alignment do not relate these two sets semantically, there is a



... (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 presidential 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. 2.** 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

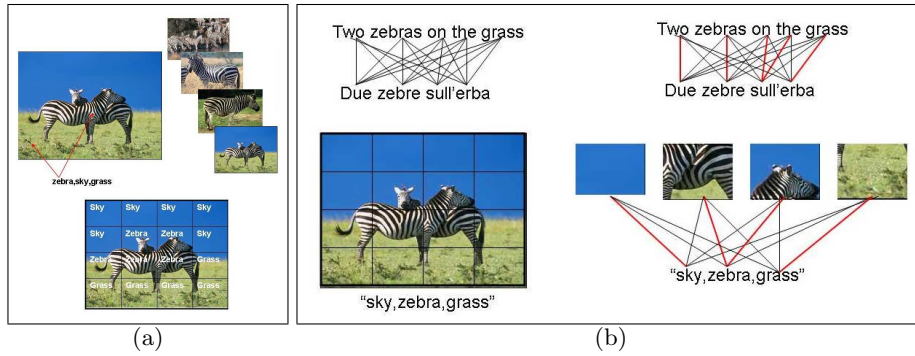
unit which provides a semantic association: a video story. Therefore, we can redefine the video alignment problem as finding the links between frames and words of a video story.

Then, in both situations, the problem is same. The correspondences between visual elements (image regions or video frames) and textual elements (keywords or speech transcript text) are unknown. (Figure 3-a). This correspondence problem is very similar to the correspondence problem faced in statistical machine translation literature (Figure 3-b) as first described in [30].

Brown *et.al* [32] suggested that it may be possible to construct automatic machine translation systems by learning from large datasets (aligned bitext) which 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 these 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.

Due to the similarity of the problems, learning the associations between visual elements and words can be attacked as a problem of translating visual features into words, as first proposed in [30]. Given a set of training images, the problem is to create a probability table that associates visual elements and words.

First, the visual features are transformed into discrete elements, called blobs, using a vector quantization technique such as k-means. The associations between blobs and words are then learned in the form of a probability table (also referred as the translation table), in which each entry indicates the



**Fig. 3.** (a) The correspondence problem between image regions and words. The words **zebra**, **grass** and **sky** 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. (b) The analogy with the statistical machine translation. We want to transform one form of data (image regions or English words) to another form of data (concepts or French words).

probability that a blob matches with a word. We use the Giza++ tool [54, 55] to learn the probabilities and adapt Model 1 of Brown et al. [32] in the form of direct translation.

Once learned, the translation table can be used to find the corresponding words for the given test images (**auto-annotation**), to label the image components with words (**region labeling**), and for ranked retrieval of images. For region naming, given a blob corresponding to a region, the word with the highest probability is chosen. For auto-annotation, the word posterior probabilities for an image are obtained by marginalizing the word posterior probabilities of all the blobs in the image and the first  $N$  words with the highest posterior probabilities are used to automatically annotate the image.

The translation approach to learn the associations between image regions and annotation words is then modified to solve the **video alignment** problem. Each story is taken as the basic unit, and the problem is turned into finding the associations between the key-frames and the speech transcript words of the story segments. To make the analogy with the association problem between image regions and annotation keywords, the stories correspond to images, the key-frames correspond to image regions and speech transcript text corresponds to annotation keywords. The features extracted from the key-frames are vector quantized to represent each image with labels which are again called blobs. Then, the translation tables are constructed similar to the one constructed for annotated images. The associations can then be used either to align the key-frames with the correct words or for predicting words for the entire stories.

### 3.2 Translation using manual annotations

In the experiments we use the TRECVID 2004 corpus provided by NIST which contains over 150 hours of CNN and ABC broadcast news videos. The shot boundaries and the keyframes extracted from each shot are provided by NIST. 114 videos are manually annotated with a collaborative effort of the TRECVID participants with a few keywords [56]. The annotations are usually incomplete.

In total, 614 words are used for annotation, most of which have very low frequencies, and with spelling and format errors. After correcting the errors and removing the least frequent words we pruned the vocabulary down to 62 words. We only use the annotations for the key-frames, and therefore eliminate the videos where the annotations are provided for the frames which are not key-frames, resulting in 92 videos with 17177 images, 10164 used for training and 7013 for testing.





The key-frames are divided into 5X7 rectangular grids and each grid is represented with various color (RGB, HSV mean-std) and texture (Canny edge orientation histograms, Gabor filter outputs) features. These features are then vector quantized using k-means to obtain the visual terms (blobs). The correspondences between the blobs and manual annotation words are learned in the form of a probability table using **Giza++** [54, 55]. The translation probabilities in the probability table are used for auto-annotation, region-labeling and ranked retrieval.





Fig. 4. Examples for blob-to-word matches.

Figure 4 and Figure 5 shows some region-labeling and auto-annotation examples. When predicted annotation words are compared with the actual annotation words, for each image, the average annotation prediction performance is around 30%. Since the manual annotations are incomplete (for example in the third example of Figure 5, although **sky** is in the picture and predicted it is not in the manual annotations), the actual annotation prediction performance should be higher than 30%.

Figure 6 shows query results for some words (with the highest rank). By visually inspecting the top 10 images retrieved for 62 words, the mean average precision (MAP) is determined to be 63%. MAP is 89% for the best

			
studio-setting female-news-person male-news-subject graphics person — female-news-person studio-setting people male-face graphics person scene-text	people basketball — people graphics basketball female-news-person scene-text male-news-subject studio-setting	water-body boat — sky graphics water-body building boat person male-news-person	sky building road car graphics — road man-made-object people sky building car man-made-scene

**Fig. 5.** Auto-annotation examples. The manual annotations are shown at the top, and the predicted words, top 7 words with the highest probability, are shown at the bottom.

	
weather-news	basketball
	
cartoon	female-news-person
	
meeting-room-setting	flower
	
monitor	food

**Fig. 6.** Ranked query results for some words using manual annotations in learning the correspondences.

(with highest precision) 30 words, and 99% for the best 15 words. The results show that when the annotations are not available the proposed system can effectively be used for ranked retrieval.

### 3.3 Translation using speech transcripts in story segments

The automatic speech recognition (ASR) transcripts for TRECVID2004 corpus are provided by LIMSI and are aligned with the shots on the time basis

[57]. The speech transcripts (ASR) are in the free text form and need preprocessing. Therefore, we applied tagging, stemming and stop word elimination and used only the nouns having frequencies more than 300 as our final vocabulary resulting in 251 words.

The story boundaries provided by NIST are used. We remove the stories associated with less than 4 words, and use the remaining 2503 stories consisting of 31450 key-frames for training and 2900 stories consisting of 31464 key-frames 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.

The key-frames are represented by blobs obtained by vector quantization of HSV, RGB color histograms, Canny edge orientation histograms, bags of SIFT keypoints extracted from entire images. The correspondences between the blobs and speech transcript words in each story segment are learned in the form of a probability table again using **Giza++**.



**Fig. 7.** Top three words predicted for some shots using ASR.

The translation probabilities are used for predicting words for the individual shots (Figure 7) and for the stories (Figure 8). 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.

Story based query results in Figure 9 show that the proposed system is able to detect the associations between the words (objects) and scenes. In these examples, the shots within each story are ranked according to the marginalized word posterior probabilities, and the shots matching the query word with highest probability are retrieved; a final ranking is done among all shots retrieved from all stories and all videos and final ranked query results are returned to the user.



Fig. 8. For sample stories corresponding ASR outputs and top 10 words predicted.

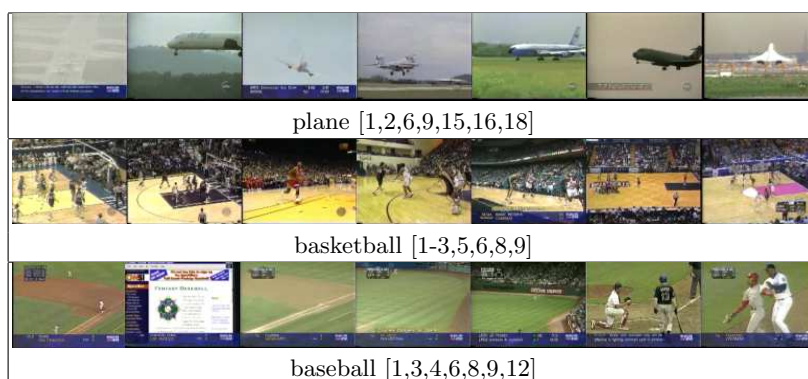


Fig. 9. Ranked story based query results for ASR. Numbers in square brackets show the rank of retrieval.

#### 4 Naming Faces in News

Another important problem which requires the integration of text and visual features is the retrieval of specific people. News, consist of stories about people, are the good sources for applying this problem. In news photographs on the web, a person’s face is likely to appear when his/her name is mentioned in the caption. Similarly, in a news video a person appears more frequently when the name is mentioned in speech. Now, we present a method to link the names with faces to recognize large number of faces appearing in news.

We first use the queried person’s name to limit our search space. We select a subset of faces for each person by searching for his/her name in the caption or in the speech transcript text. Although, there might be faces in this subspace corresponding to other people in the story, or some non-face images due to the errors of the face detection method used, the faces of the query name are likely to be the most frequently appearing ones than any other person in the same space. And also, even if the expressions or poses vary, different appearances of the face of the same person tend to be more similar to each other than to the faces of others.

In the second step, we find the similarities among the faces in the subspace of a queried person using visual information. Once a similarity measure is assigned between each pair of faces, the subspace can be represented a graph in which nodes correspond to faces and edges correspond to similarities. Then, the problem transforms into a graph problem. Hence, in the third step we find the biggest densest component in this subspace corresponding to the group of most similar faces, which are the faces belonging to the queried name.

#### 4.1 Integrating Names and Faces

The first step in our method is to integrate the face and the name information. We use the name information mainly to limit the search space, since a person is likely to appear in news around when his/her name is mentioned. Using this assumption, we reduce the face set for a queried person by choosing only the photographs that include the name of that person in the associated caption.

However, in news videos there is mostly a time shift between the visual appearance of a person and his/her name. In order to handle this alignment problem, we also choose one preceding and two succeeding shots along with the shot in which the name of the queried person is mentioned.

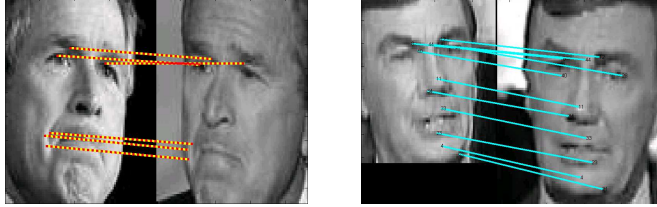
#### 4.2 Constructing the Similarity Graph

The similarity of faces are defined using the interest points extracted from the detected face areas. Lowe’s SIFT operator [58] are used for extracting the interest points.

The dissimilarity of two faces are computed based on the matching interest points. To find the matching interest points on two faces, each point on one face is compared with all the points on the other face and the points with the least Euclidean distance are selected. Since this method produces many matching points including the wrong ones, we apply two constraints to obtain only the correct matches, namely the geometrical constraint and the unique match constraint.

Geometrical constraint expects the matching points to appear around similar similar positions on the face when the normalized positions are considered. The matches whose interest points do not fall in close positions on the face are eliminated. Unique match constraint ensures that each point matches to only

a single point by eliminating multiple matches to one point and also by removing one-way matches. Example of matches after applying these constraints are shown in Fig. 10.



**Fig. 10.** Sample matching points for two faces from news photographs on the left and news videos on the right. Note that, even for faces with different size, pose or expressions the method successfully finds the corresponding points.

After applying the constraints, the distance between the two faces is defined as the average distance of all matching points between these two faces. A dissimilarity graph for all the faces in the search space is then constructed using these distances.

### 4.3 Finding the Densest Component

In the dissimilarity graph, faces represent the nodes and the distances between the faces represent the edge weights. We assume that, in this graph the nodes of a particular person will be close to each other (highly connected) and distant from the other nodes (weakly connected). Hence, the problem can be transformed in to finding the densest subgraph (component) in the entire graph. To find the densest component we adapt the method proposed by Charikar [59] where the density of subset  $S$  of a graph  $G$  is defined as

$$f(S) = \frac{|E(S)|}{|S|},$$

in which  $E(S) = \{i, j \in E : i \in S, j \in S\}$  and  $E$  is the set of all edges in  $G$  and  $E(S)$  is the set of edges induced by subset  $S$ . The subset  $S$  that has maximum  $f(S)$  is defined as the densest component.

Initially, the algorithm presented in [59] starts from the entire graph and in each step, the vertex of minimum degree is removed from the set  $S$ . The  $f(S)$  value is also computed for each step. The algorithm continues until the set  $S$  is empty. Finally, the subset  $S$  with maximum  $f(S)$  value is returned as the densest component of the graph.

In order to apply the above algorithm to the constructed dissimilarity graph, we need to convert it into a binary form, in which 0 indicates no edge and 1 indicates an edge between the two nodes. This conversion is carried out

by applying a threshold on the distance between the nodes. For instance, if 0.5 is used as the threshold value, then edges in the dissimilarity graph having higher value than 0.5 are assigned as 0, and others as 1. In other words, the threshold can be thought of an indicator of two nodes being near-by and/or remote.

#### 4.4 Experiments

First data set consists of a total of 30,281 detected faces from half a million captioned news images collected from Yahoo! News on the Web, which is constructed by Berg *et al.* [60]. Each image in this set is associated with a set of names. In the experiments, the top 23 people, whose name appears with the highest frequencies (more than 200 times) are used.

Average precision value for the baseline method is 48%, which assumes that all the faces appearing around the name is correct. With the proposed, method we achieved 68% recall and 71% precision values on the average. The method can achieve up to 84% recall and 100% precision for some people. We had initially assumed that, after associating names, true faces of the queried person appear more than any other person in the search space. However, when this is not the case, the algorithm gives bad retrieval results. For example, there is a total of 913 images associated with name *Saddam Hussein*, but only 74 of them are true *Saddam Hussein* images while 179 of them are *George Bush* images. Some sample images retrieved for three people are shown in Fig. 11.



**Fig. 11.** Sample images retrieved for three person queries in experiments. Each row corresponds to samples for George Bush, Hans Blix, and Colin Powell queries respectively.

Second data set is the broadcast news videos provided by NIST for TRECVID video retrieval evaluation competition 2004 [53]. It consists of 229 movies (30 minutes each) from ABC and CNN news. The shot boundaries and the key-frames are provided by NIST. Speech transcripts extracted by LIMSI [61] are used to obtain the associated text for each shot. The face detection algorithm provided by Mikolajczyk [62] is used to extract faces from key-frames. Due to high noise levels and low image resolution quality, the face detector produces many false alarms.

For the experiments, we choose 5 people, namely Bill Clinton, Benjamin Netanyahu, Sam Donaldson, Saddam Hussein and Boris Yeltsin. In the speech transcript text, their names appear 991, 51, 100, 149 and 78 times respectively.

When the shots including the query name are selected as explained above, the faces of the anchorpeople appear more frequently making our assumption that the most frequent face will correspond to the query name wrong. Hence, before applying the proposed method, we detect the anchorpeople and remove them from the selected shots by applying the densest component based method to each news video separately. The idea is based on the fact that, the anchorpeople are usually the most frequently appearing people in one broadcast news video. If we construct a similarity graph for the faces in a news video, the densest component in this graph will correspond to the faces of the anchorperson. We run the algorithm on 229 videos in our test set, and obtained average recall and precision values as 0.90 and 0.85 respectively.

We have recorded the number of true faces of the query name and total number of images retrieved as in Table 1. The first column of the table refers to total number of true images retrieved vs. total number of true images retrieved by using only the speech transcripts -selecting the shots within interval  $[-1,2]$ . The numbers after removing the detected anchorpeople by the algorithm from the text-only results are given in the second column. And the last column is for applying the algorithm to this set, from which the anchorpeople are removed. Some sample images retrieved for each person are shown in Fig. 12.



**Fig. 12.** Sample images retrieved for five person queries in experiments. Each row corresponds to samples for Clinton, Netanyahu, Sam Donaldson, Saddam, Yeltsin queries respectively.

As can be seen from the results, we keep most of the correct faces (especially after anchorperson removal), and we get reject many of the incorrect faces. Hence the number of images presented to the user is decreased. Also, our improvement in precision values are relatively high. Average precision of only text based results, which was 11.8% is increased by 29% to 15% after

**Table 1.** Numbers in the table indicate the number of correct images retrieved/total number of images retrieved for the query name.

Query name	Clinton	Netanyahu	Sam Donaldson	Saddam	Yeltsin
Text-only	160/2457	6/114	102/330	14/332	19/157
Anchor removed	150/1765	5/74	81/200	14/227	17/122
Method applied	109/1047	4/32	67/67	9/110	10/57

ancherperson removal, and by 152% to 29.7% after applying the proposed algorithm.

## 5 Conclusion and Discussion

In this chapter, we present recent approaches for semantic labeling of images and videos and describe two methods in detail: (i) translation approach for solving the correspondences between visual and textual elements and (ii) naming faces using a graph based method.

There are various number of other multimodal datasets which can make use of the methods discussed here.

## 6 Acknowledgements

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

## References

1. L. H. Armitage and P. Enser, “Analysis of user need in image archives,” *Journal of Information Science*, vol. 23, no. 4, pp. 287–299, 1997.
2. S. Ornager, “View a picture, theoretical image analysis and empirical user studies on indexing and retrieval,” *Swedish Library Research*, vol. 2-3, pp. 31–41, 1996.
3. M. Markkula and E. Sormunen, “End-user searching challenges indexing practices in the digital newspaper photo archive,” *Information retrieval*, vol. 1, pp. 259–285, 2000.
4. A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.
5. S. Chang and A. Hsu, “Image information systems: Where do we go from here?” *IEEE Trans. on Knowledge and Data Engineering*, vol. 4, no. 5, pp. 431–442, October 1992.
6. D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice-Hall, 2002.

7. A. A. Goodrum, "Image information retrieval: An overview of current research," *Informing Science*, vol. 3, no. 2, pp. 63–66, 2000.
8. Y. Rui, T. S. Huang, and S. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 4, pp. 39–62, 1999.
9. C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, January 2005.
10. C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, August 2002.
11. M. L. Cascia, S. Sethi, and S. Sclaroff, "Combining textual and visual cues for content-based image retrieval on the world wide web," in *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara CA USA, June 1998.
12. R. Zhao and W. I. Grosk, "Narrowing the semantic gap improved text-based web document retrieval using visual features," *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 189–200, 2002.
13. F. Chen, U. Gargi, L. Niles, and H. Schuetze, "Multi-modal browsing of images in web documents," in *Proceedings of SPIE Document Recognition and Retrieval VI*, 1999.
14. A. B. Benitez and S.-F. Chang, "Perceptual knowledge construction from annotated image collections," in *IEEE International Conference On Multimedia and Expo (ICME-2002)*, Lausanne, Switzerland, August 2002.
15. ———, "Semantic knowledge construction from annotated image collections," in *IEEE International Conference On Multimedia and Expo (ICME-2002)*, Lausanne, Switzerland, August 2002.
16. R. Srihari, "Extracting visual information from text: Using captions to label human faces in newspaper photographs," Ph.D. dissertation, Department of Computer Science, SUNY at Buffalo, 1991.
17. R. K. Srihari and D. Burhans, "Visual semantics: Extracting visual information from text accompanying pictures," in *AAAI 94*, Seattle, WA, 1994.
18. R. Srihari, R. Chopra, D. Burhans, M. Venkataraman, and V. Govindaraju, "Use of collateral text in image interpretation," in *ARPA Workshop on Image Understanding*, Monterey, CA, 1994, pp. 897–905.
19. S.-F. Chang, R. Manmatha, and T.-S. Chua, "Combining text and audio-visual features in video indexing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 18–23 2005.
20. 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*, Tokyo Japan, July 9–13 2001, pp. 326–333.
21. A. Dorado and E. Izquierdo, "Semi-automatic image annotation using frequent keyword mining," in *Proceedings of the Seventh International Conference on Information Visualization (IV03)*, 2003.
22. E. Izquierdo and A. Dorado, "Semantic labelling of images combining color, texture and keywords," in *Proc. IEEE Int. Conf. on Image Processing (ICIP2003)*, Barcelona, Spain, September 2003.

23. J. Li and J. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, September 2003.
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. O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *The Fifteenth International Conference on Machine Learning*, 1998.
26. J. Argillander, G. Iyengar, and H. Nock, "Semantic annotation of multimedia using maximum entropy models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 18-23 2005.
27. G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, 2005.
28. H. Feng, R. Shi, and T.-S. Chua, "A bootstrapping framework for annotating and retrieving www images," in *Proceedings of the 12th annual ACM international conference on Multimedia*, New York, NY, USA, 2004, pp. 960 – 967.
29. 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.
30. 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)*, vol. 4, Copenhagen Denmark, May 27 - June 2 2002, pp. 97–112.
31. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, August 2000.
32. P. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
33. 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.
34. J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME2004)*, Taipei, Taiwan, June 27-30 2004.
35. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *26th Annual International ACM SIGIR Conference*, Toronto, Canada, July 28-August 1 2003, pp. 119–126.
36. V. Lavrenko and W. B. Croft, "Relevance-based language models," in *24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, New Orleans, LA, September 7-12 2001.
37. V. Lavrenko and M. C. W. Croft, "Cross-lingual relevance models," in *25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Tampere, Finland, August 11 - 15 2002.
38. 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*, vol. 16, 2003, pp. 553–560.

39. V. Lavrenko, S. L. Feng, and R. Manmatha, "Statistical models for automatic video annotation and retrieval," in *the Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2004)*, Montreal, QC, Canada, 2004.
40. 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)*, vol. 2, 2004, pp. 1002–1009.
41. K. Barnard and D. A. Forsyth, "Learning the semantics of words and pictures," in *Int. Conf. on Computer Vision*, 2001, pp. 408–415.
42. K. Barnard, P. Duygulu, and D. A. Forsyth, "Clustering art," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2001, pp. 434–439. [Online]. Available: <http://kobus.ca/research/publications/CVPR-01/clustering-art.pdf>
43. K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003. [Online]. Available: <http://kobus.ca/research/publications/JMLR-03/JMLR-03.pdf>
44. T. Hofmann, "Learning and representing topic. a hierarchical mixture model for word occurrence in document databases," in *Proceedings of the Conference for Automated Learning and Discovery (CONALD)*, Pittsburgh, 1998.
45. T. Hofmann and J. Puzicha, "Statistical models for co-occurrence data," AI Memo 1625, CBCL Memo 159, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, MIT, Tech. Rep., February 1998.
46. D. Blei and M. I. Jordan, "Modeling annotated data," in *26th Annual International ACM SIGIR Conference*, Toronto, Canada, July 28-August 1 2003, pp. 127–134.
47. F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: Constraining the latent space," in *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, New York, October 2004.
48. 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.
49. J. Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation," in *the Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR 2004)*, Dublin City University, Ireland, July 21-23 2004, pp. 24–32.
50. 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.
51. 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*, Seattle, WA, August 22-25 2004.
52. R. Jin, J. Y. Chai, and S. Luo, "Automatic image annotation via coherent language model and active learning," in *The 12th ACM Annual Conference on Multimedia (ACM MM 2004)*, New York, USA, October 10-16 2004.
53. "Trec viedo retrieval evaluation," <http://www-nlpir.nist.gov/projects/trecvid>. [Online]. Available: <http://www-nlpir.nist.gov/projects/trecvid>
54. "Giza++," <http://www.fjoch.com/GIZA++.html>. [Online]. Available: <http://www.fjoch.com/GIZA++.html>

55. F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 1, no. 29, pp. 19–51, 2003.
56. 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.
57. J. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.
58. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.
59. M. Charikar, "Greedy approximation algorithms for finding dense components in a graph," in *APPROX '00: Proc. of the 3rd International Workshop on Approximation Algorithms for Combinatorial Optimization*, London, UK, 2000.
60. T. Berg, A. C. Berg, J. Edwards, and D. Forsyth, "Who is in the picture," in *Neural Information Processing Systems (NIPS)*, 2004.
61. J. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, 2002.
62. K. Mikolajczyk, "Face detector," INRIA Rhone-Alpes, 2004, ph.D Report.