Recognizing objects and scenes in news videos

Muhammet Baştan and Pınar Duygulu

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

Abstract. We propose a new approach to recognize objects and scenes in news videos motivated by the availability of large video collections. This approach considers the recognition problem as the translation of visual elements to words. The correspondences between visual elements and words are learned using the methods adapted from statistical machine translation and used to predict words for particular image regions (region naming), for entire images (auto-annotation), or to associate the automatically generated speech transcript text with the correct video frames (video alignment). Experimental results are presented on TRECVID 2004 data set, which consists of about 150 hours of news videos associated with manual annotations and speech transcript text. The results show that the retrieval performance can be improved by associating visual and textual elements. Also, extensive analysis of features are provided and a method to combine features are proposed.

1 Introduction

Due to the rapidly growing quantities of digital image and video archives, effective and efficient indexing, retrieval and analysis of such data have received significant attention. Being an important information source, applications on broadcast news videos are especially challenging. This challenge is also acknowledged by NIST and news videos are chosen as the data set for the TRECVID Video Retrieval Evaluation [2].

It is common to use speech transcript or closed caption text and perform text-based queries to retrieve the relevant information. 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. 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. 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. Recognition of objects and scenes is the ultimate solution but recognition on the large scale is still a challenge.

Recently, it has been shown that large number of objects can be recognized without supervision by using large annotated image collections [3, 4, 6, 7]. In general, the proposed models are based on learning the associations between image regions and annotation words.

In this study, we extend these methods to recognize objects and scenes in news videos. To learn the associations, we adapt the translation approach [7] inspired from the models proposed for statistical machine translation [5]. Our method learns the correspondences between visual features extracted from the video shots with the annotation words from a small number of videos. Then the correspondences are used for predicting words for individual regions (**region-labeling**) and for entire images (**auto-annotation**) in the rest of the data.

Methods which use manual annotations to automatically annotate the video shots are also proposed in [9, 11]. However, in those methods the associations between image regions and words are not explicitly learned and labeling of individual regions for recognition of objects is not provided.

Since the annotation words are not always available and reliable, as an alternative, we propose to use the speech transcript or closed caption text, which is the main contribution of this work. There is an alignment problem between the text and the visual content, and taking the text temporally aligned with the shot is problematic. One solution is to also use the words aligned with the preceding and following shots as in [8, 16]. However, the speech transcript text a few shots before or after may correspond to other stories that are not related with the current shot resulting in association of irrelevant words with the shot.

As a solution, we propose a story-based approach where we treat each story as a document containing associated elements. The translation approach is modified to find the correspondences between the key-frames and the speech transcript words of the story segments. This process, which we refer as **video alignment** enables a textual query to return semantically more accurate images.

While, the effect of features extracted from entire images or from image regions are heavily experimented for automatic annotation, the features extracted around salient points -which are shown to be successful for recognition of objects and scenes- is not well investigated. In this study, we provide an extensive analysis of features by (i)investigating the effect of extracting features from entire images, from fixed sized grids and around interest points, (ii) by experimenting the SIFT descriptors [13] besides the commonly used color, texture and edge features and (iii) applying the bag-of-visterms approach [14] -which has recently been proposed for classification of scenes- to the association problem. Moreover, we propose a new method to combine the features using the prediction probabilities and show that the performance improves.

In this study, we use videos from TRECVID 2004 corpus [2] which consists of over 150 hours of CNN and ABC broadcast news videos provided by NIST. The results show that retrieval performance can be improved by associating the visual elements with words as a way of recognizing objects and scenes on the large scale.

First, we will describe the method to translate visual elements to words briefly and explain our performance evaluation measures. Then, we will present the results for two separate cases: using the manual annotation words and using the speech transcript text. Finally, we will present a detailed analysis of features used in the study.

2 Translating visual elements to words

Learning the associations between visual elements and words can be attacked as a problem of translating visual features into words, as first proposed in [7]. 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 to as translation table), in which each entry indicates the probability that a blob matches with a word. In this study, we use the Giza++ tool [1, 15] to learn the probabilities and adapt Model1 of Brown et al. [5] 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 can be 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 using k-means 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 Performance measurement

We define the annotation performance for an image as the number of correct predictions divided by the number of actual annotation words for that image. The annotation performance is averaged over all test images to obtain the average annotation performance (aap) for an image. We similarly define recall and precision for each word. A word is defined to be predicted correctly, if it matches with one of the actual annotation words. Recall is the number of times that the word is correctly predicted over the number of times that the word is used as an annotation word throughout the entire data set, and precision is the number of times that the word is predicted correctly over the total number of times it is predicted. Average recall and precision are calculated by considering the words that are predicted at least once.

For each image, we can choose to predict as many words as there are in the actual annotation, which we refer to as $case_1$, or a fixed number of words, which we refer to as $case_2$.

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

4 Translation using manual annotations

In the TRECVID 2004 corpus, there are 229 videos in the training set and 128 videos in the test set. We use the shot boundaries and the key-frames provided by NIST. On the average, there are around 300 key-frames for each video. 114 videos from the training set are manually annotated with a collaborative effort of the TRECVID participants with a few keywords [12]. In total, there are 614 words used for annotation, most of which have very low frequency, 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.

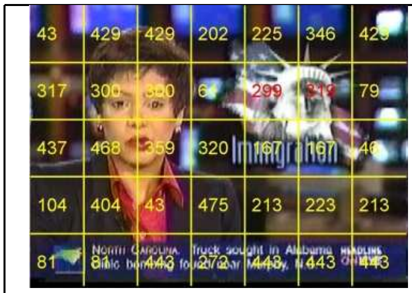
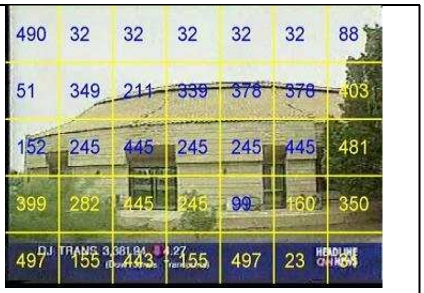
	
studio-setting female-news-person	**tree greenery sky building graphics**
300,225: female-news-person 468,359,213: female-face 202,429,320,43,46,79: studio-setting 167,272,346,443: graphics; 81,299: scene-text; 104,404:person 223, 475,317: male-face; 437: people 61: flag; 319: basketball	445,245: building; 403: man-made-object, 32: sky; 350: greenery; 152: tree 23,31,443: graphics; 497,490: scene-text 378: water-body; 99: road; 349: snow 51,88,339: person; 211: female-face 282,481: male-news-subject; 160: people 155: female-news-person; 399: male-face

Fig. 1. Example region labeling results. Manual annotations are shown for comparison.

We use the manually annotated data set to learn the correspondences between blobs and words for region naming and for auto-annotation. Figure 1 shows some region labeling results. Note that words like `female-news-person`, `female-face`, `studio-setting`, `sky` and `building` are correctly predicted. Example blobs corresponding to some words with high prediction accuracies are shown in Figure 2.

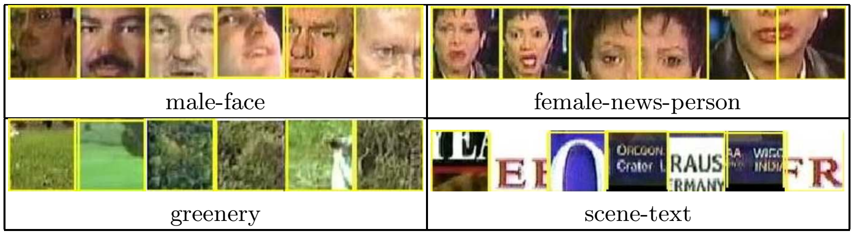


Fig. 2. Examples for blob-to-word matches.

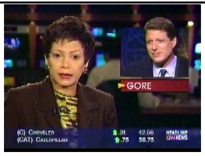
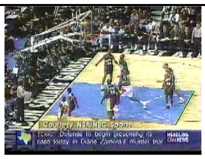
			
studio-setting female-news-person male-news-subject graphics person — female-news-person studio-setting people male-face graphics person scene-text	people basketball — people graphics basketball female-news-person scene-text male-news-subject studio-setting	water-body boat — sky graphics water-body building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

Fig. 3. 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.

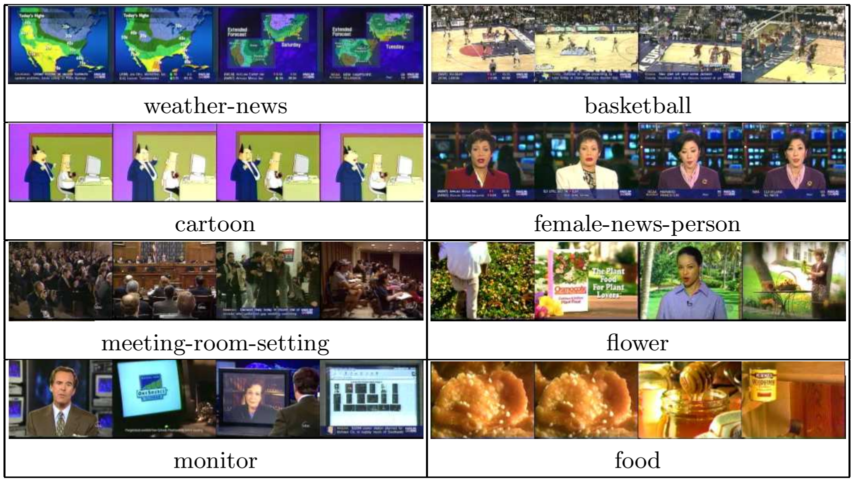


Fig. 4. Ranked query results for some words using manual annotations.

Some auto-annotation examples are shown in Figure 3. On the average, we obtain an annotation performance around 30%. We should note that the performances are calculated by comparing the predicted annotations with the manual annotations. Since manual annotations are incomplete (for example in the third example of Figure 3, although **sky** is in the picture and predicted it is not in the manual annotations) the calculated values may be lower than the actual ones.

Figure 4 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 (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.

5 Translation using speech transcripts in story segments

For the experiments using speech transcript text, 111 videos are used for training and 110 videos are used for testing. The automatic speech recognition (ASR) transcripts provided by LIMSI are aligned with the shots on the time basis [10]. The speech transcripts (ASR) are in the free text form and requires preprocessing. Therefore, we applied tagging, stemming and stop word elimination steps 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 translation probabilities are used for predicting words for the individual shots (Figure 5) and for the stories (Figure 6). 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.

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. Story based query results in Figure 7 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.

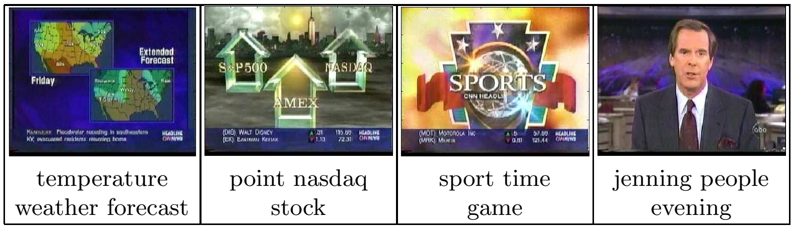
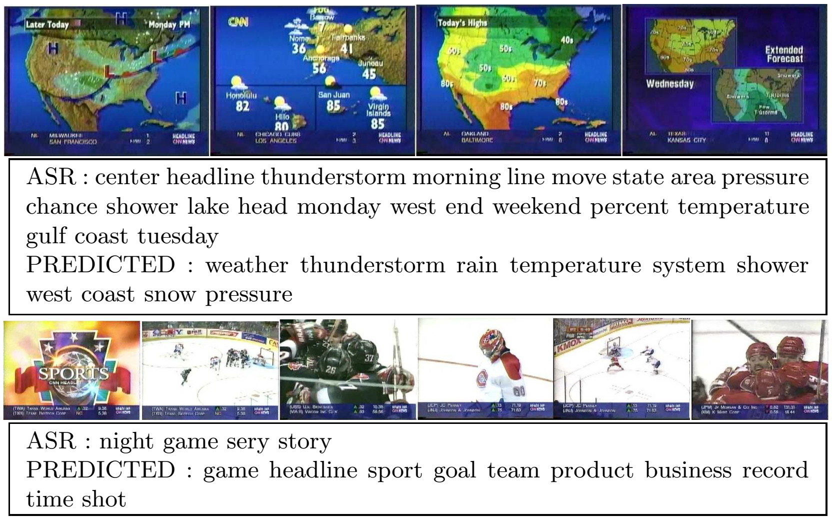


Fig. 5. Top three words predicted for some shots using the ASR outputs.



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

Fig. 6. For sample stories corresponding ASR outputs and top 10 words predicted.

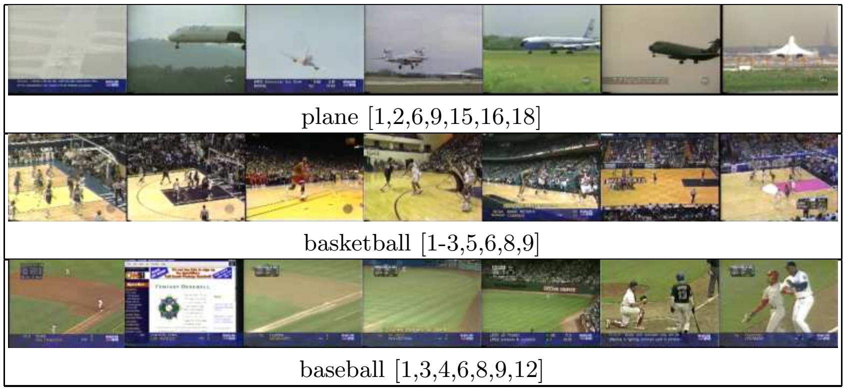


Fig. 7. Ranked story based query results for ASR. Numbers in square brackets show the rank of retrieval.

6 Analysis of Features

For manual annotations and ASR experiments, the key-frames are represented by a set of features including global histograms extracted from entire images, and local statistics extracted from grids or around keypoints.

Color features are extracted for RGB and HSV color spaces, texture is represented as Gabor filter outputs, and Canny edge detector outputs are used for edges. Global features are represented by 64 bin RGB, 162 bin HSV and 16 bin edge histograms from entire images; while local features are extracted from 5x7 fixed sized grids as mean and standard deviation of color, Gabor filter output, and 8 bin Canny edge histogram.

The keypoints are detected and represented using Lowe's SIFT operator[13]. Using the binaries provided by the author, large number of keypoints are extracted. In order to keep the number of features in the order of those extracted from grids, we chose 35 keypoints with maximum scale. In addition to the 128-element SIFT descriptor vectors, mean and standard deviation of color, texture and edge features are also extracted around keypoints similar to features extracted from grids.

We also experimented with the bag-of-visterms approach [14] by taking about 600.000 keypoints extracted from 5 videos, vector quantizing them and forming a keypoint histogram with 1000 bins for each image. The keypoint histograms, as feature vectors for each image, are quantized to obtain the final blobs.

Some words may be predicted better using one feature than others. For example, color is a good cue for commercial and cartoon scenes while edge or texture is good for basketball or studio scenes. If the outputs of multiple features, some of which can predict some words better than others, are combined, then the prediction performance of the system is expected to improve. We combined the outputs of several features at the word prediction step by marginalizing the word posterior probabilities (over all blobs) obtained from several features. If the output one feature is high it is reflected on the final output. As shown in Table 1 and Table 2, on the average, the prediction performance of the system is always improved. The improvement is more notable in the average word precision values.

In Table 1, the results are shown for different features in the form of annotation performance and average word recall and precision values for the case of translation with manual annotations. Note that the performance is always better if the outputs of multiple features are combined as explained above. The performance when SIFT descriptors are used is inferior to the grid based features. Although average word recall and precision values are close to those of other features, the number of words with nonzero prediction is significantly less. The reason is mainly due to the lost color information which is very important for the discrimination of most objects and scenes, and also due to using only the maximum scale 35 keypoints.

Using the number of faces detected per image as additional information does not improve the performance significantly. Increasing the number of blobs improves the performance but the computational cost also increases; therefore, we choose 2500 as an appropriate number.

The prediction performances obtained by comparing the predicted words for a given story with the original ASR words for some features are summarized in Table 2. The performance with the bag-of-visterms approach is better com-

case	Performance	1	2	3	4	5	6	7	8	9
$case_1$	aap	0.266	0.267	0.274	0.277	0.276	0.288	0.295	0.235	0.271
$case_1$	recall	0.144	0.149	0.142	0.169	0.170	0.178	0.183	0.155	0.148
	precision	0.218	0.231	0.217	0.294	0.322	0.334	0.390	0.266	0.245
$case_2$	recall	0.323	0.328	0.331	0.327	0.330	0.333	0.344	0.275	0.319
	precision	0.082	0.082	0.081	0.087	0.089	0.089	0.110	0.081	0.079

Table 1. Automatic annotation performances using manual annotations. For details, please see section 3. For $case_2$, 10 words are predicted per image. Numbers 1 through 9 at the top stands for the following features: 1,4,9: mean&std of color, 2,5: mean&std of color + edge, 3,6: mean&std of color + texture, 7: combination of outputs of the first 6 features; 8: SIFT descriptors. In (1-7) features are extracted from 5x7 grids, and in (8-9) features extracted around maximum scale 35 keypoints. In (1-3) HSV and in (4-6,9) RGB is used as the color feature.

case	Performance	1	2	3	4	5	6	7	8	9
$case_1$	aap	0.156	0.155	0.172	0.173	0,182	0,183	0,194	0.190	0.200
$case_1$	recall	0.141	0.142	0.160	0.162	0,149	0,150	0,207	0.152	0.170
	precision	0.159	0.164	0.187	0.195	0,207	0,214	0,275	0.226	0.236
$case_2$	recall	0.192	0.193	0.218	0.221	0,169	0,165	0,189	0.200	0.224
	precision	0.102	0.102	0.118	0.119	0,107	0,108	0,136	0.127	0.136

Table 2. Automatic story annotation performances using ASR. Number of blobs = 1000. For $case_2$, 25 words are predicted per story. Numbers 1 through 7 at the top stands for the features: 1,3: global HSV, RGB histograms, 2,4: global HSV, RGB + Canny edge histograms, 5,6: mean&std of HSV, RGB + texture from 5X7 grids, 7: combination of (1-6), 8: bag-of-visterms approach, 9: combination of (1-6,8).

pared to the color and texture features although only 5 videos are used in the construction of the bag-of-visterms due to large computational cost. As in the manual annotation case, performance is improved when multiple feature outputs are combined.

7 Conclusion and Future Work

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

In video data, motion information also plays an important role. Usually, moving objects are important 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 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 people.

8 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 video 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. 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.
5. 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.
6. 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.
7. 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.
8. 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*, Canada, 2003.
9. 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.
10. J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.
11. A. Ghoshal and P. Ircing and S. Khudanpur. Hidden Markov Models for Automatic Annotation and Content Based Retrieval of Images and Video. The 28th International ACM SIGIR Conference, Brazil, 2005.
12. 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.
13. David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004.
14. P. Quelhas and F. Monay and J.-M. Odobez and D. Gatica-Perez and T. Tuytelaars and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *International Conference on Computer Vision (ICCV)*, 2005.
15. F. J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, volume 1, pages 19–51, 2003.
16. 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.