Recognizing objects and scenes in news videos

Muhammet Bastan & Pınar Duygulu

RETINA Vision & Learning Group

Bilkent University

Ankara, **TURKEY**

http://retina.cs.bilkent.edu.tr

{bastan, duygulu}@cs.bilkent.edu.tr

Motivation

Goal: effective retrieval and analysis of large video collections

Solution required: labeling of objects and scenes

Problems: manual annotation is not practical

recognition on the large scale is still a challenge

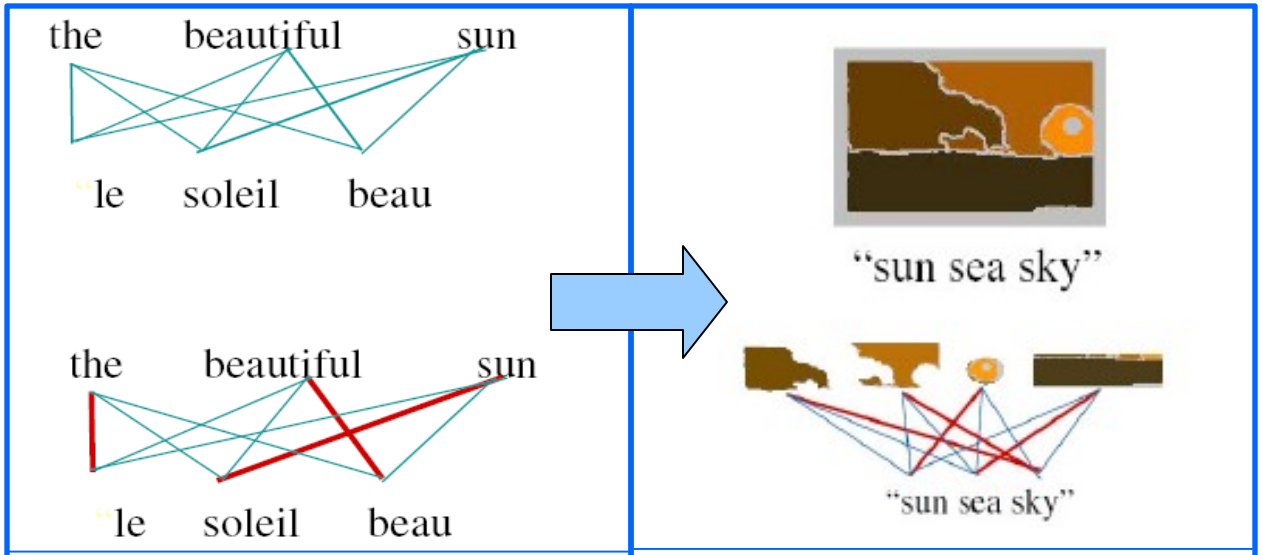
Our Approach: Translation of visual elements to words for

- recognition of objects and scenes on the large scale
- automatic annotation
- better retrieval and analysis

Data Set:

- TRECVID 2004 news videos (~ 350 videos, each ~ 300 keyframes)
- 114 videos manually annotated by TRECVID participants
- shot & story boundaries, key-frames and speech transcripts are provided by NIST

Statistical machine translation methods are adapted



Statistical machine translation

- Learn the correspondences between the words of the source and target language using a large corpus
- Translate from source to target language using the learned correspondences (Brown 1993)

Translation of visual elements to words

- Learn the correspondences between image regions and annotation words using a set of annotated images
- Annotate new images using the learned correspondences (Duygulu 2002)

Procedure:

- Extract features from video frames
- Vector quantize the features using k-means to transform into discrete elements (blobs)
- Learn the associations between the blobs and words using **Giza++** to obtain a probability table:

	word 1	word 2	…
blob 1	p(1,1)	p(1,2)	…
blob 2	p(2,1)	…	…
…	…	…	…

Probability that blob2 matches with word1, P(word1 | blob2)

- Use this table to label new images, image regions, etc.

Translating visual elements to words

- similar approach is applied to automatically annotate video frames and to label regions
- however, manual annotations are not always available

Alternative: use speech transcript text associated with the videos

Problem: speech transcript text is aligned with the shots on the time basis and usually does not correspond to visual information

Solution: learn the correspondences between video frames and speech transcript text inside the news stories

Story 1: weapon inspector iraq president saddam secretary ...

Story 2: air plane transportation ...

Story 3: game final goal ...

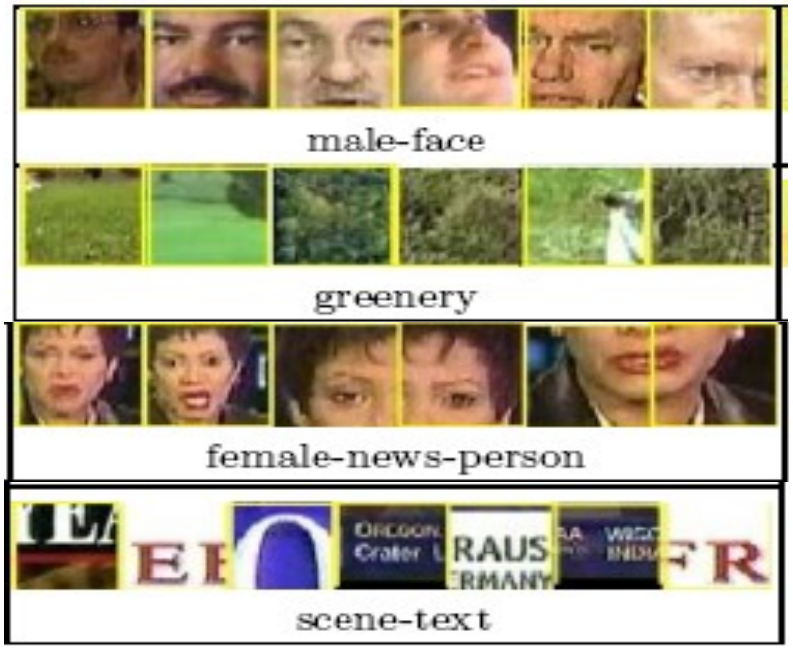
Story 3: weather ...

News stories and associated preprocessed speech transcript text

Labeling results using manual annotations

Data: training set : 10164 images, test set : 7013 images from TRECVID2004
Features: HSV, RGB mean-std, texture (Gabor, Canny) from 5X7 grids; SIFT descriptors, color features around keypoints
Keywords: 62 keywords corresponding to nouns

Auto-annotation examples
(top : actual, bottom : predicted)



Region labeling examples

Annotation Performance

- Average annotation prediction performance per image : 30%
- For words that are predicted at least once, average recall 18%, average precision 33%

Retrieval results using manual annotations

Ranked query results for: weather-news, cartoon, meeting-room-setting, basketball, female-news-person, food, monitor

Queries are performed for all the keywords and first 10 images that are retrieved are examined manually

Retrieval Performance – MAP

1st 10 images	MAP (%)
62 words	63
best 30 words	89
best 15 words	99

Labeling results using speech transcript text

Data: 31450 frames for training, 31464 frames for test

Features: global HSV, RGB, edge histograms, SIFT descriptors

- Speech transcripts are in free text form and needs preprocessing. Apply tagging, stemming and stop-word elimination to obtain 251 keywords corresponding to nouns

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

Shot and story annotation examples

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

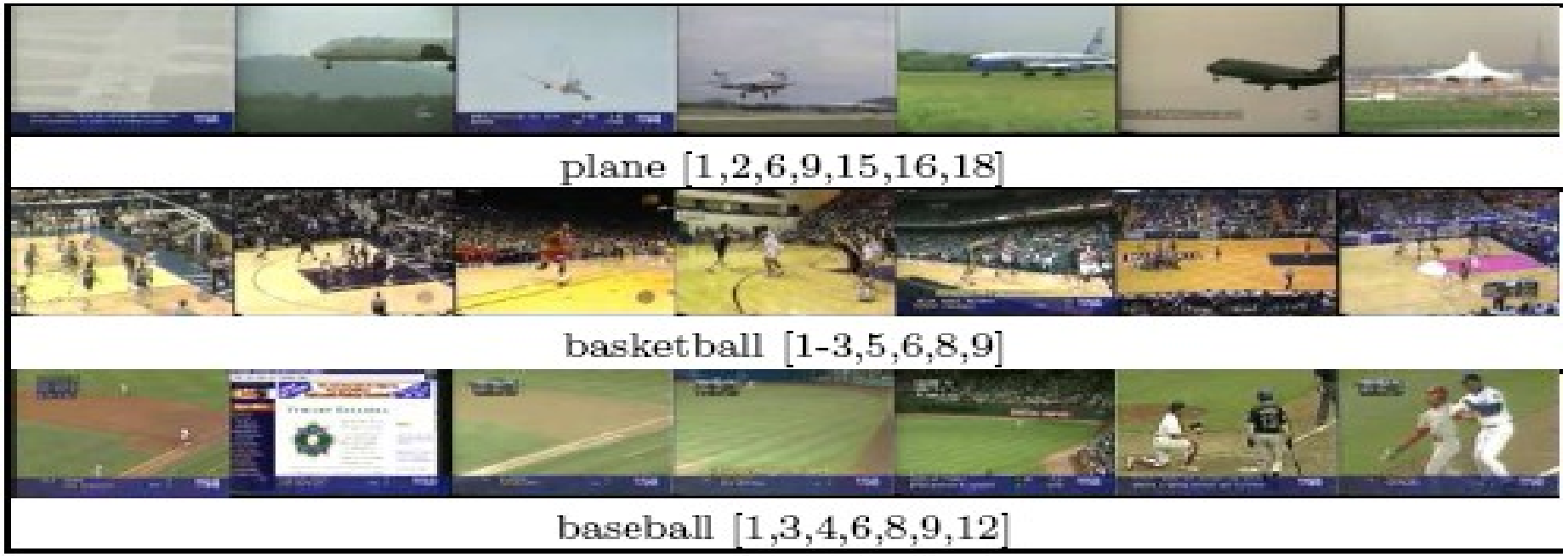
ASR : night game sery story

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

Story annotation performance

- For each news story, average annotation prediction performance 17%
- Average word recall 16%
- Average word precision 20%

Retrieval results using speech transcript text



Ranked query examples [numbers show the ranks]

Future Work

- novel approach to naming many faces by learning the correspondences between the names and faces
- using motion information inherent in video data
  - moving objects are usually more important
  - Use motion information to segment the moving objects
  - learn correspondences between motion information and motion verbs for naming actions and hence supply richer query capability