What's news what's not?
Associating news videos with words

Pinar Duygulu[1], Alex Hauptmann[2]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey
[2]Informedia Project, Carnegie Mellon University, Pittsburgh, PA, USA

duygulu@cs.bilkent.edu.tr, alex@cs.cmu.edu

http://www.cs.bilkent.edu.tr/~duygulu/Research/VideoAssociation.html

Problem

Retrieval of broadcast news videos based on text is unsatisfactory

A transcript word frequently does not directly `describe' the shot when it was spoken.

If we only look at the shots where a keyword was spoken, we find that the anchor/reporter might be introducing a story, while the following shots being relevant, but not current one

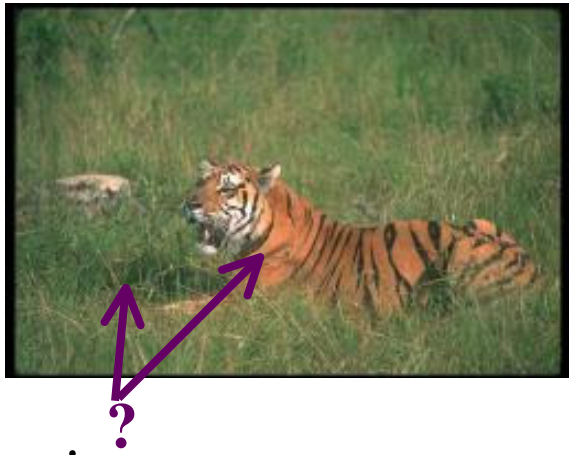
Query : "President Bush"

Arrows show the key-frames corresponding to shots including the query keyword

He appears only when his name is not mentioned

Association problem

This problem occurs due to missing correspondences between visual elements and textual information

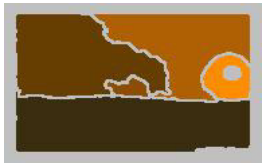


tiger grass cat

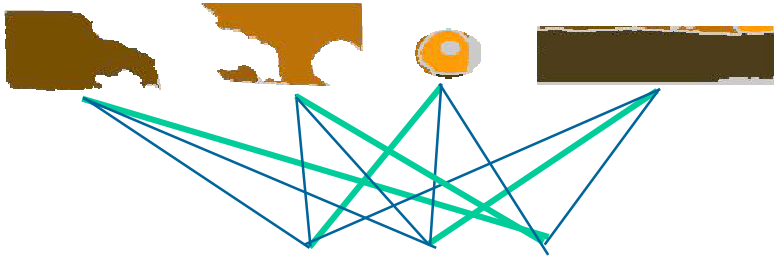
Annotated image collections:
Correspondence between image regions and keywords is unknown

News videos:
Correspondence between frames and audio transcripts is unknown.

Solution: Multimedia Translation

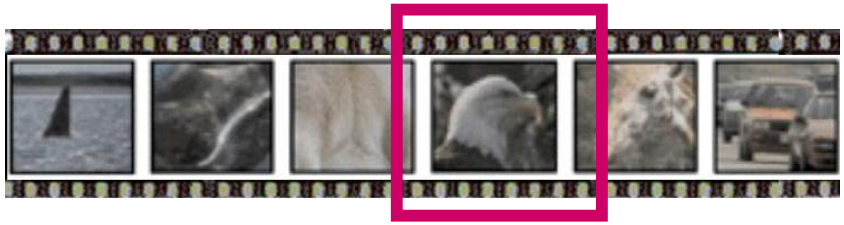


"sun sea sky"



"sun sea sky"

…despite heroic efforts many of the worlds wild creatures are doomed the loss of species is now the same as when the great dinosaurs become extinct will these creatures become the dinosaurs of our time today…



…efforts many | of the worlds | wild creatures | are doomed | the loss of | species …

This problem can be seen as translation of visual terms to text

Solution is adapted from statistical machine translation literature

Approach

Segment the the visual stream into coherent story-like units

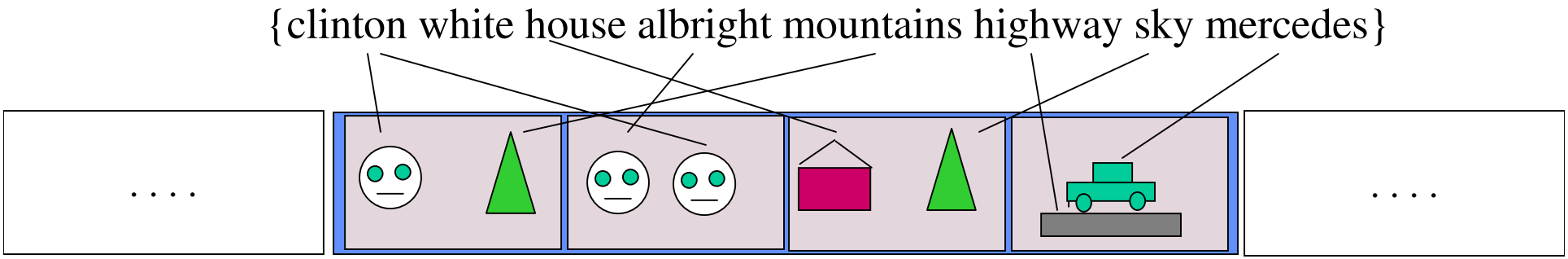
Relate the visual content in the story to the transcript words

In machine translation, sentences are the basic units
For multimedia translation, basic units will be the story segments

Extract word tokens and visual tokens from each story segment

Obtain the joint statistics of visual tokens and word tokens

Approach



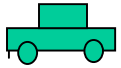
Visual tokens

 faces

 building

 road

 outdoor

 car

Word tokens

w1: fire , w2 : plane, w3: school, w4: clinton, w5: economy, ….

Story segmentation

anchor *anchor – reporter dialogs* *logos* *News story*

News story *weather* *commercials* *sports*

News videos are structured

Use the structure information to segment the videos into story-like units

Anchors, graphics and commercial frames are good cues for segmentation

Segmentation using delimiters

- Start a new story after a graphics or commercial frame.

- If there is a graphics or commercial in the next frame end the story.

- Start a new story with an anchor frame which follows a non-anchor frame.

- End a story with an anchor frame if the following frame is a non-anchor frame.

Graphics and commercials are hard boundaries for story segmentation, while anchors/reporters are soft boundaries and included into both preceding and following stories.

Segmentation results

A : anchor/reporter

G: graphics

C: commercial

N: news story

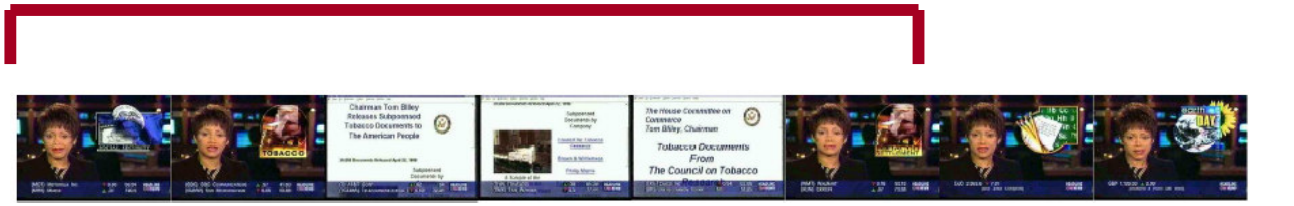
G N A N A N N N

G A A N A N N N

N G N N N G G C

Segmentation results

Soft boundary problem: icons on the top-right corners shows that the news are different, but they are put into same story



Anchor on the second frame is missed, therefore the stories are not segmented

However, these problems are not very harmful for association, since associating a visual element with the words in the neighbor segment will create only a slight noise

Anchor detection

- Multimodal classification

  -color histogram, face information, speaker similarity

- Fisher Linear Discriminant (FLD) to select features

Commercial detection

Combination of two methods

1) based on duplicate sequences

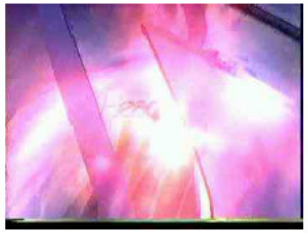
Commercials are repeated multiple times. Therefore, Sequences whose frames have duplicates has a high chance to be commercials

2) Based on distinctive color and audio features

Most commercials contain background music while news contains mostly speech. Colors are also distinguishable.

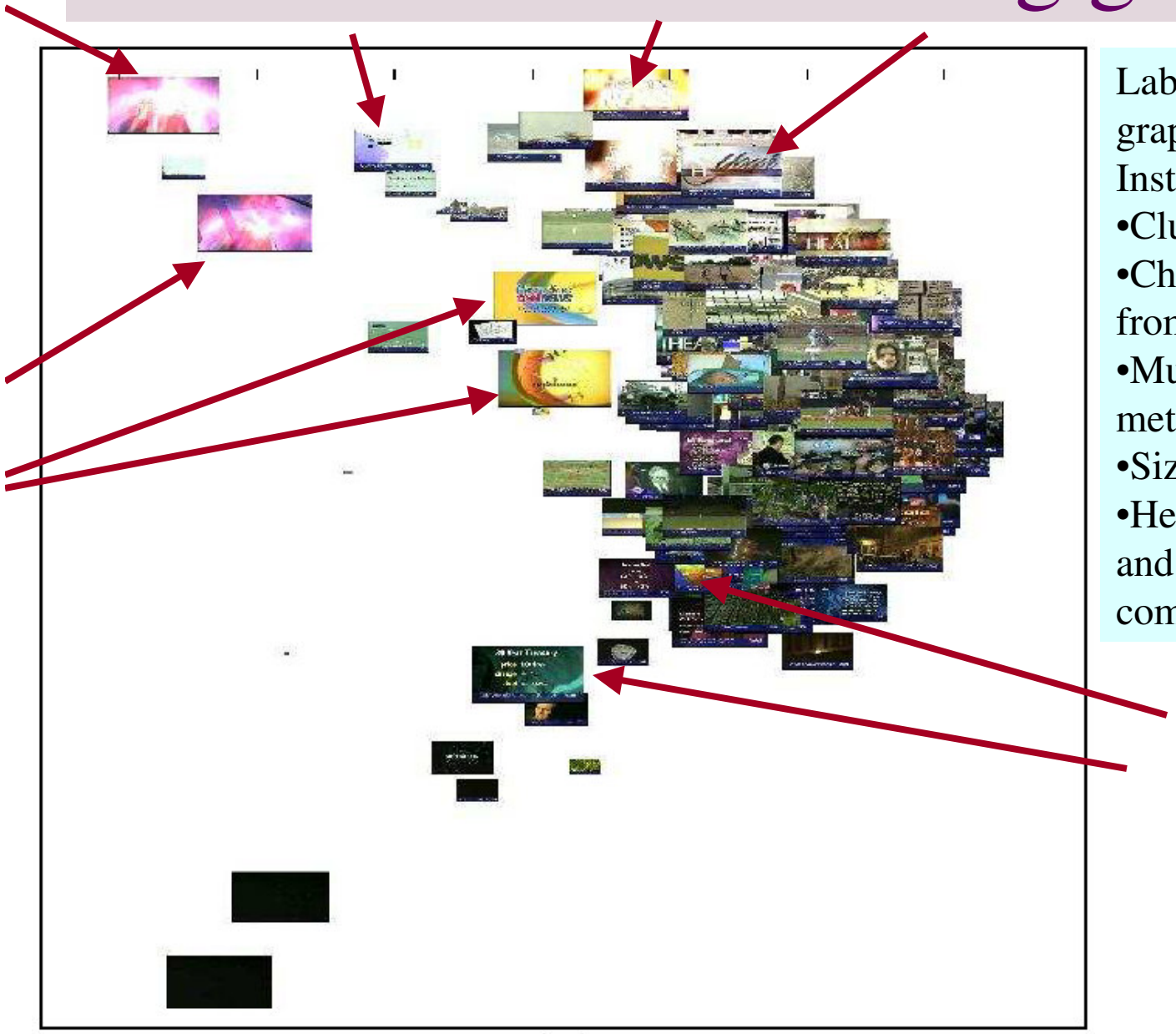
High level SVM is build to combine two strategies.

Graphics




graphics and logos are used separate one story from another, or as a starting shot for a particular news category (for example characteristic logos appear before sports, weather, health or financial news) or as corporate self-identification such as the “CNN headline news” logo.

Interface for detecting graphics



Labeling each different graphics is not feasible
Instead,

- Cluster – using color info
- Choose one representative from each cluster
- Multi dimensional scaling method for visualization
- Size ~ inv(variance)
- Helps to choose graphics and un-detected commercials and anchors

Detection results

On TRECVID 2003 CNN test data

Over 16650 shots

	anchors	commercials	graphics	in-studio
# elements	909	4347	1404	525
# correct	818 (90%)	4304 (99%)	1303 (93%)	456 (87%)

(e.g. 909 anchor frames are detected, and 90% of them are correct)

Data corresponding to news stories

After segmenting the data into story units, only the frames corresponding to news-stories are taken for a better association and retrieval

Commercial and graphics frames are removed from the data

Anchor and studio/reporter images are also deleted but the text corresponding to them is still used to find relevant stories.

Extracting tokens

Word tokens

Transcripts are aligned with shots by determining when each word was spoken

The vocabulary consists of only the nouns, originally there were 10201 words

Vocabulary is pruned to remove stop words: words occurring more than 10 times or less than 150 times are kept

→ 579 word tokens

Extracting tokens

Visual tokens

- Low level feature - Color

  the mean and variance of each color channel in HSV color space in a 5*5 image tessellation. (Hue was quantized into 16 bins. Both saturation and value were quantized into 6 bins)

  Feature vectors are clustered using G-means algorithm which determines the number of clusters adaptively

  Each cluster label is used as a visual token

  → 160 color tokens

Extracting tokens

Visual tokens (cont'd)

Mid-level features → classifier results

faces → 3 tokens

adapted from Schneiderman et.al.

according to number of faces (1 face, 2 face, 3 or more face)

Outdoor → 1 token

based on color, texture and edge features

Buildings → 1 token

adapted from Kumar and Hebert's man-made object detector

car → 1 token

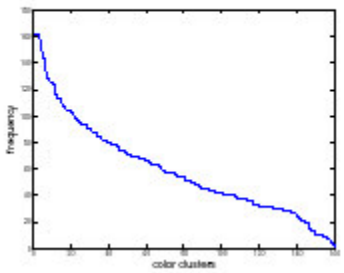
adapted from Schneiderman et.al.

Road → 1 token

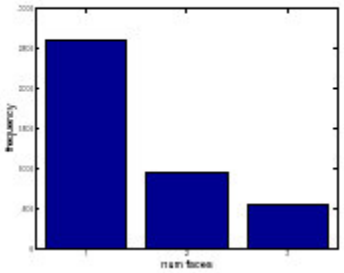
•based on color, texture and edge features

Classifier accuracy and distribution

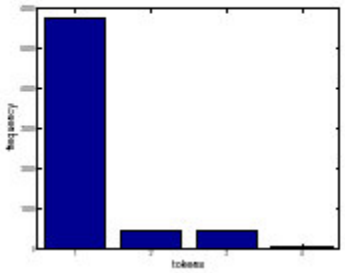
outdoor	building	car	road
1419 / 4179 (34%)	126 / 924 (14%)	26 / 78 (33%)	71 / 745 (9%)



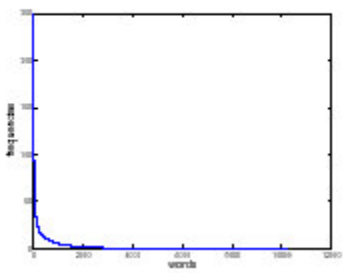
Color tokens



face



outdoor,
building,
road,
car



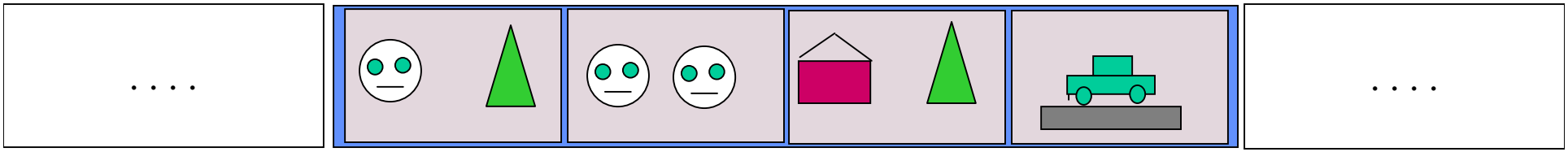
words

As it can be seen, the classifiers are errorful, and distributions are very different

Therefore, we are trying to build an association method on an incomplete and errorful data

Associating text with frames

w1 w2 w10 w1 w5 w6 w2 w1 w4 w10 w5 w3 w11



For each story segment

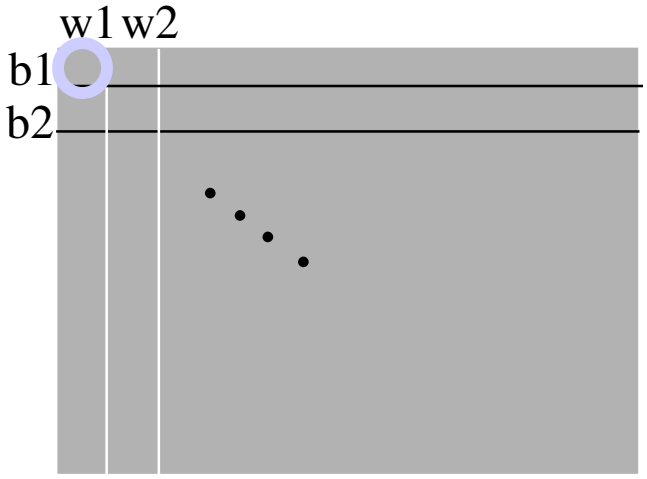
count the number of joint occurrences of visual tokens and word tokens to build a co-occurrence table

Associating text with frames

Initialize translation table with this co-occurrence table

Apply TF_IDF

Apply Expectation Maximization (EM) based algorithm (adapted from machine translation literature - Brown et.al.'s model 1) to obtain a better translation table



Use of translation table

Auto-annotation : Given a key-frame corresponding words can be predicted by finding the words associated to the visual token with high probability

Better retrieval : If only text is used, the shots aligned based on time information may not correspond to the query. The correct shot corresponding to the query word inside a story segment can be found by the associations obtained from the translation table. This will produce a better retrieval

Token words

First 20 words with the highest probability for some visual tokens
First three are color tokens, last one is building token

stock, wall, market, street, investor, report, news, business, jones, industrials, interest, deal, thanks, cnnfn, company, susan, yesterday, morris, number, merger

pilot, veteran, family, rescue, foot, effort, crew, search, security, troop, fact, affair, member, survivor, tobacco, field, department, health, communication, leader

series, bull, jazz, playoff, game, conference, final, karl, lead, indiana, utah, difference, combination, board, night, ball, point, pair, front, team

company, market, line, worker, street, union, profit, wall, cost, news, strike, yesterday, rate, quarter, stock, check, report, level, fact, board

Semantic retrieval

Query on “Clinton”

Time based alignment  20 / 130 (15%)

Proposed method  27 / 133 (20%)

Query on “fire”

Time based alignment  11 / 44 (25%)

Proposed method  15 / 38 (40%)

Note : only single occurrence per segment is taken with the proposed system

Summary

For a better retrieval, association of visual elements with text is necessary

Machine translation approaches can be adapted for multimedia translation by defining the story segments as basic units, and associating mid-level and low-level visual features to the transcript words extracted from the speech recognition

This method can be used for better retrieval by suggesting the right shots corresponding to the query word, and also for auto-annotation of video frames

Discussion & Future Work

Structure of the news is very important to get a good segmentation

However, text should be integrated for a better system

Low level and mid-level features can be combined for association

However, the classifiers should be better and more specific

Outdoor is very general – noisy

Although performance of car detector is very high,
number of examples is very low to learn the association

On the word side

instead of using only the nouns, topics can be used

Motion-tokens can be included and related to verbs