Linking Visual Features with Text for Multimedia Data Mining

Pinar Duygulu

Informedia Project, Carnegie Mellon University
Bilkent University, Turkey
Visual data with annotated text

Keywords: rose flower plant leaves
Textual Query

Query on “Rose”

Example from Berkeley Blobworld system
Visual Query

Example from Berkeley Blobworld system
Query using both text and visual features

Query on “Rose” and

Example from Berkeley Blobworld system
Combination of text with visual features

Appearance counts!

Semantics counts!
What can be done by combining text with features

- Information retrieval
- Browsing
- Auto illustration
- Auto annotation
- Multimedia translation

Annotated data sets

Pinar Duygulu, March 2004

Linking visual features with text for multimedia data mining
Annotation vs Recognition

Cannot be learned from a single image
Learning recognition from large data

Object recognition on large scale is linking image regions with words.
Use joint probability of words and images in large data sets.

Pinar Duygulu, March 2004
Statistical Machine Translation

Data: aligned sentences
But word correspondences are unknown

• Given the correspondences, we can estimate the translation \( p(\text{sun} \mid \text{soleil}) \)
• Given the probabilities, we can estimate the correspondences

Solution: enough data + EM

Brown et. al 1993
Multimedia Translation

Data:

Words are associated with the images
But correspondences between image regions and words are unknown

“sun sea sky”

Duygulu et.al, ECCV 2002
Pinar Duygulu, March 2004
Linking visual features with text for multimedia data mining
Each blob is a large vector of features:
- Region size
- Position
- Colour
- Oriented energy (12 filters)
- Simple shape features

k-means to cluster features
For each blob label of the Closest cluster \(\rightarrow\) blob tokens

sun sky waves sea

word tokens

Pinar Duygulu, March 2004

Linking visual features with text for multimedia data mining
Overview of the system

Training:
- Training images with annotated text
- Tokenization
- Expectation Maximization Algorithm
- Probability table

Testing:
- Test image
- Tokenization
- Word Prediction
- Auto annotation
- Region naming
Auto-Annotation
Region Naming
Results

- plane sky
- people ruins stone
- sunset tree water
Model Selection

- Clustering models
- Aspect models
- Hierarchical models
- Bayesian models
- Co-occurrence models

Many of these based on models proposed for text [Brown, Della Pietra, Della Pietra & Mercer 93; Hofmann 98; Hofmann & Puzicha 98]

A comparison paper is published in JMLR

‘Matching words and Pictures’, Barnard, Duygulu, Forsyth, Freitas, Blei, Jordan

Pinar Duygulu, March 2004

Linking visual features with text for multimedia data mining
## Other data sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel Image Data</td>
<td>40,000 images</td>
</tr>
<tr>
<td>Fine Arts Museum of San Francisco</td>
<td>83,000 images online</td>
</tr>
<tr>
<td>Cal-flora</td>
<td>20,000 images, species information</td>
</tr>
<tr>
<td>News photos with captions</td>
<td>1,500 images per day available from yahoo.com</td>
</tr>
<tr>
<td>(yahoo.com)</td>
<td></td>
</tr>
<tr>
<td>Hulton Archive</td>
<td>40,000,000 images (only 230,000 online)</td>
</tr>
<tr>
<td>internet.archive.org</td>
<td>1,000 movies with no copyright</td>
</tr>
<tr>
<td>TV news archives</td>
<td>Several terabytes already available</td>
</tr>
<tr>
<td>(televisionarchive.org, informedia.cs.cmu.edu)</td>
<td></td>
</tr>
<tr>
<td>Google Image Crawl</td>
<td>&gt;330,000,000 images (with nearby text)</td>
</tr>
<tr>
<td>Satellite images</td>
<td>(And associated demographic information)</td>
</tr>
<tr>
<td>(terrarserver.com, nasa.gov, usgs.gov)</td>
<td></td>
</tr>
<tr>
<td>Medical images</td>
<td>(And associated with clinical information)</td>
</tr>
</tbody>
</table>
FAMSF Data (83,000 images online)

Web number: 4359202410830012

rec number: 2
Description:
serving woman stands in a
dressing room, in front of vanity
with chair, mirror and mantle,
holding a tray with tea and toast

Title: Le Matin
Primary class: Print
Display date: 1886

Artist: Tissot
Country: France
“The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship ... “

Extracted Query

large importance attached fact old dutch century more command whale ship was person was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch ...
Linking visual features with text for multimedia data mining
Organizing Image Collections

Linking visual features with text for multimedia data mining

Pinar Duygulu, March 2004
Propose good features to differentiate words that are not distinguishable (e.g., eagle and jet)

On Corel data set color is the dominant feature
Low level segmenters split up objects and cannot group disparate regions belonging to semantic entities

Using word prediction gives a way to incorporate high-level semantic information in the merging process

Propose a merge between regions that have similar posterior distributions over words

Pinar Duygulu, March 2004
Sense Disambiguation

26078 water grass trees bank

125090 bank machine money currency bills

125084 piggy bank coins currency money

212001 bank buildings trees city

173044 mink rodent bank grass

151096 snow bank hills winter

Linking visual features with text for multimedia data mining
IDVL interface returned for "El Nino" query along with different multimedia abstractions from certain documents.
IDVL interface returned for “bin ladin” query

The results can be tuned using many classifiers
Associating video frames with text

Query on “president”

Association problem

Pinar Duygulu, March 2004

Linking visual features with text for multimedia data mining
…despite heroic efforts many of the world’s 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…
Associating video frames with text

Position,
Color
(RGB and Lab, mean and std)
Texture
(Oriented energy filters, DoG)

...efforts many of the worlds wild creatures are doomed the loss of species ...
Query for “Statue of Liberty” on current Informedia system

Auto-annotation results

- space (6), astronaut (7)
- plane (2)
- space (1), telescope (10)

Corrected with auto-annotation

- statue (1), liberty (3)
- statue (1), liberty (3)
Taking the surrounding words are problematic
Segments are defined in some close caption text
If it is not available use structure to obtain segments
Get only news stories

Remove commercials

Remove graphics

Remove anchor images but use text
Associating text with frames

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

Color tokens: 1-230

Num faces (1/2/ >=3)

building

road

outdoor

road
Token words

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

!! only single occurrence per segment

Search on clinton

Search on fire

20 / 130 (15%)

27 / 133 (20%)

11 / 44 (25%)

15 / 38 (40%)
Future work

Solving correspondences in broadcast news for better retrieval

..tanks on the street … ..start attacking on houses by helicopters and tanks… ..fuel tank…

Face Recognition by resolving correspondences between named entities and faces
When text and visual features are combined it is possible to do many interesting tasks
Including better retrieval, browsing, auto-annotation and auto-illustration

Object recognition on the very large scale can be viewed as translation of regions to words

There are many other available multi-modal data sets

Video is a huge source of information where audio, text, and visual features appear together

Current systems that are based on text should be improved with the help of multi-modal data