Associating video frames with text

Pinar Duygulu and Howard Wactlar

Informedia Project
School of Computer Science
Carnegie Mellon University
IDVL interface returned for "El Nino" query along with different multimedia abstractions from certain documents.
The IDVL interface returned for "bin ladin" query.

The results can be tuned using many classifiers.
Query on “president” : association problem

Obviously, this is all awkward for the President. Whether he'll like, or not, he can't come out on opposite sides of it. It's not realistic. At the same time, he's made it clear he doesn't want an independent commission; the woman commission of this tragedy, and today he made it clear that one investigation is enough, there's other work to be done. And, as it underscoring the point, he made the comments at the national security agency. We go back to the White House and our senior White House correspondent John King with the President side of things, Oh, Good evening. President visited agency, the chief executive agency for the intelligence community. Private pep talk for the employees there. Came out to speak to reporters. Clearly looking to put the stomp in the investigation. President saying for all the revelations, some would say lies, about what the government knew but did not share. Obviously, we're at the White House and our senior White House correspondent John King with the President side of things. Oh, Good evening. President visited agency, the chief executive agency for the intelligence community. Private pep talk for the employees there. Came out to speak to reporters. Clearly looking to put the stomp in the investigation. President saying for all the revelations, some would say lies, about what the government knew but did not share. Obviously, we're at the White House and our senior White House correspondent John King with the President side of things.
Approach

• Combine textual and visual features to understand how to link semantics with appearance

• Model joint statistics of visual features and words using a large collection of visual data with annotated text
Outline

• Correspondence problem between image regions and text using annotated image collections

• Correspondence problem between video frames and video

• Experiments on TREC 2001 data set

• Experiments on Chinese Cultural data set
Outline

• Correspondence problem between image regions and text using annotated image collections

• Correspondence problem between video frames and video

• Experiments on TREC 2001 data set

• Experiments on Chinese Cultural data set
Associating image regions with words

tiger  grass  cat

Duygulu, Barnard, de Freitas, Forsyth, ECCV2002
Associating image regions with words

Duygulu, Barnard, de Freitas, Forsyth, ECCV2002
Statistical machine translation

Data: Aligned sentences, but word correspondences are unknown

“the beautiful sun”

“le soleil beau”
Statistical machine translation

• Given the correspondences we can estimate $p(\text{sun}|\text{soleil})$

• Given the probabilities we can estimate the correspondences

“the beautiful sun”

“le soleil beau”

Enough data + EM, we can obtain the translation $p(\text{sun}|\text{soleil})=1$
Multimedia translation

“sun   sea   sky”
Input Representation

Each blob is a large vector of features:
- Region size
- Position
- Colour
- Oriented energy (12 filters)
- Simple shape features

sun sky waves sea
Tokenization

- Words $\rightarrow$ word tokens

- Image segments

  • represented by 40 features
  (size, position, color, texture and shape)
  k-means to cluster features
  • best cluster for the blob $\rightarrow$ blob tokens
Associations

$p(a_1 = 1) \xrightarrow{\text{p(a_1=2)}} p(a_1 = 3) \xrightarrow{\text{p(a_1=4)}} \text{“sun sea sky”}$

$B_n \sum_{i=1}^{n} p(a_1 = i) = 1$
Initialization

Initialize translation table to blob-word co-occurrences (empirical joint distribution of blobs and words)
Expectation Maximization Algorithm

Given the translation probabilities estimate the correspondences

Given the correspondences estimate the translation probabilities
EM Algorithm

**E step:** Predicting correspondences from translation probabilities
(for one pair)

---

**translation probabilities**

<table>
<thead>
<tr>
<th>b1</th>
<th>w1</th>
<th>w2</th>
</tr>
</thead>
<tbody>
<tr>
<td>b2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**correspondences**

\[
\begin{align*}
&b1 & b3 & b4 \\
&b2 & b1 & b5 \\
&b1 & b2 \\
&b1 & b2 & b5 \\
\end{align*}
\]

---

Carnegie Mellon
EM Algorithm

M step (for one pair)

Predicting translation probabilities from correspondences

correspondences

\[
\begin{array}{c}
\{ b_1 \ b_3 \ b_4 \} \\
| \\
\{ w_1 \ w_5 \} \\
| \\
\{ b_2 \ b_1 \ b_5 \} \\
| \\
\{ w_1 \ w_2 \ w_4 \} \\
| \\
\{ b_1 \ b_2 \} \\
| \\
\{ w_1 \ w_2 \ w_6 \} \\
\end{array}
\]

translation probabilities

\[
\begin{array}{c}
\{ w_1 \ w_2 \} \\
\{ b_1 \} \\
\{ b_2 \} \\
\end{array}
\]
Word prediction

On a new test image
• segment the image
• extract the features from the regions
• then, for each region
  find the corresponding blob token b
use word posterior probabilities $p(w|b)$ for predicting words

Use predicted words for
• region naming
• auto-annotation
Region naming
Results

plane sky

people ruins stone

sunset tree water
Auto-annotation
Using annotation performance as a proxy

Actual Keywords: GRASS  TIGER  CAT  FOREST

Predicted Words: CAT  HORSE  GRASS  WATER
Using annotation performance as a proxy

Actual Keywords:
- GRASS
- TIGER
- CAT
- FOREST

Predicted Words:
- CAT
- HORSE
- GRASS
- WATER
Outline

• Correspondence problem between image regions and text using annotated image collections

• Correspondence problem between video frames and video

• Experiments on TREC 2001 data set

• Experiments on Chinese Cultural data set
...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...
...efforts many of the worlds wild creatures are doomed the loss of species ...
Input

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 ...

Brill’s tagger is used to extract nouns

The text only corresponding to the shot can be used
Or also the surrounding text within a window size can be used
Outline

• Correspondence problem between image regions and text using annotated image collections

• Correspondence problem between video frames and video

• Experiments on TREC 2001 data set

• Experiments on Chinese Cultural data set
TREC 2001 data

Number of images = 2232

7 x 7 blocks
56 features extracted from each block
Number of blob tokens = 500

Number of words = 1938

Window size for surrounding text = 5
Auto-annotations

- Statue: 1
- Liberty: 2
- Plane: 2
- Space: 1
- Telescope: 10
- Robot: 5

Carnegie Mellon
Auto-annotations

- space (6), astronaut (7)
- plane (2)
- water (1), research (3)
- space (1), world (6)
Query for “Statue of Liberty”
Associating frames and text

statue(1) liberty(2)

statue(1) liberty(3)

statue(1) liberty(2)

statue(1) liberty(3)
Outline

• Correspondence problem between image regions and text using annotated image collections

• Correspondence problem between video frames and video

• Experiments on TREC 2001 data set

• Experiments on Chinese Cultural data set
Query on “panda”
Query on “Great Wall”
Chinese Culture data set

Number of images = 3745

5 x 5 blocks
56 features extracted from each block
Number of blob tokens = 1000

Number of words = 2597
Pruning data

Eliminate the words if frequency $\leq 5$ or frequency $> 250$

Number of words = 626
Number of images = 2785
Recall and Precision

When only the words that are associated with the frame are taken

Recall: number of correct predictions / number of actual occurrence

Precision: number of correct predictions / number of all predictions
Panda (when predicted in the first 3)
Wall (when predicted in the first 3)
Emperor (when predicted in the first 3)
## Correspondence results for “panda”

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of annotations</td>
<td>127</td>
</tr>
<tr>
<td>number of correct annotations</td>
<td>42</td>
</tr>
<tr>
<td>number of incorrect annotations</td>
<td>85</td>
</tr>
<tr>
<td>number of predictions</td>
<td>121</td>
</tr>
<tr>
<td>number of correct predictions</td>
<td>51</td>
</tr>
<tr>
<td>number of incorrect predictions</td>
<td>70</td>
</tr>
<tr>
<td>annotation present not predicted</td>
<td>64</td>
</tr>
<tr>
<td>annotation correct not predicted</td>
<td>11</td>
</tr>
<tr>
<td>annotation incorrect not predicted</td>
<td>53</td>
</tr>
<tr>
<td>not annotated but predicted</td>
<td>61</td>
</tr>
<tr>
<td>not annotated but correctly predicted</td>
<td>25</td>
</tr>
<tr>
<td>not annotated and incorrectly predicted</td>
<td>36</td>
</tr>
<tr>
<td>annotated and predicted and correct</td>
<td>26</td>
</tr>
<tr>
<td>annotated and predicted but incorrect</td>
<td>31</td>
</tr>
</tbody>
</table>
## Correspondence results for “panda”

<table>
<thead>
<tr>
<th>Image</th>
<th>Correspondence Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image 1]</td>
<td>3</td>
</tr>
<tr>
<td>![Image 2]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 3]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 4]</td>
<td>13</td>
</tr>
<tr>
<td>![Image 5]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 6]</td>
<td>17</td>
</tr>
<tr>
<td>![Image 7]</td>
<td>18</td>
</tr>
<tr>
<td>![Image 8]</td>
<td>111</td>
</tr>
<tr>
<td>![Image 9]</td>
<td>6</td>
</tr>
<tr>
<td>![Image 10]</td>
<td>5</td>
</tr>
<tr>
<td>![Image 11]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 12]</td>
<td>13</td>
</tr>
<tr>
<td>![Image 13]</td>
<td>127</td>
</tr>
<tr>
<td>![Image 14]</td>
<td>4</td>
</tr>
<tr>
<td>![Image 15]</td>
<td>10</td>
</tr>
<tr>
<td>![Image 16]</td>
<td>4</td>
</tr>
<tr>
<td>![Image 17]</td>
<td>82</td>
</tr>
<tr>
<td>![Image 18]</td>
<td>620</td>
</tr>
<tr>
<td>![Image 19]</td>
<td>2</td>
</tr>
<tr>
<td>![Image 20]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 21]</td>
<td>22</td>
</tr>
<tr>
<td>![Image 22]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 23]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 24]</td>
<td>16</td>
</tr>
<tr>
<td>![Image 25]</td>
<td>8</td>
</tr>
<tr>
<td>![Image 26]</td>
<td>4</td>
</tr>
<tr>
<td>![Image 27]</td>
<td>89</td>
</tr>
<tr>
<td>![Image 28]</td>
<td>12</td>
</tr>
<tr>
<td>![Image 29]</td>
<td>620</td>
</tr>
<tr>
<td>![Image 30]</td>
<td>1</td>
</tr>
<tr>
<td>![Image 31]</td>
<td>3</td>
</tr>
</tbody>
</table>
Recall and precision as a function of window size

Single frame

Window size = 1

Window size = 2

Window size = 3
## Recall and precision for some selected words as a function of window size

<table>
<thead>
<tr>
<th></th>
<th>panda</th>
<th>wall</th>
<th>emperor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>single frame</strong></td>
<td>0.7183 - 0.2044</td>
<td>0.8492 - 0.2281</td>
<td>0.6885 - 0.2242</td>
</tr>
<tr>
<td>wsize = 1</td>
<td>0.8182 - 0.2432</td>
<td>0.9180 - 0.2923</td>
<td>0.8744 - 0.2860</td>
</tr>
<tr>
<td>wsize = 2</td>
<td>0.8857 - 0.2560</td>
<td>0.9517 - 0.3149</td>
<td>0.9446 - 0.3257</td>
</tr>
<tr>
<td>wsize = 3</td>
<td>0.9154 - 0.2604</td>
<td>0.9720 - 0.3186</td>
<td>0.9693 - 0.3631</td>
</tr>
</tbody>
</table>
Conclusions

While text and images are separately ambiguous, jointly they tend not to be.

Linking visual information with text improves performance

The proposed method
• Can learn correspondence between visual data and text
• Unsupervised – uses the available large data sets efficiently

Can be used
• For region naming – object recognition on the large scale
• Auto-annotation – predicting words
• Associating frames and text
Future Directions

Applying on broadcast news

A better set of features including face detectors, and moving objects

Better linguistic analysis (e.g. taking noun phrases)

Finding the best window size by statistical analysis
Thank you

http://www.informedia.cs.cmu.edu/
# Prediction measure as a function of window size

<table>
<thead>
<tr>
<th>wsize</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>single frame</td>
<td>0.1851</td>
</tr>
<tr>
<td>wsize = 1</td>
<td>0.2469</td>
</tr>
<tr>
<td>wsize = 2</td>
<td>0.2783</td>
</tr>
<tr>
<td>wsize = 3</td>
<td>0.2975</td>
</tr>
</tbody>
</table>

Prediction measure (PR):

\[
\frac{1}{N} \sum_{N} \left( \frac{\text{#correct}}{\text{# actual}} \right)
\]