Data mining for large scale image and video analysis

Pinar Duygulu
Large volumes of video

- For YouTube alone
  - More than 1 billion unique user visits each month
  - Over 6 billion hours of video are watched each month
  - 100 hours of video are uploaded every minute


Pinar Duygulu, November 2016, Ankara
Applications

- Analyzing video archives
  - First appearance of N. Sarkozy on TV
  - Sociology research: Influence of character smoking in movies
  - Education: How do I make a pizza?

- Surveillance
  - Where is my cat?
  - Predicting crowd behavior
  - Counting people

- Graphics
  - Motion capture and animation

Pinar Duygulu, November 2016, Ankara
# Available Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH</td>
<td>6</td>
</tr>
<tr>
<td>Weizmann</td>
<td>9</td>
</tr>
<tr>
<td>IXMAS</td>
<td>11</td>
</tr>
<tr>
<td>Hollywood</td>
<td>8</td>
</tr>
<tr>
<td>UCF Sports</td>
<td>9</td>
</tr>
<tr>
<td>Hollywood2</td>
<td>12</td>
</tr>
<tr>
<td>UCF YouTube</td>
<td>11</td>
</tr>
<tr>
<td>MSR</td>
<td>3</td>
</tr>
<tr>
<td>Olympic</td>
<td>16</td>
</tr>
<tr>
<td>UCF50</td>
<td>50</td>
</tr>
<tr>
<td>HMDB51</td>
<td>51</td>
</tr>
</tbody>
</table>

Yahoo! Recently released 100 million Flick data

http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

Pinar Duygulu, November 2016, Ankara
**Recent datasets**

<table>
<thead>
<tr>
<th>Veri Kümesi</th>
<th>#Sınıf</th>
<th>#Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF-101</td>
<td>101</td>
<td>13320</td>
</tr>
<tr>
<td>ActivityNet</td>
<td>200</td>
<td>20000</td>
</tr>
<tr>
<td>FCVID</td>
<td>239</td>
<td>91223</td>
</tr>
<tr>
<td>Sports-1M</td>
<td>487</td>
<td>1M</td>
</tr>
</tbody>
</table>

Yahoo! Flickr 100M dataset
418.507 labeled video

Pinar Duygulu, November 2016, Ankara
Videos in the wild

- Unrestricted type of events with various activities

Harlem Shake: [http://www.youtube.com/watch?v=4hpEnLqUDg](http://www.youtube.com/watch?v=4hpEnLqUDg)

Pinar Duygulu, November 2016, Ankara
Our attempts

- Videos as sequence of frames
  - Detect concepts in each frame
  - Utilize image search engines

- Discover important knowledge from videos itself
  - Discriminate parts

- Understand actions in videos
  - Simple but effective descriptors

Pinar Duygulu, November 2016, Ankara
Utilizing large volumes of weakly labeled images
Utilize image search results

Query : Ankara
Single Dominant Category

Query: Turing
Among the faces associated with a name find the correct subset:
The most similar subset of faces

Ozkan, D., Duygulu, P., "Interesting Faces: A Graph Based Approach for Finding People in News", Pattern Recognition, 2010
Finding Densest component

Node with the minimal degree is removed at each iteration (Charikar, 2000)
Image Re-ranking


Pinar Duygulu, November 2016, Ankara
Multiple meanings/variations

The concepts are observed in different forms requiring grouping and irrelevant elements to be eliminated.

Pinar Duygulu, November 2016, Ankara
CMAP for Concept Learning

Golge, E., Duygulu, P., “Concept Maps: Mining Noisy Web Data for Concept Learning”, accepted to ECCV 2014
Color and Texture Attributes

Salient Clusters

Outlier Elements

Outlier Clusters

Brown

Vegetation

Pinar Duygulu, November 2016, Ankara
Scene Concepts

Outdoor

Bedroom

Pinar Duygulu, November 2016, Ankara
Attribute and Scene Learning

Attribute learning for object recognition  Attribute based scene recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>RSOM-M</th>
<th>RSOM</th>
<th>PLSA-reg [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td>cars</td>
<td>0.97</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>shoes</td>
<td>1.0</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>dresses</td>
<td>1.0</td>
<td>1.0</td>
<td>0.99</td>
</tr>
<tr>
<td>pottery</td>
<td>0.98</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>overall</td>
<td>0.99</td>
<td>0.95</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RSOM-A</td>
<td>46.2%</td>
<td>82.7%</td>
</tr>
<tr>
<td>RSOM-S</td>
<td>-</td>
<td>80.7%</td>
</tr>
<tr>
<td>RSOM-S+HM</td>
<td>-</td>
<td>81.3%</td>
</tr>
<tr>
<td>Li et al. [12] VQ</td>
<td>47.6%</td>
<td>82.1%</td>
</tr>
<tr>
<td>Pandey et al. [16]</td>
<td>43.1%</td>
<td>-</td>
</tr>
<tr>
<td>Kwitt et al. [9]</td>
<td>44%</td>
<td>82.3%</td>
</tr>
</tbody>
</table>

On ImageNet: 37.4% (RSOM), 36.8% (Russakovsky & Fei-Fei, 2012)

Learning scene concepts directly

Comparison with other clustering methods

---

**Legend:**
- CMAP
- SOM
- k-means
- MeanShift
- DBSCAN
- Baseline

**Axes:**
- Y-axis: Value range from 0.2 to 0.8
- X-axis: Bing, Google, ImageNet, EBAY
Faces

Pinar Duygulu, November 2016, Ankara
FAME: Face Association Through Model Evolution
Capture unusualness in unusual videos

Pinar Duygulu, November 2016, Ankara
Boiman and Irani
ICCV 2005

Roshtkhari and Levine, CVPR 2013

Ito, Kitani, Bagnell, Hebert, 2012

Zhao, Fei-Fei, Xing, CVPR 2011
Usual versus unusual

Pinar Duygulu, November 2016, Ankara
Usual versus unusual
Rapid motions

Fast speed
Large spatial extension
Trajectory Snippet Histograms
Velocity and spatial extension of the motion

\[
H_S^l = \sum_{t=s-(\|S\|/2)}^{s+(\|S\|/2)} H^l_S(t) \quad H_S = (H^l_S, H^x_S, H^y_S)
\]

\[
H^l_S(t) = (H^l_S(t)_{[1,1]}, \ldots H^l_S(t)_{[1,N]}, \ldots H^l_S(t)_{[N,N]})
\]

\[
T^* = (P_t, \ldots, P_{t+D-1}) \quad P_t^* = (x_t, y_t)
\]

\[
m_x = \frac{1}{D} \sum_{t}^{t+D-1} x_t, \quad v_x = \frac{1}{D} \sum_{t}^{t+D-1} (x_t - m_x)^2
\]

\[
m_y = \frac{1}{D} \sum_{t}^{t+D-1} y_t, \quad v_y = \frac{1}{D} \sum_{t}^{t+D-1} (y_t - m_y)^2,
\]

\[
l = \sum_{t}^{t+D-1} \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2}
\]
Classification

Best SH: people falling 75%, funny videos 76.25%
HOG3D: people falling 65%, funny videos 73.75%
Failure cases
Limitations
Finding discriminative parts in videos

Pinar Duygulu, November 2016, Ankara
Weakly labeled videos

Pinar Duygulu, November 2016, Ankara
Birthday event

Mum üfleme

November 2016, Ankara
Finding discriminative parts

Singh ECCV 2012

Unsupervised Discovery of Mid-Level Discriminative Patches

Our Discriminative Patches

Visual Words

Jain CVPR 2013

Pinar Duygulu, November 2016, Ankara
Dense sampling in each spatial scale

Tracking in each spatial scale separately

Trajectory description

Input Video

CNN FC-6
Layer Features

Classification

AlexNet [8,9,10]

3D ConvNets [11,12]
Example of successful eliminated instances by AME\cite{1} for ActivityNet action classes. (1) "Archery" class. Baseline: 31.57\%, AME: 44.73\%. (2) "Checking tires" class. Baseline: 26.82\%, AME: 41.46\%. (3) "Platform diving" class. Baseline: 56.66\%, AME: 73.33\%
Example of unsuccessful eliminated instances for ActivityNet class "Windsurfing" with AME$^{[1]}$. Baseline: 74.07%, AME$^{[1]}$: 66.66%
Prototypes
Human Activity Analysis

Pinar Duygulu, November 2016, Ankara
What do these people do?

- Pose tells a lot about the actions.
- How can we describe the pose?

running  
walking  
throwing  
crouching
Pose as a Collection of Rectangles

- Human body is composed of cylindrical parts.
- The projection of a cylinder on 2D is a rectangle.
- Body can be thought as a collection of rectangular regions
- We can represent the pose based on the orientation of these rectangles
Rectangular regions are extracted over silhouettes using convolution of a zero-padded rectangular 2D Gaussian on different orientations and scales.

- 12 angles 15° apart
Use snippets of frames and form histogram of oriented rectangles over a window (HORW)
Action Recognition in Still Images

- Pose estimation by Ramanan’s method
- Form Circular HORs (CHORs)
- Classification based on LDA+SVM

Still Image Results

ActionWeb dataset - 467 images collected from the web

Correctly classified action images

- running
- walking
- throwing
- catching
- crouching
- kicking
Boundary-fitted Lines

- In the absence of silhouettes, we can use lines fitted to the boundaries (Pb) (Martin PAMI 2004) of human figures.
Dense block-based optical flow calculation

- L₁ block distance
- 5x5 template size with a window size of 3
Recognition with LHist and OFHist

Pose as line segments

Line pairs

(a) pose extraction

(b) projection of edge_img to extract bounding box

(c) noise elimination on cen_img

Pinar Duygulu, November 2016, Ankara
Multiple camera views
Oriented cylinders

Pehlivan, S., Duygulu, P. "3D Human Pose Search using Oriented Cylinders", IEEE Workshop on Search in 3D and Video (S3DV), in conjunction with ICCV 2009
Projections as circles

Pehlivan, S., Duygulu, P., "A new pose-based representation for recognizing actions from multiple cameras", Computer Vision and Image Understanding, volume 115, number 2, pages 140-151, February 2011
Assistive systems for Patient and Elderly care

Pinar Duygulu, November 2016, Ankara
Cooking Activities:
High Intra-class Variance

Iscen, A., Armagan, A., Duygulu, P., "Knives are picked before slices are cut: Recognition through Activity Analysis", Workshop on Cooking and Eating Activities, in conjunction with ACM Multimedia 2013.
Low Inter-class Variance

Cut apart, cut ends. cut slices, cut stripes, cut dice
Solution

\[ y = \underset{i}{\text{argmax}} \ P(c_i | x) \]

\[ P(c_i | x) = T(c_i) \cdot A(c_i, x) \]
Put in Pan or Put in Bowl?

\[ P(\text{“put in bowl”} \mid \text{“cut dice”}) > P(\text{“put in pan”} \mid \text{“cut dice”}) \]

\[ P(\text{“put in pan”} \mid \text{“spread”}) > P(\text{“put in bowl”} \mid \text{“spread”}) \]
Medical Device Use

**Asthma Inhaler**

- Breathe out slowly
- Hold your breath for 10 seconds
- Breathe in and push down the button at the same time
- Put the inhaler about 2 inches in front of your mouth
- Shake the inhaler (for 5 seconds)

---

<table>
<thead>
<tr>
<th></th>
<th>Trajectory</th>
<th>HOG</th>
<th>HOF</th>
<th>MBH</th>
<th>Snippet Hist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>95.31</td>
<td>50.00</td>
<td><strong>100.00</strong></td>
<td>87.50</td>
<td>98.44</td>
</tr>
<tr>
<td>Precision</td>
<td>91.04</td>
<td>22.70</td>
<td>91.43</td>
<td>71.79</td>
<td><strong>100.00</strong></td>
</tr>
<tr>
<td>F-score</td>
<td>93.13</td>
<td>31.22</td>
<td>95.52</td>
<td>78.87</td>
<td><strong>99.21</strong></td>
</tr>
</tbody>
</table>

Pinar Duygulu, November 2016, Ankara
### Infusion Pump

(a) front  
(b) side  
(c) above

<table>
<thead>
<tr>
<th>Actions</th>
<th>Trajectory</th>
<th>HOG</th>
<th>HOF</th>
<th>MBH</th>
<th>Snippet Hist</th>
<th>ROI-BoW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turn the pump on/off</td>
<td>91.52</td>
<td>91.52</td>
<td>90.83</td>
<td>92.39</td>
<td>97.23</td>
<td>89.40</td>
</tr>
<tr>
<td>Press buttons</td>
<td>79.93</td>
<td>80.28</td>
<td>80.10</td>
<td>79.76</td>
<td>83.91</td>
<td>88.33</td>
</tr>
<tr>
<td>Uncap tube end/arm port</td>
<td>84.26</td>
<td>85.64</td>
<td>83.56</td>
<td>85.47</td>
<td>91.35</td>
<td>65.41</td>
</tr>
<tr>
<td>Cap tube end/arm port</td>
<td>84.26</td>
<td>83.91</td>
<td>83.91</td>
<td>84.26</td>
<td>89.45</td>
<td>44.55</td>
</tr>
<tr>
<td>Clean tube end/arm port</td>
<td>70.24</td>
<td>73.18</td>
<td>77.51</td>
<td>74.05</td>
<td>75.78</td>
<td>92.02</td>
</tr>
<tr>
<td>Flush using syringe</td>
<td>88.75</td>
<td>88.24</td>
<td>88.06</td>
<td>87.20</td>
<td>92.56</td>
<td>94.80</td>
</tr>
<tr>
<td>Connect/disconnect</td>
<td>90.14</td>
<td>90.31</td>
<td>88.24</td>
<td>90.14</td>
<td>92.73</td>
<td>53.35</td>
</tr>
<tr>
<td>Average</td>
<td>84.16</td>
<td>84.73</td>
<td>84.60</td>
<td>84.75</td>
<td>89.00</td>
<td>75.41</td>
</tr>
</tbody>
</table>
Contributors

- Ahmet Iscen
- Eren Golge
- Anil Armagan
- Sermetcan Baysal
- Fadime Sener
- Hilal Zitouni
- Sare Gul Sevil
- Selen Pehlivan
- Gokberk Cinbis
- Derya Ozkan

Nazli Ikizler
Pinar Duygulu, November 2016, Ankara