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, January 13, 2014, CMU
<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>


Pinar Duygulu, January 13, 2014, CMU
Videos in the wild

- Unrestricted type of events with various activities

Harlem Shake: [http://www.youtube.com/watch?v=4hpEnLtqUDg](http://www.youtube.com/watch?v=4hpEnLtqUDg)
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
Utilizing large volumes of weakly labeled images
Utilize image search results

Query: Paris
Single Dominant Category

Query: George W. Bush
Naming faces

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

Multiple Instance Learning for re-ranking

On the dataset by Schroff, F., ICCV 2007 “Harvesting Image Databases from the Web”.

Sener, F., Ikizler-Cinbis, N., Duygulu, P., “Multiple Instance Learning for re-ranking of Web image search results”, SIU 2012
the attributes are observed in different forms and in small portions requiring grouping and non-attribute parts to be eliminated.
RSOM for Concept Learning

- Collect images from web for a keyword
- Clustering and outlier detection
- RSOM (Rectifying Self Organizing Maps)
- Learn a model for each cluster

Color and Texture Attributes

Salient Clusters

Outlier Elements

Outlier Clusters

Brown

Vegetation

Pinar Duygulu, January 13, 2014, CMU
# Scene Concepts

**Outdoor**

- Image 1
- Image 2
- Image 3
- Image 4

**Bedroom**

- Image 5
- Image 6
- Image 7
- Image 8

---

Pinar Duygulu, January 13, 2014, CMU
Attribute and Scene Learning

Attribute learning

<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>

Attribute based scene recognition

<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

Utilizing videos
Movie Genre & MPAA Ratings

300 Spartans – Movie Trailer (represented as frames)

Statistical features:
- Key frame count, Black and White frame counts, Avg. in shot frame difference, Total number of frames, Avg. shot length, Avg. key frame HSV color histogram, Avg. key frame edge histogram

Visual features:
- SIFT, LAB histogram, Local Binary Patterns, GIST

Genre: action, horror, animation, comedy, drama
Video Analysis for Sports

- Track soccer players in real-time in all matches
  - Extract XY coordinates and event data

- Using the big data
  - Player/team performance analysis/comparison
  - Extract game features that is correlated with seasonal success
  - Performance indexing and market value evaluation
  - Fatigue and injury prediction
Utilize video data: Capture usualness in unusual videos
Usual versus unusual
Trajectory Snippet Histograms

Representation

Velocity and spatial extension of the motion

\[ H_S^l(t) = \sum_{t=s-(\|S\|/2)}^{s+(\|S\|/2)} H_S^l(t) \]

\[ H_S = (H_S^l, H_S^x, H_S^y) \]

\[ T = (P_t, ..., P_{t+D-1}) \]

\[ P_t = (x_t, y_t) \]

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

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

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

- Discriminative video patch idea over snippets (short video sequences)

Singh ECCV 2012

Jain CVPR 2013
HOG3D --- people falling 25 %, funny videos 32 %
Birthday event

Flowing candles
Human Activity Analysis
What do these people do?

- Pose tells a lot about the actions.
- How can we describe the pose?
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 based on CRFs.
- 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 PAMI2004) of human figures.
..and Optical Flow

- Dense block-based optical flow calculation
  - $L_1$ 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 csp_img

\[ f_1 \]

\[ f_2 \]
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
Cooking Activities: High Intra-class Variance
Low Inter-class Variance
Solution

\[ y = \arg\max_i P(c_i | x) \]

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

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

\[ P(\text{“put in pan” | “spread”}) > P(\text{“put in bowl” | “spread”}) \]
Asthma Inhaler use

Correct steps to use an inhaler:

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

shaking
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, January 13, 2014, CMU
TagSuggestr

- Given a few initial tags, predict more.
- Give more weights to the visually similar images.

\[
W = \left\{ I_1, I_2, I_3, \ldots, I_m \right\} \times \left[ \begin{array}{c} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{array} \right]
\]

Experimental Evaluation of HOR

Comparison to other methods on the Weizzman dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOR</td>
<td>100%</td>
</tr>
<tr>
<td>Blank et al. [12]</td>
<td>99.64%</td>
</tr>
<tr>
<td>Jhuang et al. [48]</td>
<td>98.8%</td>
</tr>
<tr>
<td>Wang et al. [96]</td>
<td>97.78%</td>
</tr>
<tr>
<td>Niebles et al. [63]</td>
<td>72.8%</td>
</tr>
</tbody>
</table>

Comparison to other methods on the KTH dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jhuang et al. [48]</td>
<td>91.7%</td>
</tr>
<tr>
<td>Wong et al. [100]</td>
<td>91.6%</td>
</tr>
<tr>
<td>HORW</td>
<td>89.4%</td>
</tr>
<tr>
<td>Niebles et al. [64]</td>
<td>81.5%</td>
</tr>
<tr>
<td>Dollár et al. [24]</td>
<td>81.2%</td>
</tr>
<tr>
<td>Ke et al. [50]</td>
<td>80.9%</td>
</tr>
<tr>
<td>Schuldt et al. [84]</td>
<td>71.7%</td>
</tr>
</tbody>
</table>

Comparison to HOGs on the KTH

<table>
<thead>
<tr>
<th></th>
<th>HOG</th>
<th>HOR</th>
<th>HORW</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>76.85%</td>
<td>77.31%</td>
<td>85.65%</td>
</tr>
<tr>
<td>DTW</td>
<td>67.59%</td>
<td>74.54%</td>
<td>78.24%</td>
</tr>
<tr>
<td>v+SVM</td>
<td>82.41%</td>
<td>81.48%</td>
<td>89.35%</td>
</tr>
</tbody>
</table>
Line and Flow Results

- Shape and flow are complimentary to each other.
- Again, this depends on the nature of the actions in mention.
Still Image Results

<table>
<thead>
<tr>
<th>Action</th>
<th>Running</th>
<th>Walking</th>
<th>Throwing</th>
<th>Catching</th>
<th>Crouching</th>
<th>Kicking</th>
</tr>
</thead>
<tbody>
<tr>
<td>running</td>
<td>0.83</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.0</td>
</tr>
<tr>
<td>walking</td>
<td>0.04</td>
<td>0.94</td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>throwing</td>
<td>0.0</td>
<td>0.07</td>
<td>0.85</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>catching</td>
<td>0.15</td>
<td>0.04</td>
<td>0.04</td>
<td>0.72</td>
<td>0.0</td>
<td>0.06</td>
</tr>
<tr>
<td>crouching</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.89</td>
<td>0.01</td>
</tr>
<tr>
<td>kicking</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.0</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Total accuracy 85.1%

Misclassified action images
Multi-view

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (over 11 actions)</th>
<th>Accuracy (over 13 actions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weinland et al. [61]</td>
<td>93.33%</td>
<td>-</td>
</tr>
<tr>
<td>Our method</td>
<td>90.91%</td>
<td>88.63%</td>
</tr>
<tr>
<td>Liu et al. [32]</td>
<td>-</td>
<td>82.8%</td>
</tr>
<tr>
<td>Weinland et al. [59]</td>
<td>81.27%</td>
<td>-</td>
</tr>
<tr>
<td>Lv et al. [35]</td>
<td>-</td>
<td>80.6%</td>
</tr>
<tr>
<td>Yan et al. [62]</td>
<td>78.0%</td>
<td>-</td>
</tr>
</tbody>
</table>