Semantic Spaces for Zero-Shot Behaviour Analysis

Xun Xu

Computer Vision and Interactive Media Lab, NUS Singapore
Collaborators

Prof. Shaogang Gong

Dr. Timothy Hospedales
Outline

• Background

• Transductive Zero-Shot Action Recognition

• Multi-Task Zero-Shot Embedding

• Zero-Shot Crowd Analysis
Video Behaviour

Defined as Visually Distinguishable Activities

• Human Actions

• Crowd Behaviour
Human Actions

- Individual or multiple interactive human activities

Human Actions Tasks

• Action Recognition

Eye Makeup  Rafting  Swimming
Diving  Archery  Fencing
Human Actions Tasks

• **Action Detection (Retrieval)**
  Given query “Swimming” return ranked videos
Crowd Behaviour

- A group of people acting collectively

Crowd Behaviour Tasks

- Crowd Behaviour Profiling

<table>
<thead>
<tr>
<th>pedestrian</th>
<th>street</th>
<th>outdoor</th>
<th>walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stand</th>
<th>indoor</th>
<th>performance</th>
<th>sit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.96</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>photograph</th>
<th>outdoor</th>
<th>photographer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td>0.95</td>
<td>0.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>walk</th>
<th>stand</th>
<th>pedestrian</th>
<th>outdoor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>0.93</td>
<td>0.76</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stand</th>
<th>walk</th>
<th>passenger</th>
<th>platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>0.96</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>outdoor</th>
<th>park</th>
<th>walk</th>
<th>pedestrian</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.99</td>
<td>0.83</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>outdoor</th>
<th>watch</th>
<th>audience</th>
<th>square</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.95</td>
<td>0.94</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>indoor</th>
<th>conference</th>
<th>center</th>
<th>stand</th>
<th>walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.65</td>
<td>0.55</td>
<td>0.55</td>
<td>0.53</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Crowd Behaviour Tasks

• Crowd Anomaly Detection

Potential Applications

Surveillance

Video Sharing

Human Computer Interaction
Outline

• Background

• Transductive Zero-Shot Action Recognition

• Multi-Task Zero-Shot Embedding

• Zero-Shot Crowd Analysis
Motivation

- Ever Increasing #Categories for action recognition

- KTH 6 Classes
- Weizmann 9 Classes
- Olympic Sports 16 Classes
- ACTIVITYNET 203 Classes
- UCF101 101 Classes
- HMDB51 51 Classes
Motivation

- Ever Increasing #Categories
  - KTH 6 Classes
  - Weizmann 9 Classes
  - Olympic Sports 16 Classes
  - HMDB51 51 Classes
  - UCF101 101 Classes
  - 203 Classes

Limitations

- Expensive to collect training data
- Annotating video is costly
Zero-Shot Learning (ZSL)

- Can we use videos from known class to help predict videos from unknown classes?
Attribute Semantic Space

• Attribute Based

- Hammer Throw
- Discus Throw

Attributes:
- Throw Away
- Outdoor
- Turn Around
- Ball
- Bend
Attribute Semantic Space

• Attribute Based

Attributes
- Throw Away
- Outdoor
- Turn Around
- Ball
- Bend

Attributes

Known a priori

Discus Throw

Hammer Throw

Shot-put

17
Attribute Semantic Space

• Attribute Based

- Hammer Throw
- Discus Throw
- Shot-put
- Throw Away
- Outdoor
- Turn Around
- Ball
- Bend

Test video
Attribute Semantic Space

• Attribute Based

**Limitations**

• Ontological problem

• Manual label attributes is costly for videos

• Incompatible with other attribute sets
Word-Vector Semantic Space

Feature Space $X$

Word-Vector Space $Z$

Discus Throw = $[0.2 \ 0.5 \ 0.1 \ ...]$

Hammer Throw = $[0.1 \ 0.6 \ 0.1 \ ...]$

$z = f(x)$
Word-Vector Semantic Space

Feature Space $X$

Word-Vector Space $Z$

Discus Throw = $[0.2\ 0.5\ 0.1\ ...]$  
Hammer Throw = $[0.1\ 0.6\ 0.1\ ...]$  
ShotPut = $[0.3\ 0.4\ 0.2\ ...]$
Semantic Word-Vector

- Skip-gram model predicts adjacent words

\[
\max \left\{ \mathbf{z} \right\} \quad \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(z_{t+j} | z_t) \\
p(z_i | z_j) = \frac{\exp(z_i^T z_j)}{\sum_i \exp(z_i^T z_j)}
\]

Result of this optimization

vec(“ball”)=[-0.004 0.01 0.01 -0.03 0.05]
vec(“sword”)=[0.16 0.06 0.09 -0.06 -0.002]
vec(“archery”)=[0.02 0.01 0.02 -0.03 -0.03]
vec(“boxing”)=[-0.08 -0.01 0.15 -0.01 0.09]

Benefits

• Geometric Meaningful
Benefits

- Unsupervised Semantic Space

Machine learning

For the journal, see Machine Learning (journal).

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make deterministic predictions or decisions, rather than following strictly static program instructions.

Machine learning is closely related to and often overlaps with computational statistics, which also measures predictive accuracy. However, the differences are significant: machine learning focuses on prediction, while statistics often include the analysis of observed empirical data. Moreover, a machine learning system operates on a dataset; statistical analysis often incorporates additional data for a more extensive study.

In industrial contexts, machine learning methods are referred to as predictive analytics or predictive modeling.

Image processing

In imaging science, image processing is the process of performing operations on images using mathematical operations by using any form of signal processing for which the input is an image, such as a photograph or video frame. The output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it.

Image processing usually refers to digital image processing, but optical and analog image processing also are possible. This article is about general techniques that apply to all of them. The acquisition of images (producing the input image in the first place) is referred to as imaging.

Closely related to image processing are computer graphics and computer vision. In computer graphics, images are usually made from physical models of objects, environments, and lighting, instead of being observed (via imaging devices such as cameras) from natural scenes, or in most automated systems. Computer vision, on the other hand, is often considered high-level image processing out of which a machine/computer/smartphone intends to decipher the physical contents of an image or a sequence of images (e.g., video) or in a fully automatic way.
Benefits

• Wide coverage of words

\[
\begin{align*}
\text{Vec(“Apple”) } &= [0.2 \ 0.3 \ 0.1 \ …] \\
\text{Vec(“Bear”) } &= [0.1 \ 0.9 \ 0.1 \ …] \\
\text{Vec(“Car”) } &= [0.6 \ 0.2 \ 0.4 \ …] \\
\text{Vec(“Desk”) } &= [0.2 \ 0.8 \ 0.4 \ …] \\
\text{Vec(“Fish”) } &= [0.5 \ 0.2 \ 0.3 \ …] \\
\end{align*}
\]

…
Benefits

• Uniform across datasets

\[
\text{Dataset 1}
\]

Discus Throw = \([0.2 0.5 \ldots]\)

HammerThrow = \([0.1 0.2 \ldots]\)

\[
\text{Dataset 2}
\]

Discus Throw = \([0.2 0.5 \ldots]\)

HammerThrow = \([0.1 0.2 \ldots]\)
Challenges

• **Domain Shift**

Feature Space $X$

Semantic Vector Space $Y$

- Hammer Throw
- Discus Throw
- Sword Exercise
- Play Guitar
Challenges

- Domain Shift

Feature Space $X$

Semantic Vector Space $Y$

Hammer Throw

Discus Throw

HammerThrow

Sword Exercise

Confusion

Play Guitar
Our Solution

Our Solution

Low-Level Visual Feature

• Improved Trajectory Feature for $x$

Combinations of Multi Words

• A phrase is constructed from single word vectors

Additive Composition

\[ \text{vec(“Apply Eye Makeup”)} = \text{vec(“Apply”)} + \text{vec(“Eye”)} + \text{vec(“Makeup”)} \]

\[ \text{vec(“Brushing Teeth”)} = \text{vec(“Brushing”)} + \text{vec(“Teeth”)} \]

\[ \text{vec(“Playing Guitar”)} = \text{vec(“Playing”)} + \text{vec(“Guitar”)} \]
Our Solution

Visual to Semantic Mapping by Regularized Linear Regression

- Multi-Dimensional Regularized Linear Regression

\[
\min_{\mathbf{W}} \sum_{i=1}^{N} \left\| \mathbf{z}_i - \mathbf{Wx}_i \right\|_2^2 + \lambda \left\| \mathbf{W} \right\|_2^2
\]

\( \mathbf{x} \) is \( N \) Dimension Feature Space

\( \mathbf{z} \) is \( D \) Dimension Semantic Space
Domain Shift – Semi Supervised (Manifold Regularized) Regression

- Semi-supervised regression is applied to tackle domain shift which takes test data distribution into consideration.

\[ \sum \omega_{ij} \left\| f(x_i) - f(x_j) \right\|_2^2 : x \in [X_{tr} ; X_{te}] \]

KNN Graph to model Manifold

Train and Test Data in Feature Space

- \( X_{tr} = X_{tr}^{trg} \)
- \( X_{te} = X_{te}^{trg} \)
Domain Shift – Semi Supervised (Manifold Regularized) Regression

- Semi-supervised regression is applied to tackle domain shift which takes test data distribution into consideration

\[
\min_{W} \sum_{i=1}^{N} \| z_i -WX_i \|_2^2 + \lambda \| W \|_2^2 + \gamma \sum_{ij} \sigma_{ij} \| WX_i - WX_j \|_2^2
\]
Our Solution

Additional datasets are available

Data Augmentation

- Use more training data from Auxiliary Dataset to help learn a better regression

Augmented Train and Test Data in Feature Space

\[ X_{tr} = [X_{tr}^{trg} ; X_{aux}^{aux}] \]
\[ X_{te} = X_{te}^{trg} \]

More Data is considered to learn more robust regressor
Semantic Word Vector Approach

**Input Data**

- **Target Dataset**
  - Visual Feature $X^{trg}$
  - Class Labels $Y^{trg}$

- **Auxiliary Dataset**
  - Visual Feature $X^{aux}$
  - Class Labels $Y^{aux}$

**Semantic Embedding Space**

- $Z^{trg} = g(Y^{trg})$
- $Z^{aux} = g(Y^{aux})$

**Semantic Embedding Representation**

**Visual to Semantic Mapping**

- $X_{tr} = [X_{trg}^{trg}, X_{aux}^{aux}]$
- $Z_{tr} = [Z_{trg}^{trg}, Z_{aux}^{aux}]$
- $f: \min(\|Z_{tr} - f(X_{tr})\|)$
- $Z_{ts}^{trg} = f(X_{ts}^{trg})$

**Action Recognition Tasks**

- Multi-shot Action Recognition
- Zero-shot Action Recognition

Augment Training Data for Visual to Semantic Mapping
Zero-Shot Recognition by Nearest Neighbor

- Do nearest Neighbor search in word-vector space to predict category of test data
Domain Shift – SelfTraining

• Self-training is applied to tackle domain shift

\[ \tilde{Z}_{te} = f(x) \]

\[ Z("Taichi") = g("Taichi") \]

\[ Z^*("Taichi") = \frac{1}{K} \sum_{\tilde{Z}_{te} \in \text{NN}(Z("Taichi"),K)} \tilde{Z}_{te} \]

\( \text{NN}(Z_{\text{proto}},K) \) is the KNN function

4 NN example

\[ Z^*("Taichi") = (\tilde{Z}_5 + \tilde{Z}_6 + \tilde{Z}_7 + \tilde{Z}_8)/4 \]
Domain Shift – SelfTraining

- Self-training is applied to tackle domain shift

\[
\tilde{z}_{te} = f(x)
\]

\[
Z(\text{"Taichi"}) = g(\text{"Taichi"})
\]

\[
Z^*(\text{"Taichi"}) = \frac{1}{K} \sum_{\tilde{z}_{te} \in \text{NN}(Z(\text{"Taichi"}), K)} \tilde{z}_{te}
\]

\(\text{NN}(Z_{\text{proto}}, K)\) is the KNN function

4 NN example

\[
Z^*(\text{"Taichi"}) = (\tilde{z}_5 + \tilde{z}_6 + \tilde{z}_7 + \tilde{z}_8) / 4
\]
Experiments

Dataset:

- HMDB51 – 51 classes 6766 videos
- UCF101 – 101 classes 13320 videos
- Olympic Sports – 16 classes 786 videos
- CCV – 20 classes 9317 videos
- USAA – 8 classes (subset of CCV)

Visual Feature:

- Improved Trajectory Feature [1]
- Improved fisher vector encoding [2]

Semantic Embedding Space:

- Skip-gram neural network model trained on Google News Dataset
- 300 dimension word vector


Qualitative Insight

• How do Self-Training, Manifold Regularization and Data Augmentation perform

All data projected to 2D space via T-SNE [1]

# Zero-Shot Experiment

- Test on public human action datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>DA</th>
<th>Trans</th>
<th>Embed</th>
<th>Feat</th>
<th>HMDB51</th>
<th>UCF101</th>
<th>Olympic Sports</th>
<th>CCV</th>
<th>USAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Guess</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>RR (Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>MR (Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>SJE (Akata et al, 2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>ConSc (Norouzi et al, 2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>TMV-MLP (Fu et al, 2014a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>SVE (Xu et al, 2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>DAP (Lampert et al, 2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>IAP (Lampert et al, 2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>HAA (Liu et al, 2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>PST (Rohrbach et al, 2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>M2LTM (Fu et al, 2014b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>TMV-MLP (Fu et al, 2014a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>TMV-MLP (Fu et al, 2015a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>UDA (Kodirov et al, 2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>MR (Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>TMV-MLP (Fu et al, 2014a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
<tr>
<td>UDA (Kodirov et al, 2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>2.0</td>
<td>12.5</td>
<td>10.0</td>
<td>25.0</td>
</tr>
</tbody>
</table>
Outline

• Background

• Transductive Zero-Shot Action Recognition

• Multi-Task Zero-Shot Embedding

• Zero-Shot Crowd Analysis
Revisit Visual-to-Semantic Mapping

• Multi-Dimensional Regularized Linear Regression

\[ \min_{W} \sum_{i=1}^{N} \|z_i - Wx_i\|_2^2 + \lambda \|W\|_2^2 \]

\[ z \text{ is } D \text{ Dimension Semantic Space} \]

\[ x \text{ is } N \text{ Dimension Feature Space} \]
Visual-to-Semantic Mapping by Multi-Task Regression

• Two stage regression

\[ W = AS \]

\[ \begin{align*} x_1 & \rightarrow l_1 \\ \vdots & \rightarrow \vdots \\ x_d & \rightarrow l_T \\ \end{align*} \]

\[ \begin{align*} a_1 & \rightarrow s_1 \\ \vdots & \rightarrow \vdots \\ a_T & \rightarrow s_m \\ \end{align*} \]

Visual-to-Semantic Mapping by Multi-Task Regression

- Two stage regression

\[
\min_{\{s_t\}, A, \{l_i\}} \sum_{t=1}^{T} \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} \left( \frac{1}{2} \|z_{t,i} - s_t l_i\|^2 + \|l_i - Ax_i\|^2 \right) + \\
\lambda_S \sum_{t=1}^{T} \|s_t\|^2 + \lambda_A \|A\|_F^2 + \lambda_L \sum_{i=1}^{n_x^{tr}} \|l_i\|^2
\]

![Diagram](image)
Visual-to-Semantic Mapping by Multi-Task Regression

- Solve efficiently

Loss Function

\[
\min_{\{s_t\}, A, \{l_i\}} \sum_{t=1}^{T} \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} \left( \|z_{t,i} - s_t l_i\|_2^2 + \|l_i - Ax_i\|_2^2 \right) + \\
\lambda_s \sum_{t=1}^{T} \|s_t\|_2^2 + \lambda_A \|A\|_F^2 + \lambda_L \sum_{i=1}^{n_x^{tr}} \|l_i\|_2^2
\]

Iterative Update

\[
L = (S^T S + (\lambda_L n_x^{tr} + 1)I)^{-1} (S^T Z + AX) \\
S = ZL^T (LL^T + \lambda_S n_x^{tr} I)^{-1} \\
A = LX^T (XX^T + \lambda_A n_x^{tr} I)^{-1}
\]
Multi-Task Embedding

- Lower dimension subspace embedding

\[ z^* = \arg\min_z \|Ax - S^{-1}z\|_2^2 \]
Importance Weighting for Domain Adaptation

\[
\min_{\omega} D_{KL}(p^{te}(x) \| \omega(x) p^{tr}(x)) = \int p^{te}(x) \log \frac{p^{te}(x)}{\omega(x) p^{tr}(x)} \, dx
\]
Revisit Visual-to-Semantic Mapping

- Uniform weight is given to all training examples

Uniform Model

\[ \min_W \sum_{i=1}^{N} \| z_i - Wx_i \|_2^2 + \lambda \| W \|_2^2 \]

Weighted Model

\[ \min_W \sum_{i=1}^{N} \omega_i \| z_i - Wx_i \|_2^2 + \lambda \| W \|_2^2 \]
Experiments

Dataset:
- HMDB51 – 51 classes 6766 videos
- UCF101 – 101 classes 13320 videos
- Olympic Sports – 16 classes 786 videos

Feature:
- Improved Trajectory Feature [1]
- Improved fisher vector encoding [2]

Semantic Embedding Space:
- Skip-gram neural network model trained on Google News Dataset
- 300 dimension word vector

## MTL v.s. STL

<table>
<thead>
<tr>
<th>ZSL Model</th>
<th>MTL</th>
<th>Latent Matching</th>
<th>HMDB51</th>
<th>UCF101</th>
<th>Olympic Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR [1]</td>
<td>X</td>
<td>N/A</td>
<td>18.3±2.1</td>
<td>14.5±0.9</td>
<td>40.9±10.1</td>
</tr>
<tr>
<td>RMTL [2]</td>
<td>✓</td>
<td>X</td>
<td>18.5±2.1</td>
<td>14.6±1.1</td>
<td>41.1±10.0</td>
</tr>
<tr>
<td>RMTL [2]</td>
<td>✓</td>
<td>✓</td>
<td>18.7±1.7</td>
<td>14.7±1.0</td>
<td>41.1±10.0</td>
</tr>
<tr>
<td>GOMTL [3]</td>
<td>✓</td>
<td>X</td>
<td>18.5±2.2</td>
<td>13.1±1.5</td>
<td>43.5±8.8</td>
</tr>
<tr>
<td>GOMTL [3]</td>
<td>✓</td>
<td>✓</td>
<td>18.9±1.0</td>
<td>14.9±1.5</td>
<td>44.5±8.5</td>
</tr>
<tr>
<td>MTE(Ours)</td>
<td>✓</td>
<td>X</td>
<td>18.7±2.2</td>
<td>14.2±1.3</td>
<td>44.5±8.2</td>
</tr>
<tr>
<td>MTE(Ours)</td>
<td>✓</td>
<td>✓</td>
<td>19.7±1.6</td>
<td>15.8±1.3</td>
<td>44.3±8.1</td>
</tr>
</tbody>
</table>

## Importance Weighting

<table>
<thead>
<tr>
<th>ZSL Model</th>
<th>Weighting Model</th>
<th>HMDB51</th>
<th>UCF101</th>
<th>Olympic Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR [1]</td>
<td>Uniform</td>
<td>21.9±2.4</td>
<td>19.4±1.7</td>
<td>46.5±9.4</td>
</tr>
<tr>
<td>MTE (Ours)</td>
<td>Uniform</td>
<td><strong>23.4±3.4</strong></td>
<td><strong>20.9±1.5</strong></td>
<td><strong>49.4±8.8</strong></td>
</tr>
<tr>
<td>RR [1]</td>
<td>Visual KLI EP</td>
<td>23.2±2.7</td>
<td>20.3±1.6</td>
<td>47.2±9.3</td>
</tr>
<tr>
<td>RR [1]</td>
<td>Category KLI EP</td>
<td>23.0±2.1</td>
<td>20.2±1.6</td>
<td>51.8±8.7</td>
</tr>
<tr>
<td>RR [1]</td>
<td>Full KLI EP</td>
<td>23.7±2.7</td>
<td>20.7±1.4</td>
<td>51.3±9.0</td>
</tr>
<tr>
<td>MTE (Ours)</td>
<td>Visual KLI EP</td>
<td>23.4±2.8</td>
<td>20.8±2.0</td>
<td>51.4±9.2</td>
</tr>
<tr>
<td>MTE (Ours)</td>
<td>Category KLI EP</td>
<td>23.3±2.4</td>
<td>20.9±1.7</td>
<td>50.9±8.3</td>
</tr>
<tr>
<td>MTE (Ours)</td>
<td>Full KLI EP</td>
<td><strong>23.9±3.0</strong></td>
<td><strong>21.9±2.7</strong></td>
<td><strong>52.3±8.1</strong></td>
</tr>
</tbody>
</table>

# Ours v.s. State-of-the-Art

<table>
<thead>
<tr>
<th>Method</th>
<th>Embed Feat</th>
<th>TD</th>
<th>Aug</th>
<th>HMDB51</th>
<th>UCF101</th>
<th>Olympic Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTE</td>
<td>W</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>19.7±1.6</td>
<td>15.8±1.3</td>
</tr>
<tr>
<td>MTE+Full KLIEP</td>
<td>W</td>
<td>FV</td>
<td>✓</td>
<td>✓</td>
<td>23.9±3.0</td>
<td>21.9±2.7</td>
</tr>
<tr>
<td>MTE+Full KLIEP+PP</td>
<td>W</td>
<td>FV</td>
<td>✓</td>
<td>✓</td>
<td>24.8±2.2</td>
<td>22.9±3.3</td>
</tr>
<tr>
<td>MTE</td>
<td>A</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>N/A</td>
<td>18.3±1.7</td>
</tr>
<tr>
<td>State-of-the-art models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAP [1] CVPR09</td>
<td>A</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>N/A</td>
<td>15.9±1.2</td>
</tr>
<tr>
<td>IAP [1] CVPR09</td>
<td>A</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>N/A</td>
<td>16.7±1.1</td>
</tr>
<tr>
<td>HAA [2] CVPR11</td>
<td>A</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>N/A</td>
<td>14.9±0.8</td>
</tr>
<tr>
<td>SVE [3] ICIP15</td>
<td>W</td>
<td>BoW</td>
<td>X</td>
<td>X</td>
<td>14.9±1.8</td>
<td>12.0±1.4</td>
</tr>
<tr>
<td>SVE [3] ICIP15</td>
<td>W</td>
<td>BoW</td>
<td>✓</td>
<td>✓</td>
<td>22.8±2.6</td>
<td>18.4±1.4</td>
</tr>
<tr>
<td>ESZSL [4] ICML15</td>
<td>W</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>18.5±2.0</td>
<td>15.0±1.3</td>
</tr>
<tr>
<td>ESZSL [4] ICML15</td>
<td>A</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>N/A</td>
<td>17.1±1.2</td>
</tr>
<tr>
<td>SJE [5] ICCV15</td>
<td>W</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>13.3±2.4</td>
<td>9.9±1.4</td>
</tr>
<tr>
<td>SJE [5] ICCV15</td>
<td>A</td>
<td>FV</td>
<td>X</td>
<td>X</td>
<td>N/A</td>
<td>12.0±1.2</td>
</tr>
</tbody>
</table>

Outline

• Background

• Transductive Zero-Shot Action Recognition

• Multi-Task Zero-Shot Embedding

• Zero-Shot Crowd Analysis
Zero-Shot Crowd Analysis

- Interesting crowd behaviours, e.g. violence, are rare
Motivation

- Interesting Crowd Behaviours are Rare, e.g. ViolenceFlow Dataset.

Motivation

- Exploit Existing Crowd Video Data, e.g. WWW Crowd dataset
Zero-Shot Predict Crowd Behaviour

• Predict “Violence” in Zero-Shot Manner

Challenges

• Semantic relatedness v.s. Visual relatedness

“Outdoor” & “Indoor” highly related in word-vector space

\[ \text{vec(“Outdoor”)}^\top \text{vec(“Indoor”)} = 0.7104 \]

But Visually Never Co-Occur!!
Solution

• Exploit co-occurrence of labels to improve ZSL
Solution

• Exploit co-occurrence of labels to improve ZSL
Zero-Shot Predict Crowd Behaviour

- Visual Context Aware ZSL

\[
p(y_q^* | y_p) = \frac{\exp\left(\frac{1}{\gamma}v_q^\top v_p^S\right)}{\sum_{p=1}^{P} \exp\left(\frac{1}{\gamma}v_q^\top v_p^S\right)}
\]

\[
p(y_q^* | y_p) = \frac{\exp(v_q^\top M v_p)}{\sum_p \exp(v_q^\top M v_p)}
\]
Solution

- Exploit co-occurrence of labels to improve ZSL
Experiment

**Dataset**
- WWW Crowd dataset [1]
- Violence Flow [2]

**Visual Feature**
- Improved Trajectory Feature [3]

**Semantic Embedding Space:**
- Skip-gram neural network model trained on Google News Dataset
- 300 dimension word vector

**Setting**
- Training on WWW dataset and testing on violence flow
- Evaluate both accuracy and ROC

Performance

• Evaluation on Violence Detection Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Split</th>
<th>Feature</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>WVE[1]</td>
<td>Zero-Shot</td>
<td>ITF</td>
<td>64.27+5.06</td>
<td>64.25</td>
</tr>
<tr>
<td>ESZSL[2]</td>
<td>Zero-Shot</td>
<td>ITF</td>
<td>61.30-8.28</td>
<td>61.76</td>
</tr>
<tr>
<td>ExDAP[3]</td>
<td>Zero-Shot</td>
<td>ITF</td>
<td>54.47-7.37</td>
<td>52.31</td>
</tr>
<tr>
<td>TexCAZSL</td>
<td>Zero-Shot</td>
<td>ITF</td>
<td>67.07-3.87</td>
<td>69.95</td>
</tr>
<tr>
<td>CoCAZSL</td>
<td>Zero-Shot</td>
<td>ITF</td>
<td>80.52-4.67</td>
<td>87.22</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>5-fold CV</td>
<td>ITF</td>
<td>94.72-4.85</td>
<td>98.72</td>
</tr>
<tr>
<td>Linear SVM[4]</td>
<td>5-fold CV</td>
<td>ViF</td>
<td>81.30-0.21</td>
<td>85.00</td>
</tr>
</tbody>
</table>

TexCAZSL uses M=I
CoCAZSL learns M from attribute co-occurrence

\[
P(\text{"viol"} | \text{"mob"}) = \frac{1}{Z} \exp \left( \mathbf{v}(\text{"viol"})^\top \mathbf{M} \mathbf{v}(\text{"mob"}) \right)
\]

Qualitative Evaluation

- Relation to “Violence”

\[ p(y^*_q|y_p) = \frac{\exp(v_q^T M v_p)}{\sum_p \exp(v_q^T M v_p)} \]
Conclusion

• Zero-shot learning can overcome the challenge of labelling ever increasing data

• Unsupervised word-vector semantic space produces reasonable ZSL performance without labelling attribute

• Access to testing data could substantially improve the quality of ZSL

• ZSL underpinned by large amount of related data may perform rather close to specifically collected small training data
Thank You