

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

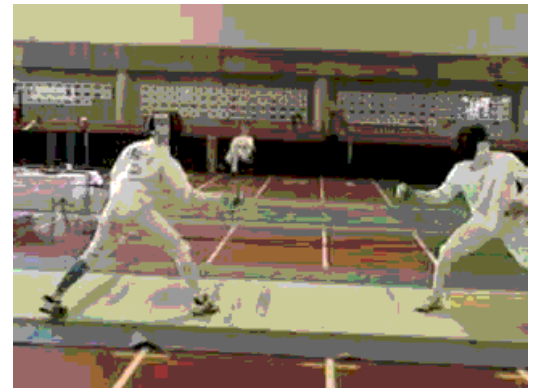
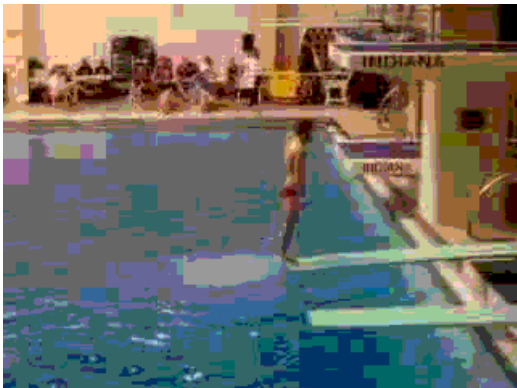
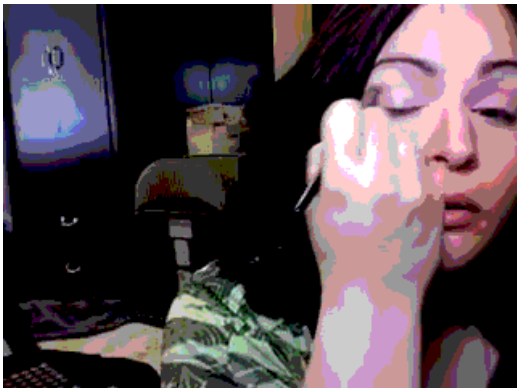
Video Behaviour

Defined as Visually Distinguishable Activities

- Human Actions
- Crowd Behaviour

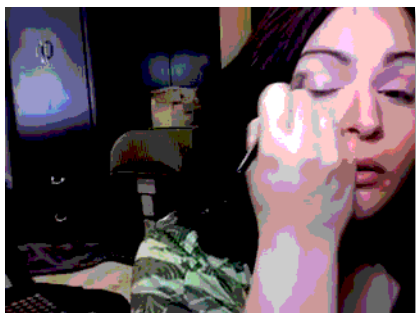
Human Actions

- Individual or multiple interactive human activities

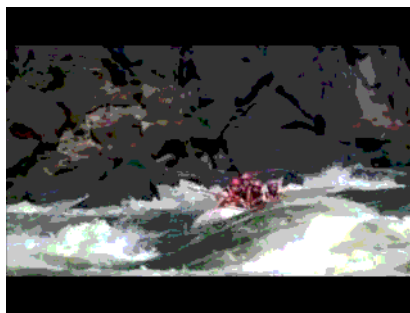


Human Actions Tasks

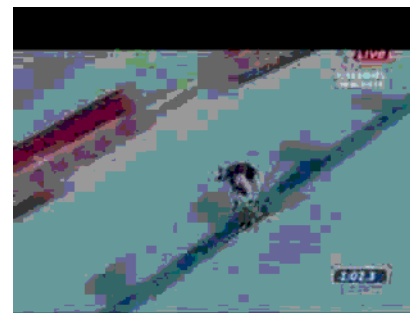
- Action Recognition



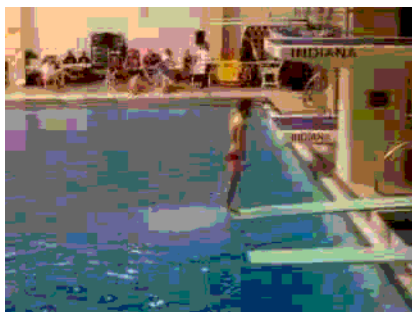
Eye Makeup



Rafting



Swimming



Diving



Archery

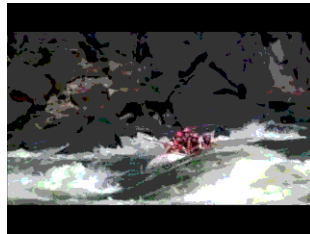


Fencing

Human Actions Tasks

- Action Detection (Retrieval)

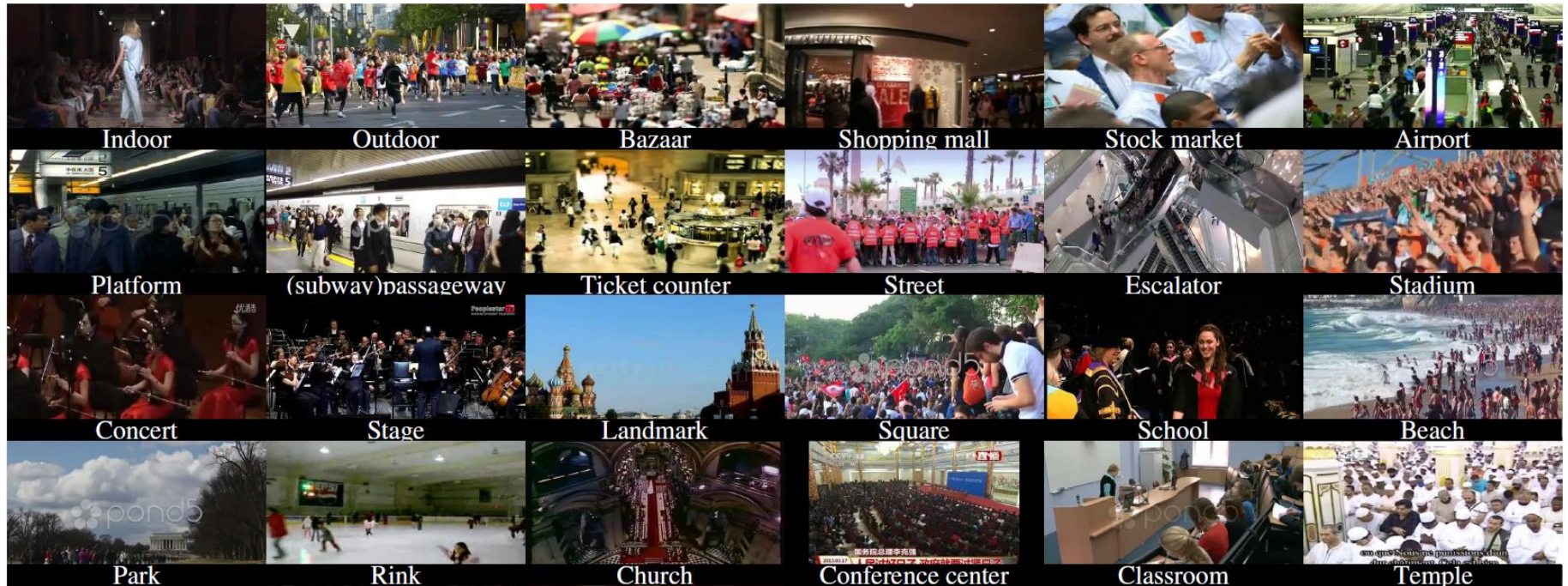
Given query “Swimming” return ranked videos



.....

Crowd Behaviour

- A group of people acting collectively



Crowd Behaviour Tasks

- Crowd Behaviour Profiling

	pedestrian 1.00 street 1.00 outdoor 0.99 walk 0.99		stand 0.99 indoor 0.96 performance 0.86 sit 0.86
	stand 0.99 photograph 0.98 outdoor 0.95 photographer 0.93		walk 0.95 stand 0.93 pedestrian 0.76 outdoor 0.66
	stand 0.99 walk 0.99 passenger 0.96 platform 0.86		outdoor 1.00 park 0.99 walk 0.83 pedestrian 0.51
	outdoor 0.99 watch 0.95 audience 0.94 square 0.67		indoor 0.65 conference center 0.55 stand 0.53 walk 0.48

Crowd Behaviour Tasks

- Crowd Anomaly Detection

Violence Detection

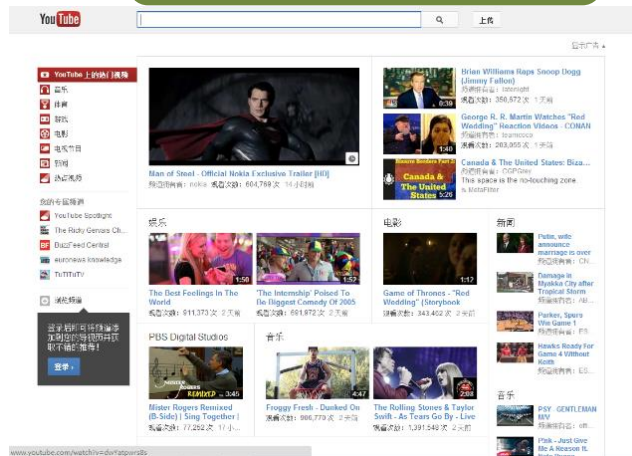


Potential Applications

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



2004

KTH 6 Classes



2005

Weizmann 9 Classes



Olympic Sports 16 Classes

2010



ACTIVITYNET

2015

203 Classes

2012



UCF101 101 Classes

2011



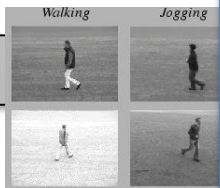
HMDB51 51 Classes

Motivation

• Ever

Limitations

- Expensive to collect training data
- Annotating video is costly



KTH 6 Classes



203 Classes



UCF101 101 Classes



HMDB51 51 Classes



2010

Classes

Zero-Shot Learning (ZSL)

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

Known Classes



Hammer
Throw

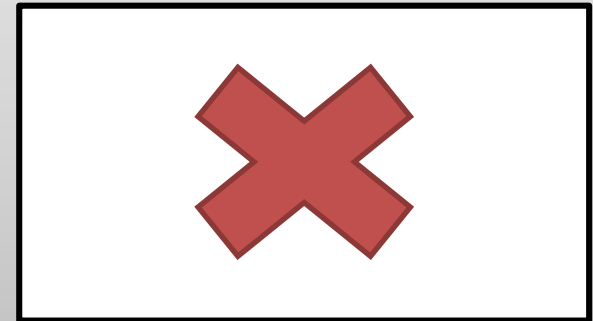


Discus
Throw



Unknown Classes

Shot-Put



Attribute Semantic Space

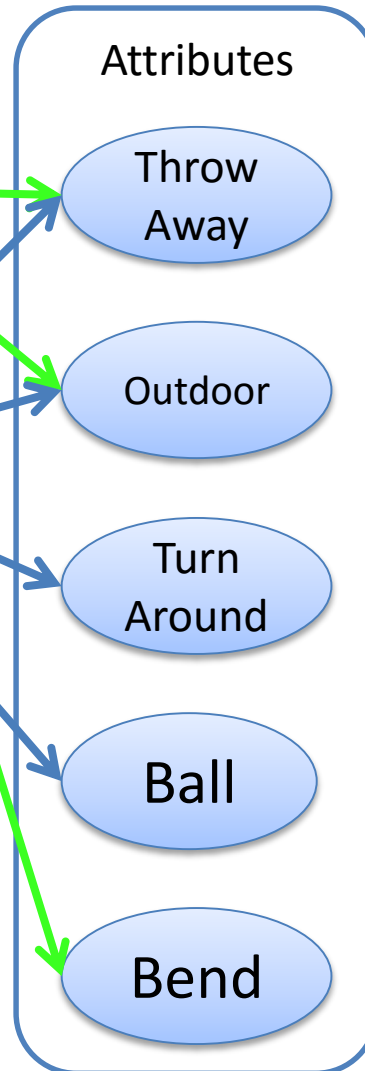
- Attribute Based



Hammer
Throw

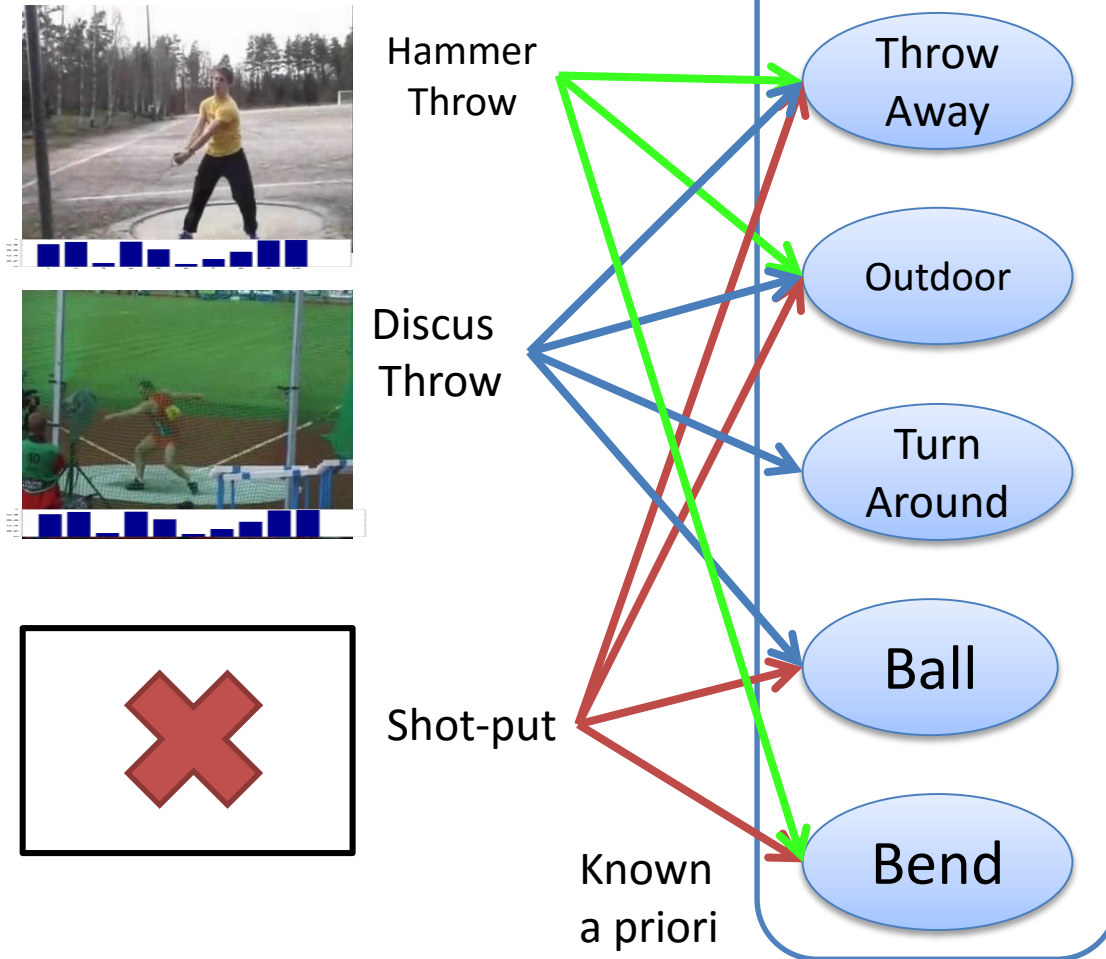


Discus
Throw



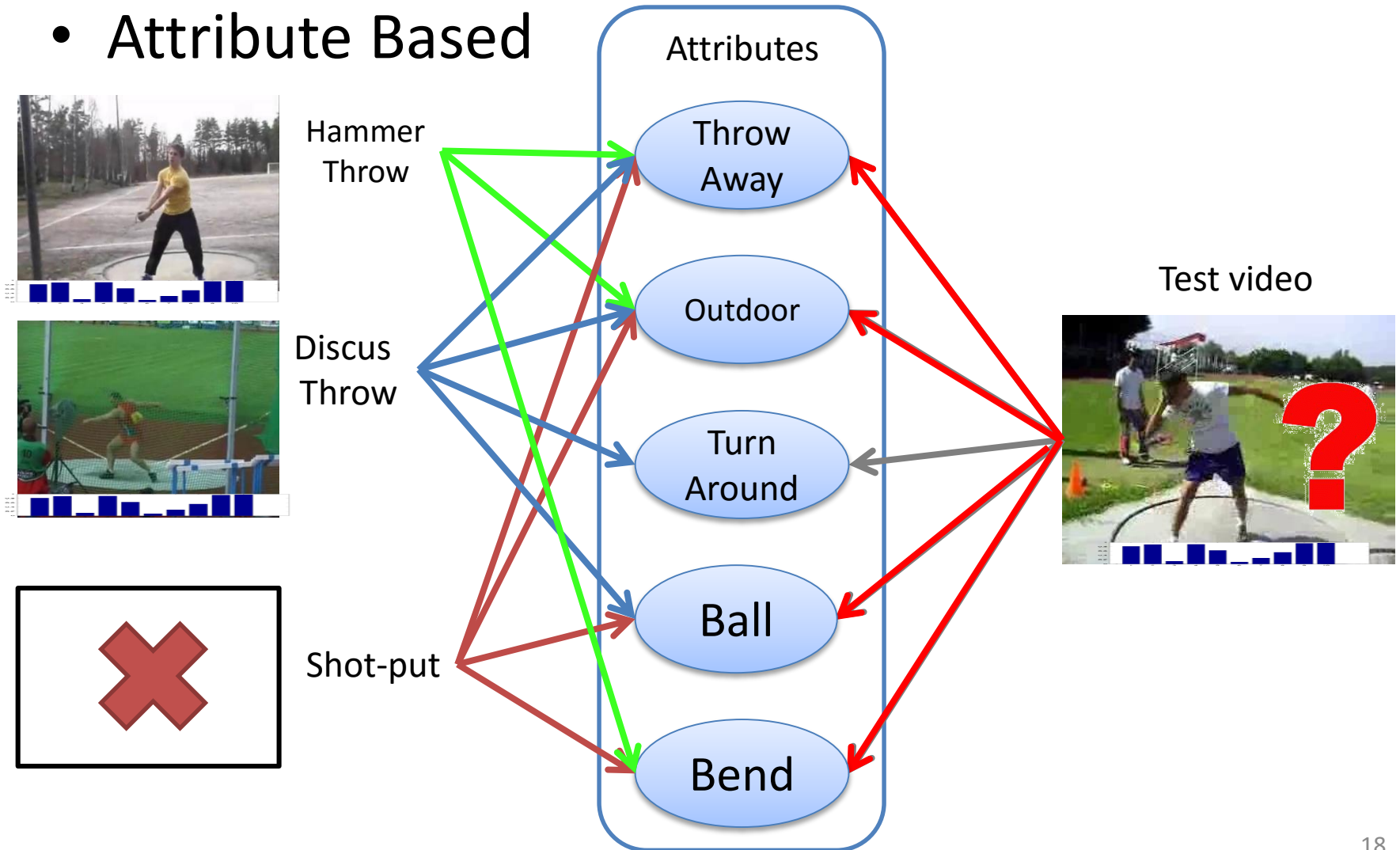
Attribute Semantic Space

- Attribute Based



Attribute Semantic Space

- Attribute Based



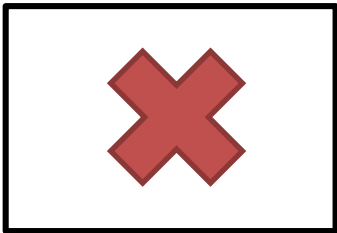
Attribute Semantic Space

- Attribute Based

Attributes

Limitations

- Ontological problem
- Manual label attributes is costly for videos
- Incompatible with other attribute sets



Bend

Word-Vector Semantic Space

Feature Space X

Word-Vector Space Z

Hammer
Throw



Discus Throw = [0.2 0.5 0.1 ...]



$$z = f(x)$$

Discus
Throw



Hammer Throw = [0.1 0.6 0.1 ...]

Word-Vector Semantic Space

Feature Space X

Word-Vector Space Z

Hammer
Throw



Discus
Throw



Discus Throw = [0.2 0.5 0.1 ...]



ShotPut = [0.3 0.4 0.2 ...]

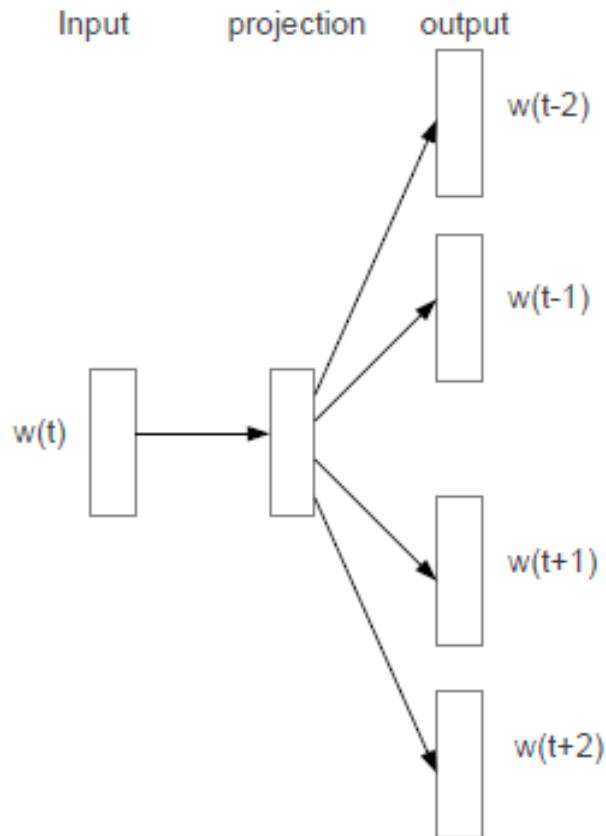


Hammer Throw = [0.1 0.6 0.1 ...]



Semantic Word-Vector

- Skip-gram model predicts adjacent words



$$\max_{\{z\}} \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(z_{t+j} | z_t)$$

$$p(z_i | z_j) = \frac{\exp(z_i^T z_j)}{\sum_i \exp(z_i^T z_j)}$$

Result of this optimization

vec("ball")=[-0.004 0.01 0.01 -0.03 0.05]

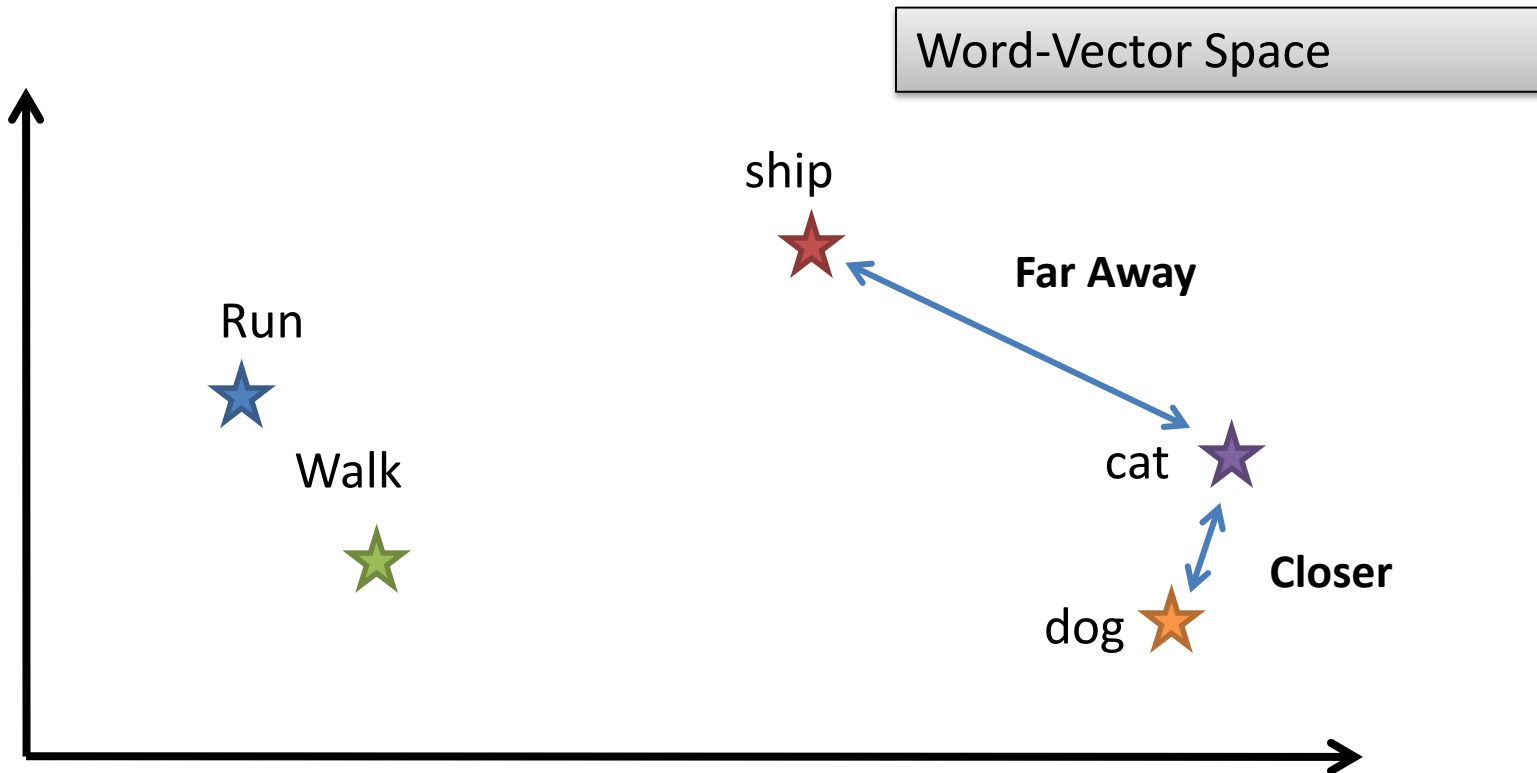
vec("sword")=[0.16 0.06 0.09 -0.06 -0.002]

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

Article [Talk](#) Read [Edit](#) [View history](#)

Institute of Electrical and Electronics Engineers

From Wikipedia, the free encyclopedia
(Redirected from [IEEE](#))

"IEEE" redirects here. It is not to be confused with [Institution of Electrical Engineers \(IEE, I-double-E\)](#).

 This article relies too much on references to primary sources. Please improve this article by adding secondary or tertiary sources. (August 2014)

The **Institute of Electrical and Electronics Engineers** (**IEEE**) is a professional association with its corporate office in New York City and its operations center in Piscataway, New Jersey. It was formed in 1963 from the amalgamation of the **American Institute of Electrical Engineers** and the **Institute of Radio Engineers**. Today, it is the world's largest association of technical professionals with more than 400,000 members in chapters around the world. Its objectives are the educational and technical advancement of electrical and electronic engineering, telecommunications, computer engineering and allied disciplines.

IEEE	
	
Founded	January 1, 1963
Type	Professional Organization
Focus	Electrical, Electronics, Communications, Computer Engineering, Computer Science and Information Technology ^[1]
Location	Piscataway, New Jersey, USA
Origins	Merger of the American Institute of Electrical

Contents [hide]	
1	IEEE
2	History
3	Organization
4	Publications

Quebec City

From Wikipedia, the free encyclopedia
(Redirected from [Quebec city](#))

Coordinates: 46° 49′ N 71° 13′ W﻿ / ﻿46.817° N 71.217° W﻿ / 46.817; -71.217

Quebec (/kɪbɛk/; French: *Québec* [kɛbɛk] (ⓘ) (ⓘ) (ⓘ)), also **Québec**, **City of Québec**,^[9] **Quebec City**, or **Québec City** (French: *Ville de Québec*),^[10] is the capital of the Canadian province of Quebec. In 2011 the city had a population of 516,622,^[4] and the metropolitan area had a population of 765,706,^[6] making it the second most populous city in Quebec after Montreal, which is about 233 km (145 mi) to the southwest.

The narrowing of the **Saint Lawrence River** proximate to the city's promontory, Cap-Diamant (Cape Diamond), and **Lévis**, on the opposite bank, provided the name given to the city, *Kébec*, an Algonquin word meaning "where the river narrows". Founded in 1608 by **Samuel de Champlain**, Quebec City is one of the oldest cities in North America. The ramparts surrounding Old Quebec (*Vieux-Québec*) are the only fortified city walls remaining in the Americas north of Mexico, and were declared a **World Heritage Site** by **UNESCO** in 1985 as the "Historic District of Old Québec".^{[11][12]}

According to the federal and provincial governments, *Québec* is the city's official name in both French and English,^[13] although *Quebec City* (or its French equivalent, *Ville de Québec*) is commonly used, particularly to distinguish the city from the province.^[10] In French, the names of the province and the city are distinguished grammatically in that the province takes the definite article (*Le Québec, du Québec, au Québec*) and the city does not (*Québec, de Québec, à Québec*).

The city's scenic landscape includes the **Château Frontenac**, a hotel which dominates the skyline, and



Machine learning

From Wikipedia, the free encyclopedia

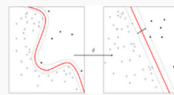
For the journal, see [Machine Learning \(Journal\)](#).

Machine learning is a subfield of **computer science**^[1] that evolved from the study of **pattern recognition** and **computational learning theory** in **artificial intelligence**.^[1] Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.^[2] Such algorithms operate by building a **model** from example inputs in order to make data-driven predictions or decisions,^{[1][2]} rather than following strictly static program instructions.

Machine learning is closely related to and often overlaps with **computational statistics**; a discipline that also specializes in prediction-making. It has strong ties to **mathematical optimization**, which delivers methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Example applications include **spam filtering**, **optical character recognition** (OCR),^[4] search engines and **computer vision**. Machine learning is sometimes conflated with **data mining**,^[5] although that focuses on exploratory data analysis.^[6] Machine learning and pattern recognition "can be viewed as two of the same field."^{[5][7]}

When employed in industrial contexts, machine learning methods may be referred to as **predictive** **analytics** or **predictive modelling**.

Machine learning and data mining



Problems

- Classification
- Clustering
- Regression
- Anomaly detection
- Association rules
- Reinforcement learning
- Structured prediction
- Feature learning
- Online learning
- Semi-supervised learning
- Unsupervised learning
- Learning to rank
- Grammar induction

Image processing

From Wikipedia, the free encyclopedia

In **imaging science**, **image processing** is processing of 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.^[1] 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**.^[2]

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 acquired (via imaging devices such as cameras) from *natural* scenes, as in most animated movies. Computer vision, on the other hand, is often considered *high-level* image processing out of which a machine/computer/software intends to decipher the physical contents of an image or a sequence of images (e.g., videos or 3D full-body semantic recognition scene).



Benefits

- **Wide coverage of words**

Vec("Apple") = [0.2 0.3 0.1 ...]

Vec("Bear") = [0.1 0.9 0.1 ...]

Vec("Car ") = [0.6 0.2 0.4 ...]

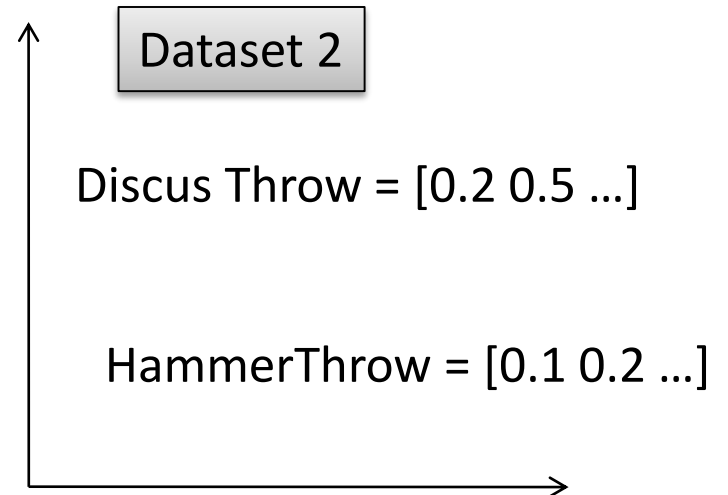
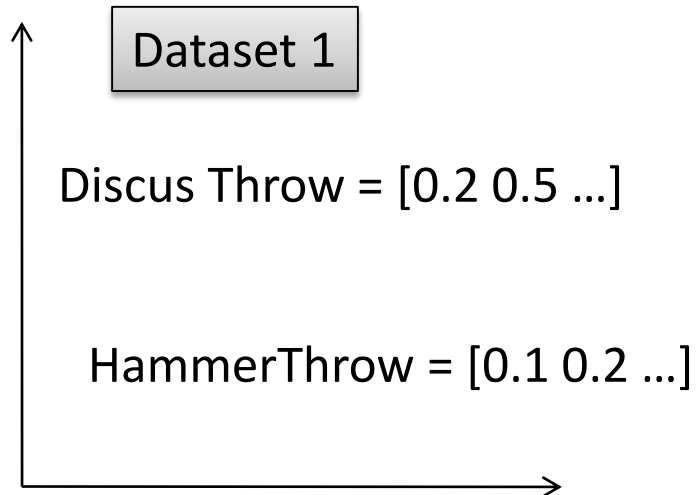
Vec("Desk") = [0.2 0.8 0.4 ...]

Vec("Fish") = [0.5 0.2 0.3 ...]

...

Benefits

- **Uniform across datasets**

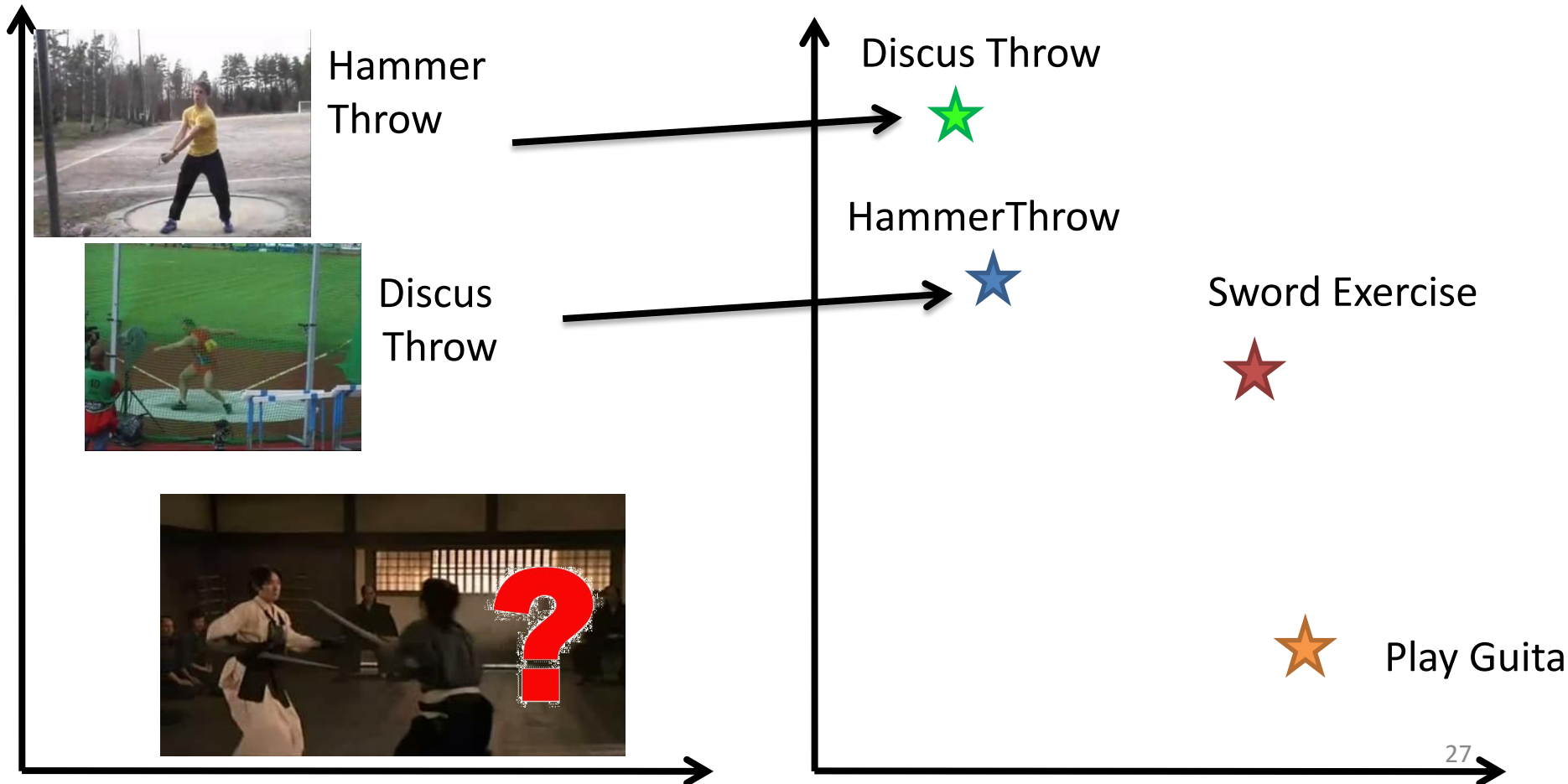


Challenges

- **Domain Shift**

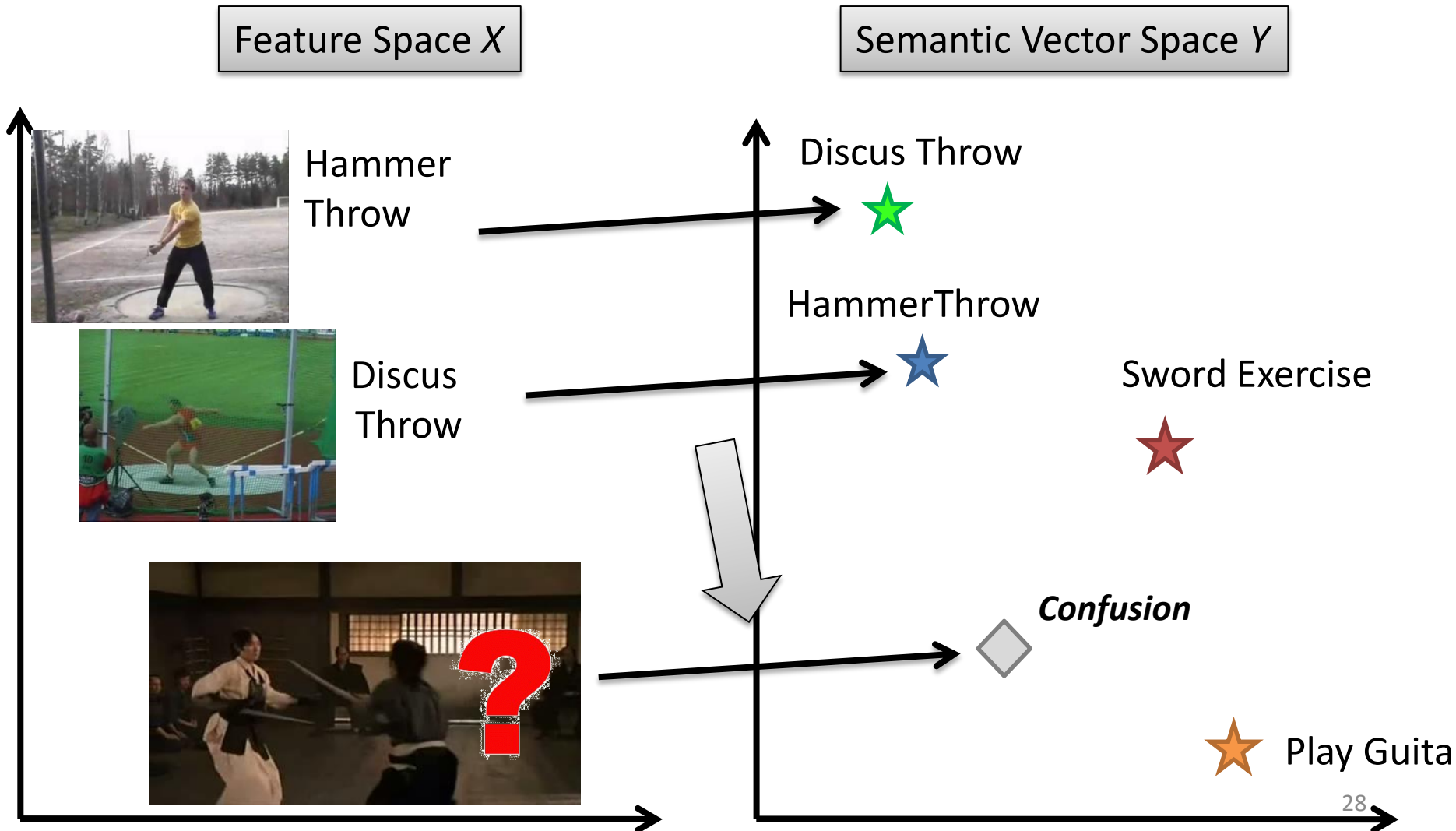
Feature Space X

Semantic Vector Space Y

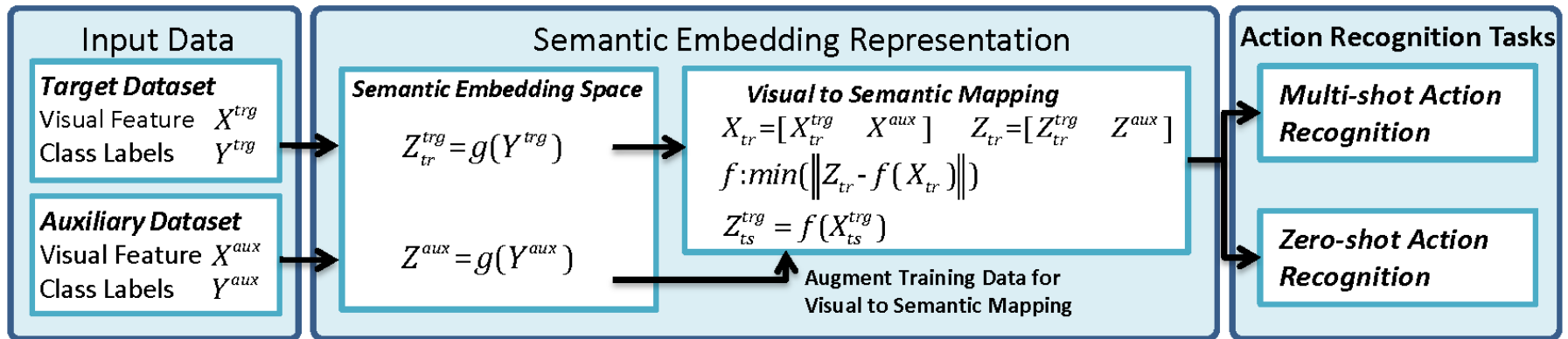


Challenges

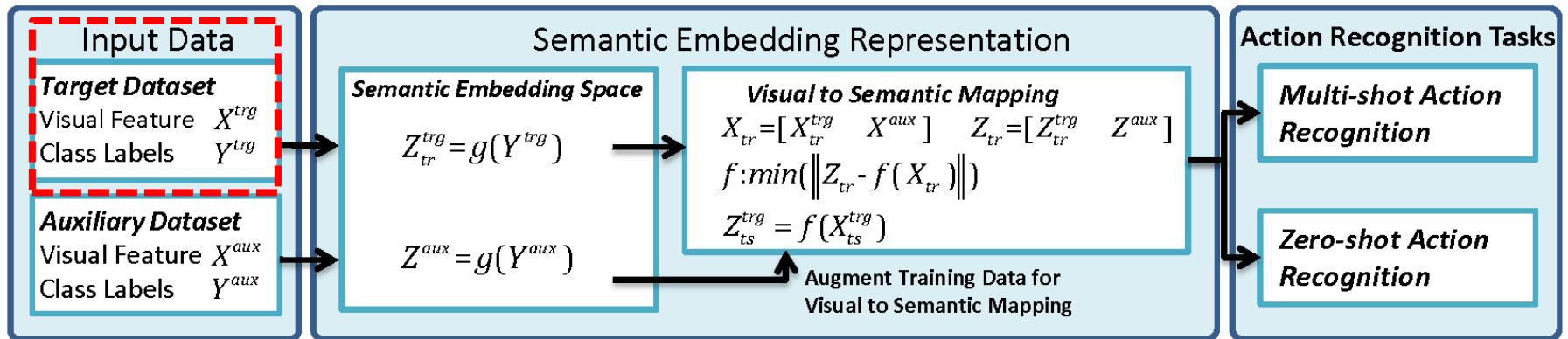
- **Domain Shift**



Our Solution



Our Solution

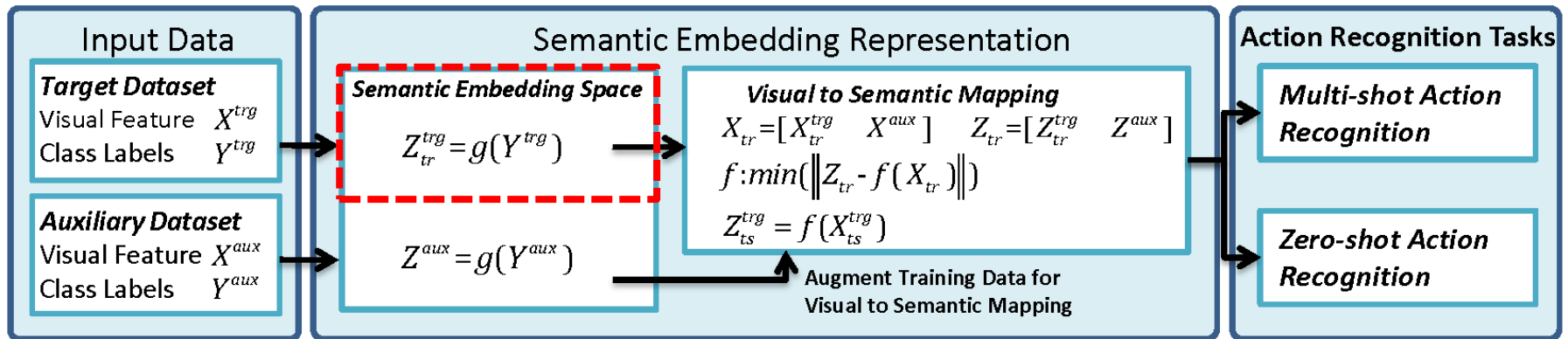


Low-Level Visual Feature

- Improved Trajectory Feature for \mathbf{x}



Our Solution



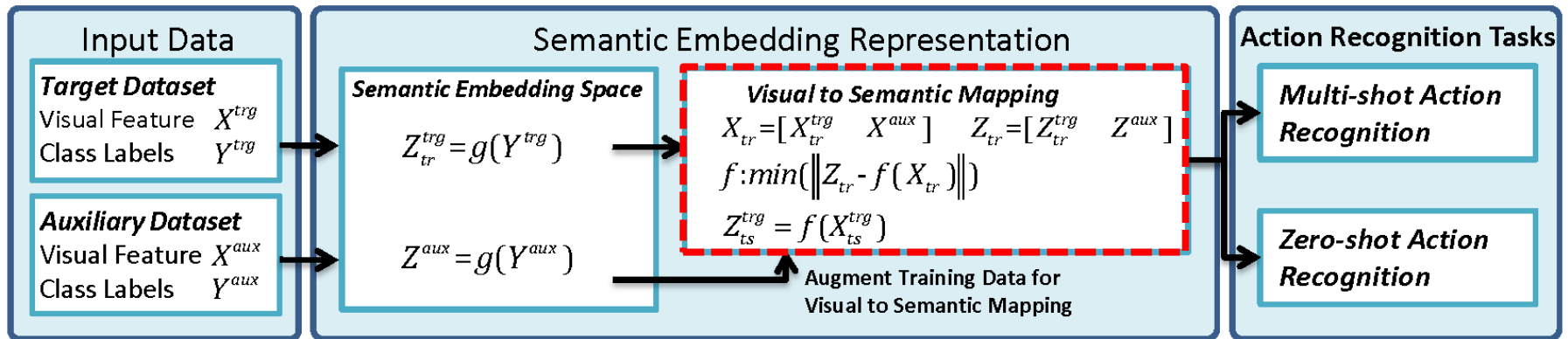
Combinations of Multi Words

- A phrase is constructed from single word vectors

Additive Composition

$$\text{vec}(\text{"Apply Eye Makeup"}) = \text{vec}(\text{"Apply"}) + \text{vec}(\text{"Eye"}) + \text{vec}(\text{"Makeup"})$$
$$\text{vec}(\text{"Brushing Teeth"}) = \text{vec}(\text{"Brushing"}) + \text{vec}(\text{"Teeth"})$$
$$\text{vec}(\text{"Playing Guitar"}) = \text{vec}(\text{"Playing"}) + \text{vec}(\text{"Guitar"})$$

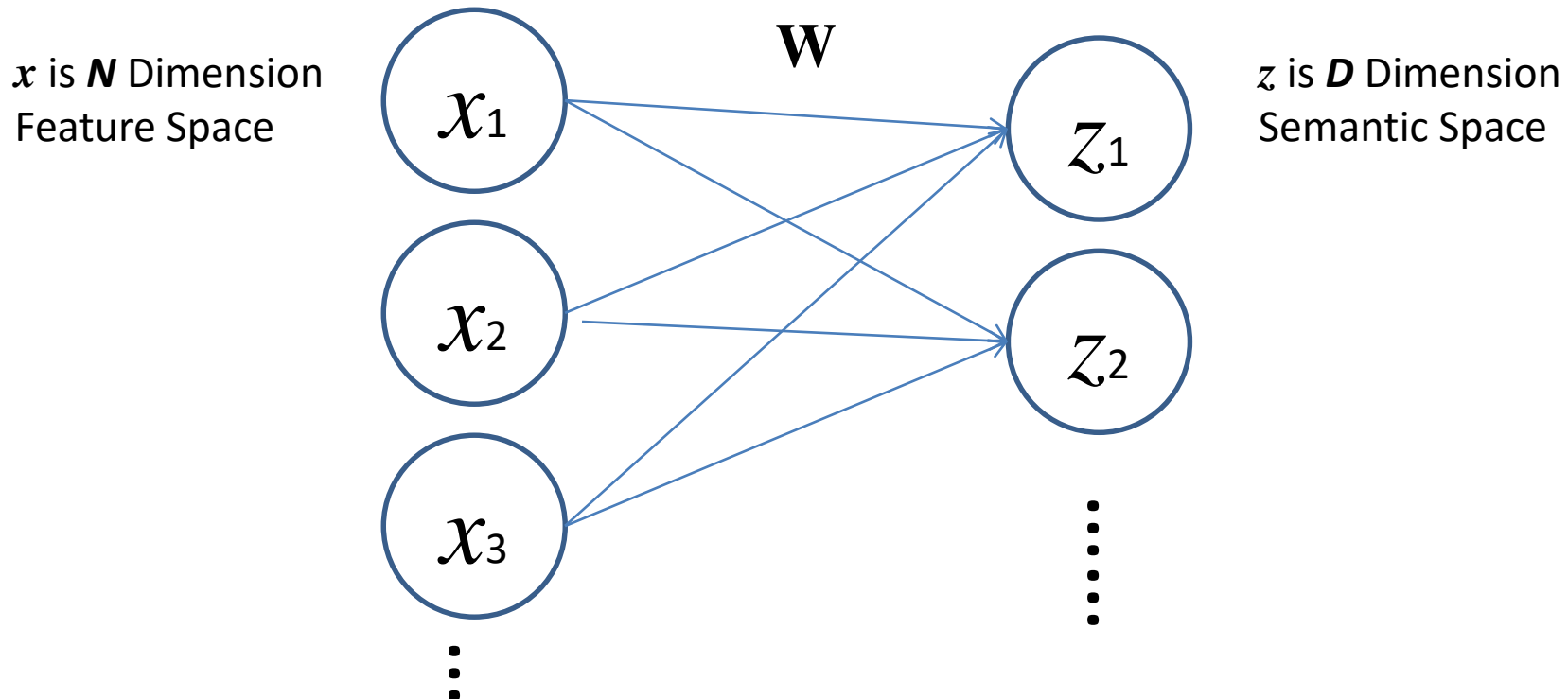
Our Solution



Visual to Semantic Mapping by Regularized Linear Regression

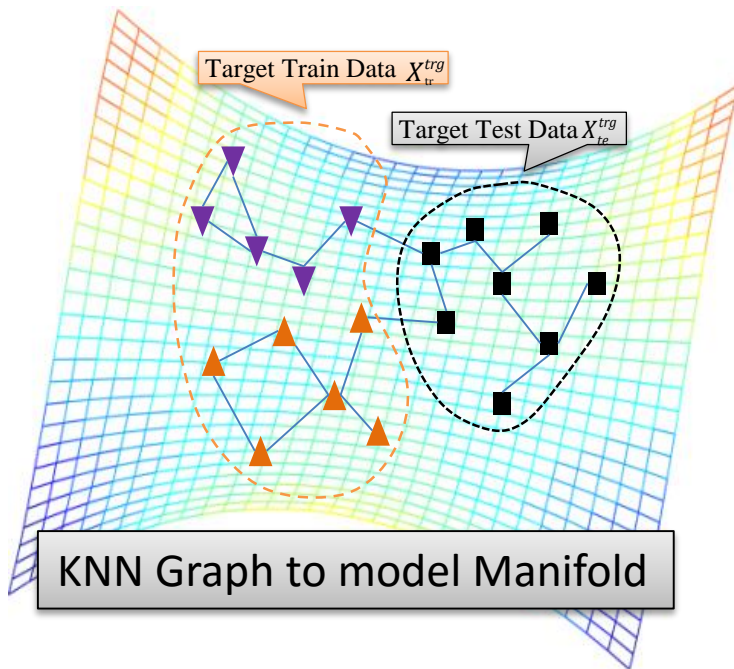
- Multi-Dimensional Regularized Linear Regression

$$\min_{\mathbf{W}} \sum_{i=1}^N \|\mathbf{z}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{W}\|_2^2$$



Domain Shift – Semi Supervised (Manifold Regularized) Regression

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



Manifold Regularizer

Train and Test Data in Feature Space



$$X_{tr} = X_{tr}^{trg}$$

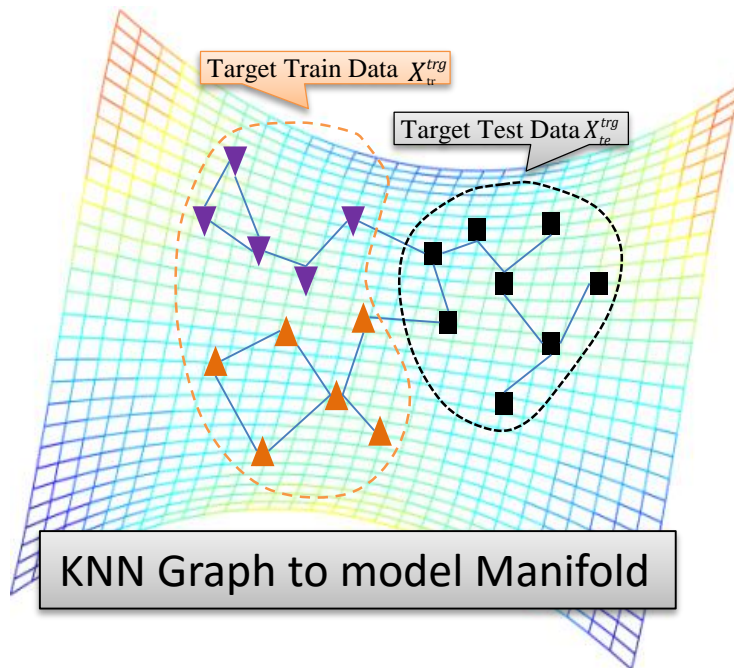
$$X_{te} = X_{te}^{trg}$$

KNN Graph
weight

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

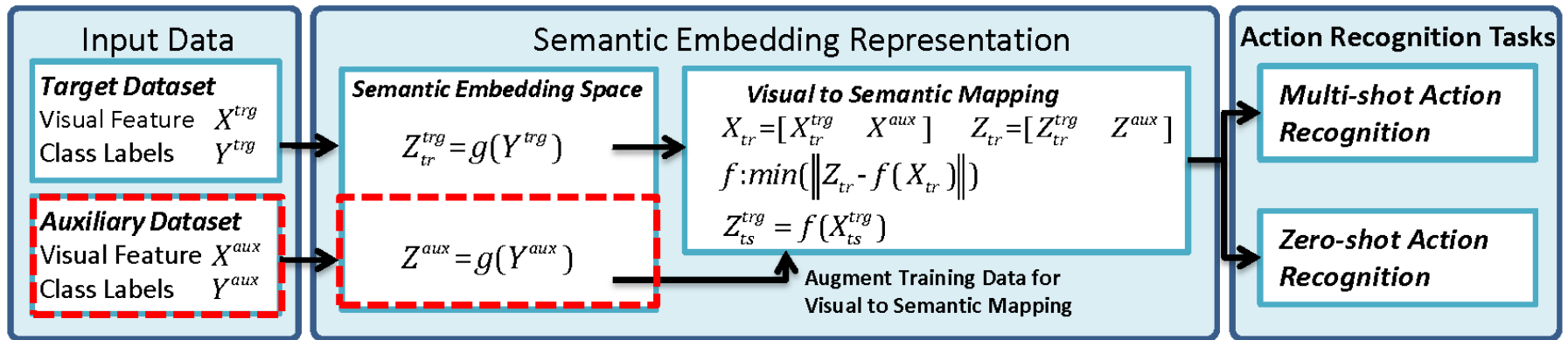
Domain Shift – Semi Supervised (Manifold Regularized) Regression

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



$$\min_{\mathbf{W}} \sum_{i=1}^N \|\mathbf{z}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{W}\|_2^2 + \gamma \sum_{ij} \varpi_{ij} \|\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_j\|_2^2$$

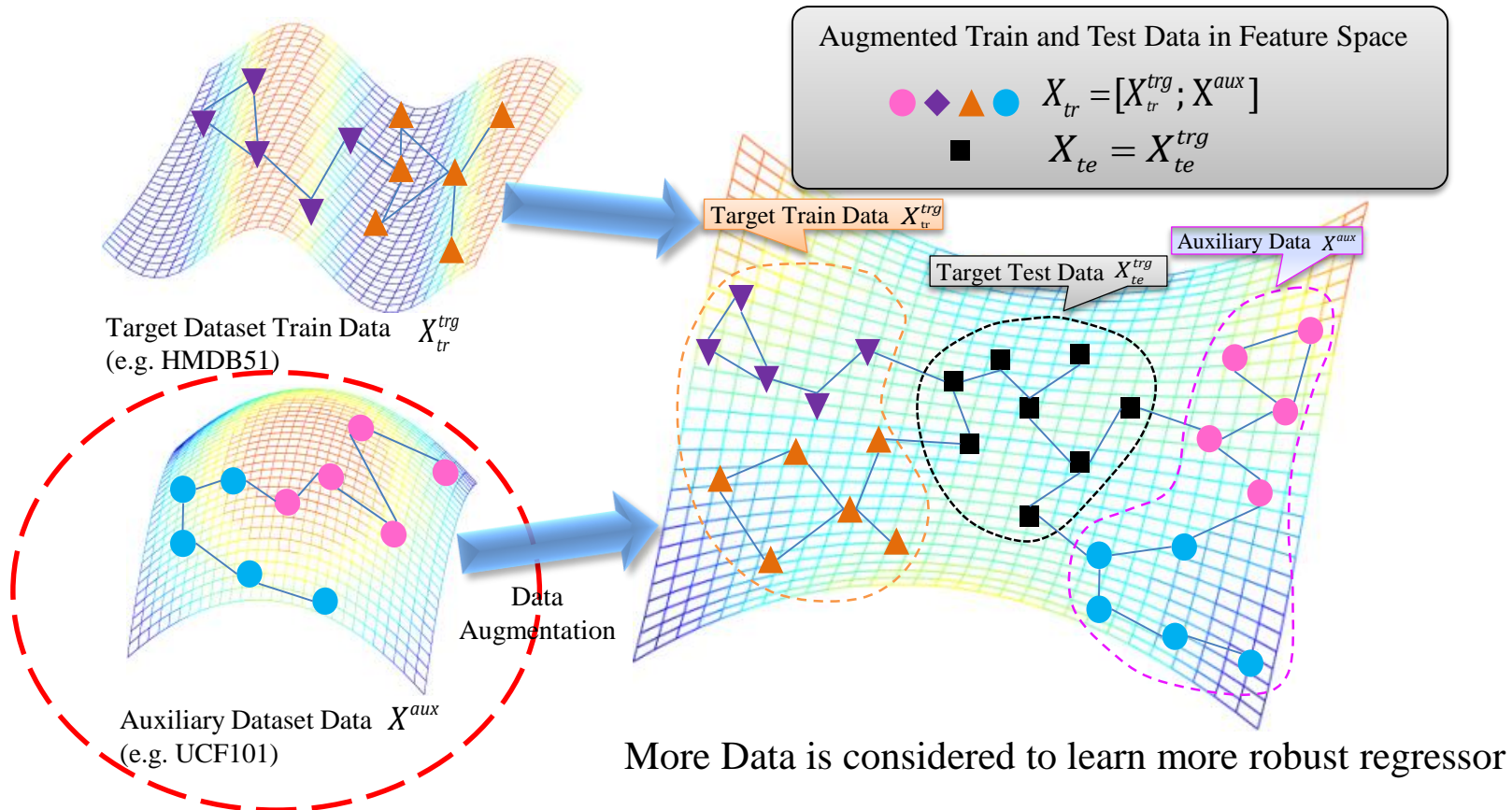
Our Solution



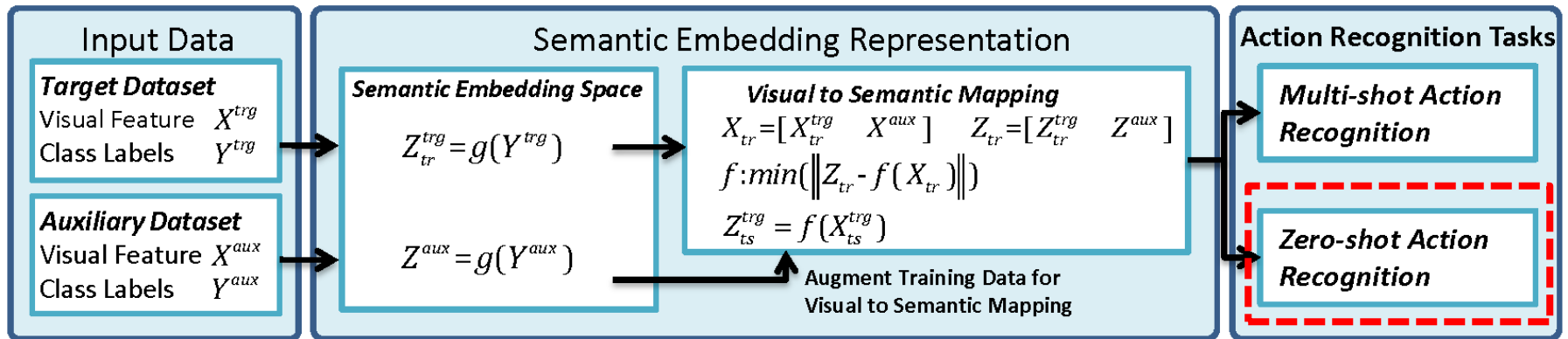
Additional datasets are available

Data Augmentation

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

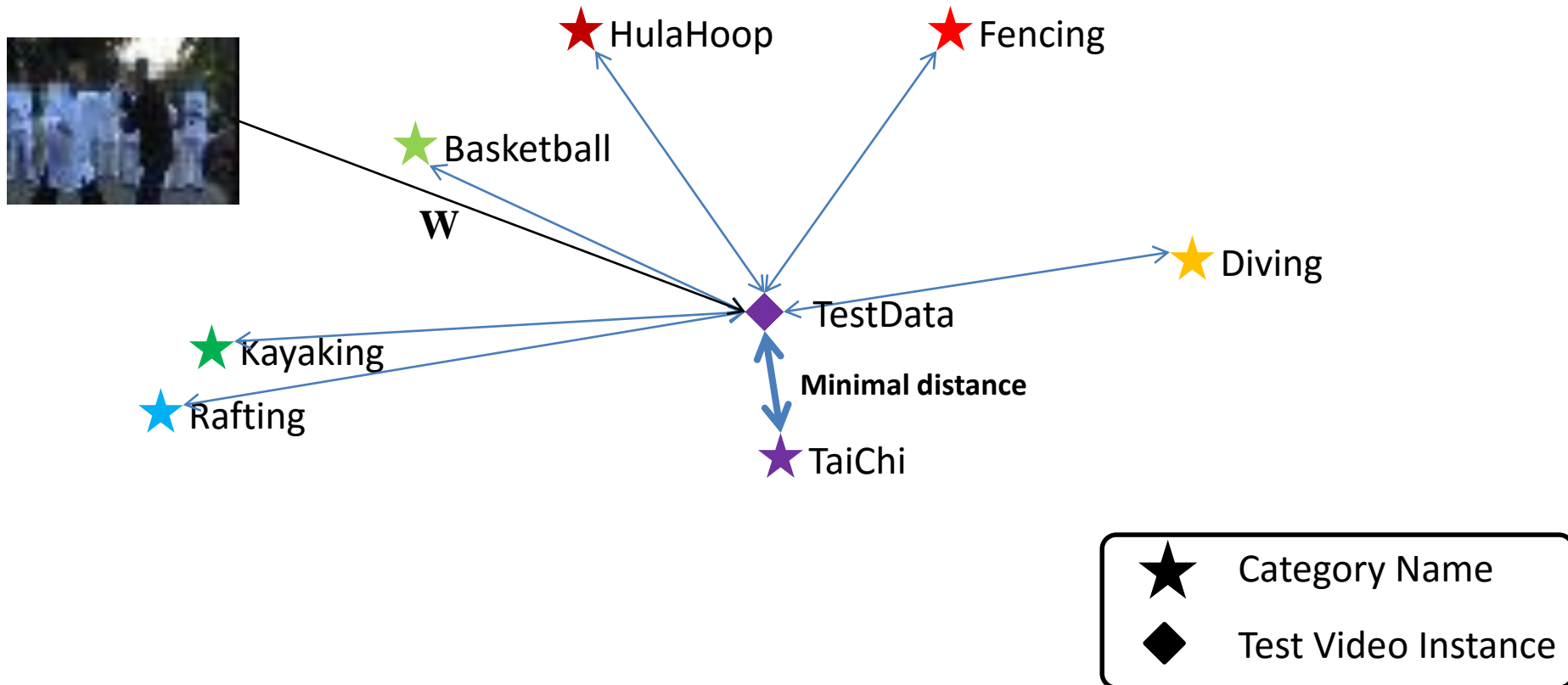


Semantic Word Vector Approach



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

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

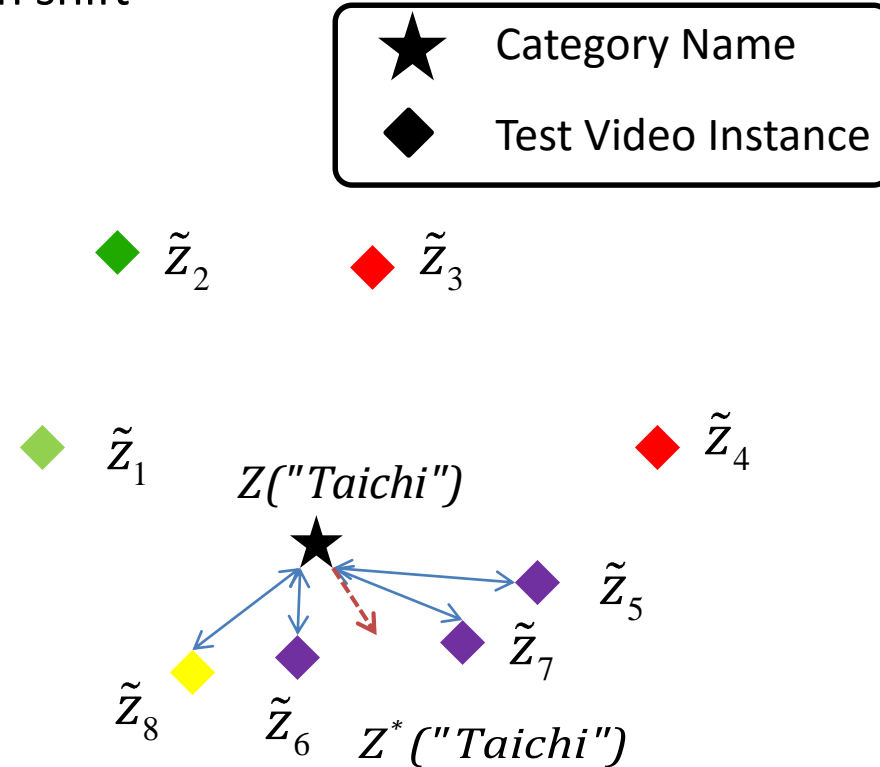
$$Z(\text{"Taichi"}) = g(\text{"Taichi"})$$

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

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

4 NN example

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



Domain Shift – SelfTraining

- Self-training is applied to tackle domain shift

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

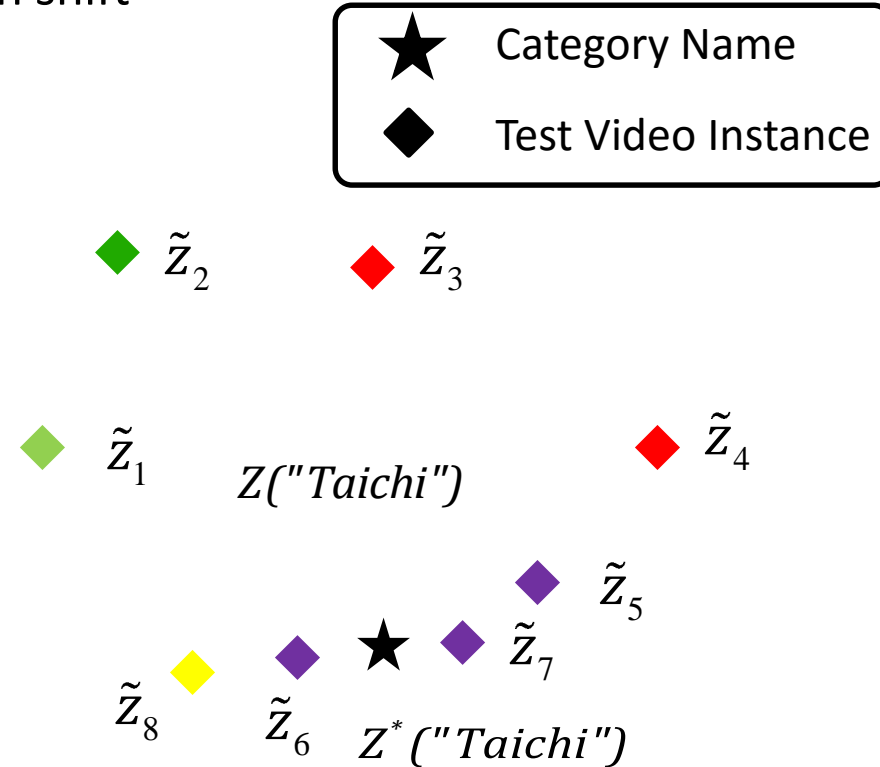
$$Z("Taichi") = g("Taichi")$$

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

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

4 NN example

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



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

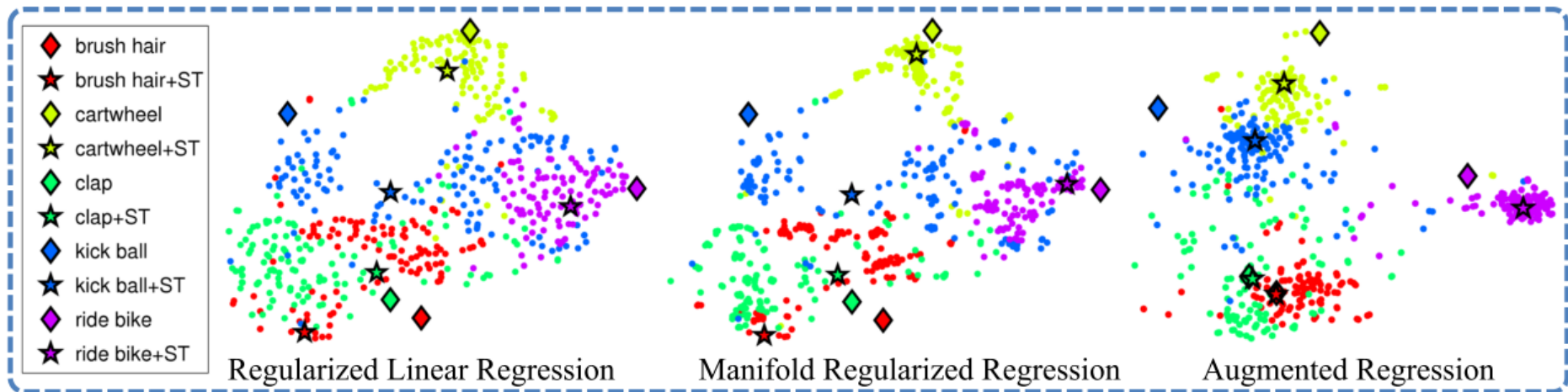
[1] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." *ICCV 2013*.

[2] Perronnin, Florent, Jorge Sánchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification." *ECCV 2010*

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

Model	DA	Trans	Embed	Feat	HMDB51	UCF101	Olympic Sports	CCV	USAA
Random Guess	X	X	X	X	4.0	2.0	12.5	10.0	25.0
RR (Ours)	X	X	W	FV	14.5±2.7	11.7±1.7	35.7±8.8	20.7±3.0	29.5±5.5
MR (Ours)	X	✓	W	FV	19.1±3.8	18.0±2.7	38.6±10.6	22.5±3.4	31.6±3.2
MR (Ours)	✓	✓	W	FV	24.1±3.8	22.1±2.5	43.2±8.3	33.0±4.8	43.3±10.9
SJE (Akata et al, 2015)	X	X	W	FV	12.0±2.6	9.3±1.7	34.6±7.6	16.3±3.1	21.3±0.6
ConSe (Norouzi et al, 2014)	X	X	W	FV	15.0±2.7	11.6±2.1	36.6±9.0	20.7±3.1	28.2±4.8
TMV-BLP (Fu et al, 2014a)*	X	✓	W	SMS	N/A	N/A	N/A	N/A	41.0
TMV-HLP (Fu et al, 2015a)**	X	✓	W	SMS	N/A	N/A	N/A	N/A	43.0
SVE (Xu et al, 2015)	X	X	W	BoW	12.9±2.3	11.0±1.8	N/A	N/A	N/A
RR (Ours)	X	X	A	FV	N/A	12.6±1.8	51.7±11.3	N/A	44.2±13.9
MR (Ours)	X	✓	A	FV	N/A	20.2±2.2	53.5±11.9	N/A	51.6±10.0
DAP (Lampert et al, 2014)	X	X	A	FV	N/A	15.2±1.9	44.4±9.9	N/A	37.9±5.9
IAP (Lampert et al, 2014)	X	X	A	FV	N/A	15.6±2.2	44.0±10.7	N/A	31.7±1.6
HAA (Liu et al, 2011)	X	X	A	FV	N/A	14.3±2.0	48.3±10.2	N/A	41.2±9.8
PST (Rohrbach et al, 2013)	X	✓	A	FV	N/A	15.3±2.2	48.6±11.0	N/A	47.9±10.6
M2LATM (Fu et al, 2014b)	X	✓	A	SMS	N/A	N/A	N/A	N/A	41.9
TMV-BLP (Fu et al, 2014a)*	X	✓	A	SMS	N/A	N/A	N/A	N/A	40.0
TMV-HLP (Fu et al, 2015a)**	X	✓	A	SMS	N/A	N/A	N/A	N/A	42.0
UDA (Kodirov et al, 2015)	X	✓	A	FV	N/A	13.2±1.9	N/A	N/A	N/A
MR (Ours)	X	✓	A+W	FV	N/A	20.8±2.3	53.2±11.6	N/A	51.9±10.1
TMV-BLP (Fu et al, 2014a)	X	✓	A+W	SMS	N/A	N/A	N/A	N/A	47.8
UDA (Kodirov et al, 2015)	X	✓	A+W	FV	N/A	14.0±1.8	N/A	N/A	N/A
TMV-HLP (Fu et al, 2015a)	X	✓	A+W	SMS	N/A	N/A	N/A	N/A	50.4

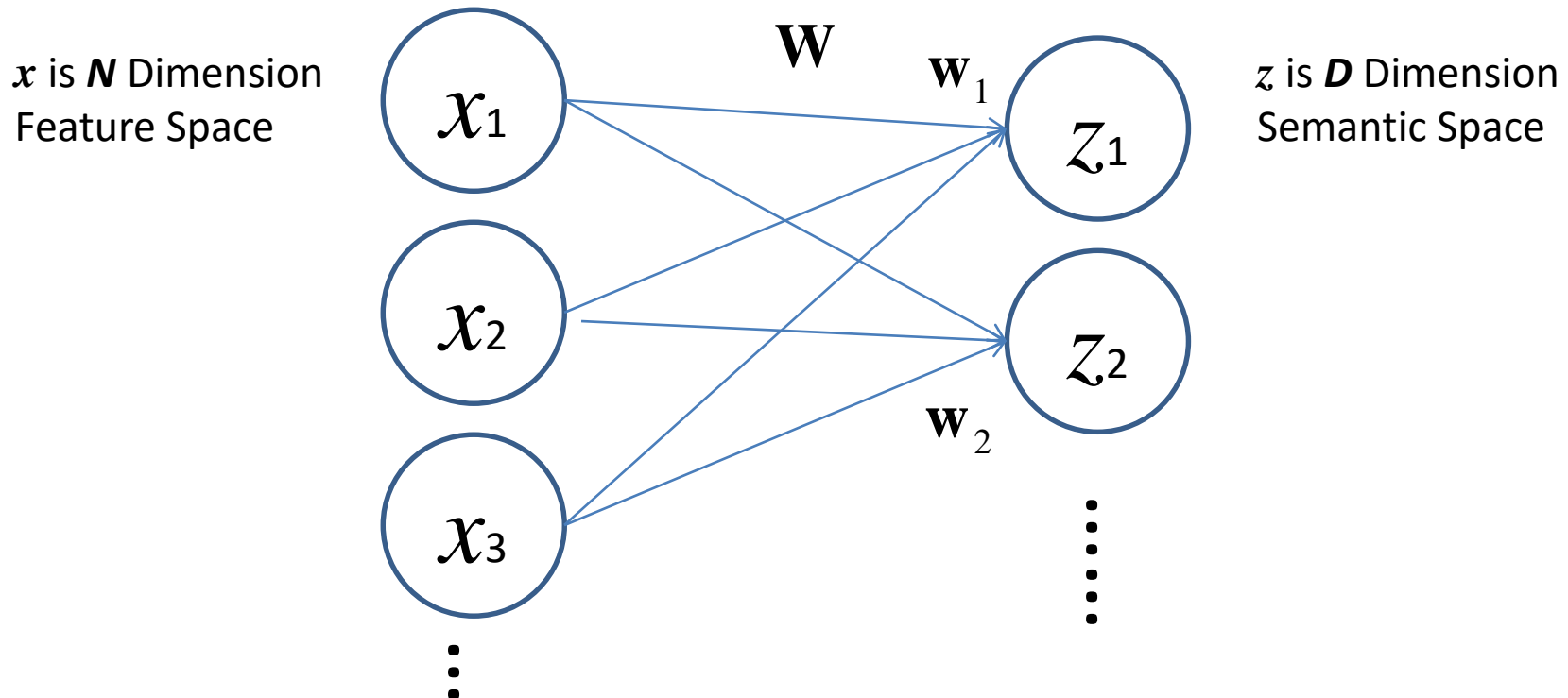
Outline

- Background
- Transductive Zero-Shot Action Recognition
- **Multi-Task Zero-Shot Embedding**
- Zero-Shot Crowd Analysis

Revisit Visual-to-Semantic Mapping

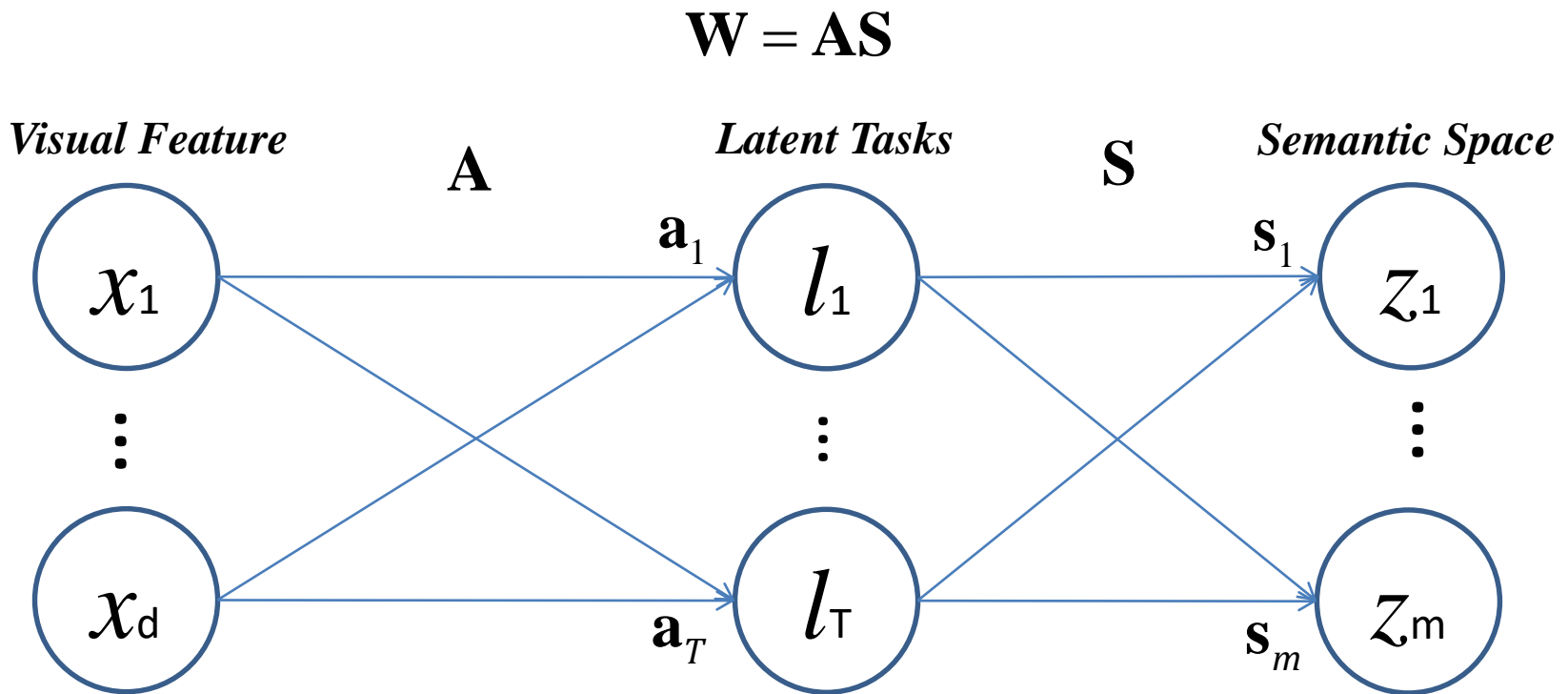
- Multi-Dimensional Regularized Linear Regression

$$\min_{\mathbf{W}} \sum_{i=1}^N \|\mathbf{z}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{W}\|_2^2$$



Visual-to-Semantic Mapping by Multi-Task Regression

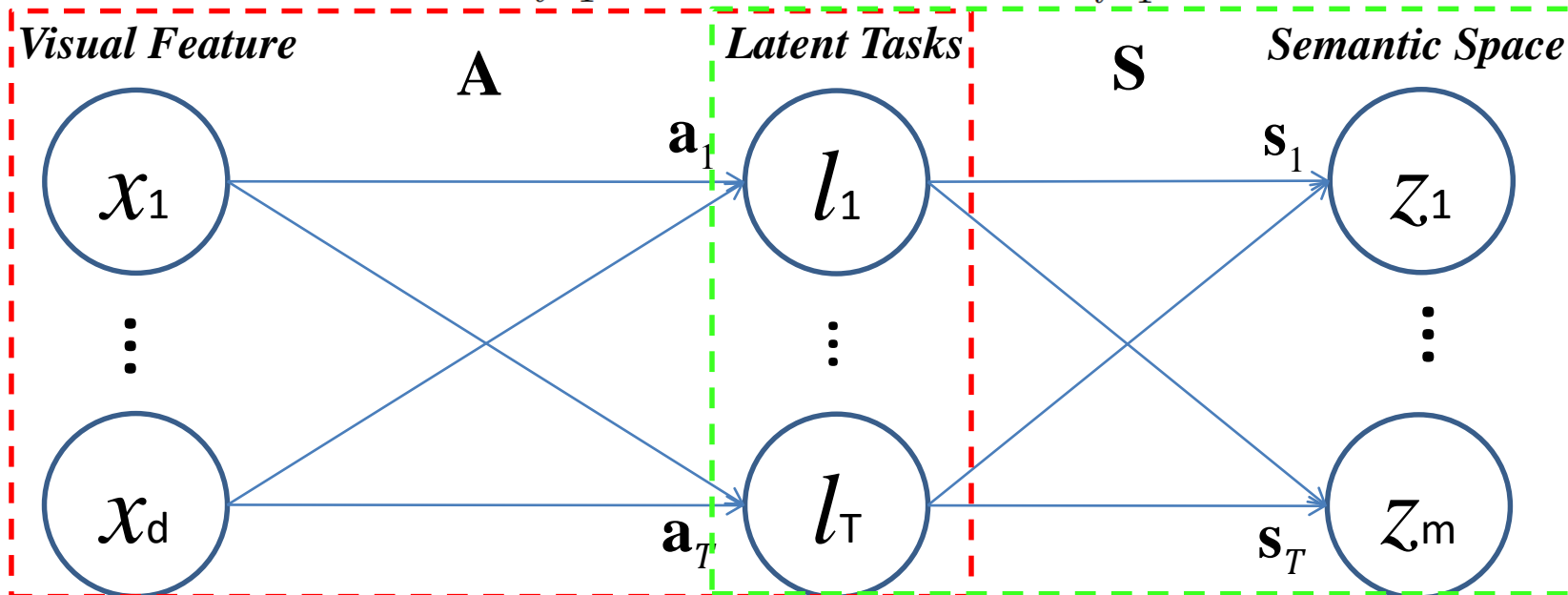
- Two stage regression



Visual-to-Semantic Mapping by Multi-Task Regression

- Two stage regression

$$\min_{\{\mathbf{s}_t\}, \mathbf{A}, \{\mathbf{l}_i\}} \sum_{t=1}^T \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} \left(\|\mathbf{z}_{t,i} - \mathbf{s}_t \mathbf{l}_i\|_2^2 + \|\mathbf{l}_i - \mathbf{A} \mathbf{x}_i\|_2^2 \right) + \lambda_S \sum_{t=1}^T \|\mathbf{s}_t\|_2^2 + \lambda_A \|\mathbf{A}\|_F^2 + \lambda_L \sum_{i=1}^{n_x^{tr}} \|\mathbf{l}_i\|_2^2$$



Visual-to-Semantic Mapping by Multi-Task Regression

- Solve efficiently

Loss Function

$$\min_{\{\mathbf{s}_t\}, \mathbf{A}, \{\mathbf{l}_i\}} \sum_{t=1}^T \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} (\|\mathbf{z}_{t,i} - \mathbf{s}_t \mathbf{l}_i\|_2^2 + \|\mathbf{l}_i - \mathbf{A} \mathbf{x}_i\|_2^2) + \lambda_S \sum_{t=1}^T \|\mathbf{s}_t\|_2^2 + \lambda_A \|\mathbf{A}\|_F^2 + \lambda_L \sum_{i=1}^{n_x^{tr}} \|\mathbf{l}_i\|_2^2$$

Iterative Update

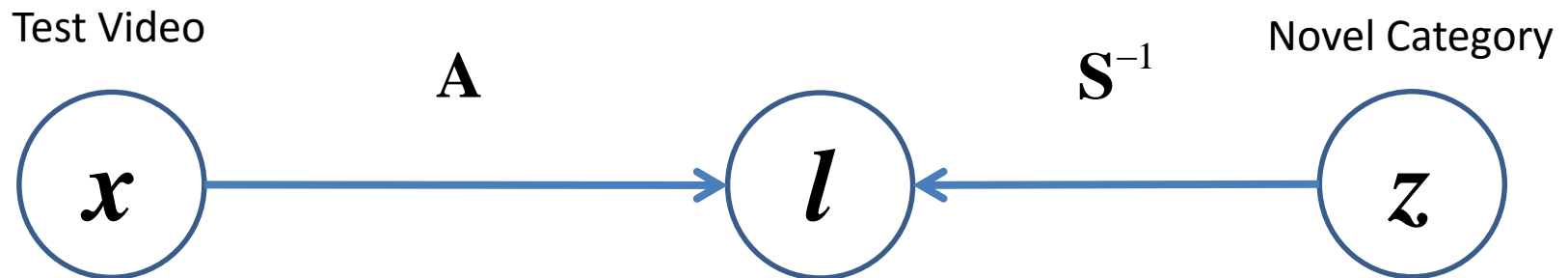
$$\mathbf{L} = (\mathbf{S}^T \mathbf{S} + (\lambda_L n_x^{tr} + 1) \mathbf{I})^{-1} (\mathbf{S}^T \mathbf{Z} + \mathbf{A} \mathbf{X})$$

$$\mathbf{S} = \mathbf{Z} \mathbf{L}^T (\mathbf{L} \mathbf{L}^T + \lambda_S n_x^{tr} \mathbf{I})^{-1}$$

$$\mathbf{A} = \mathbf{L} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \lambda_A n_x^{tr} \mathbf{I})^{-1}$$

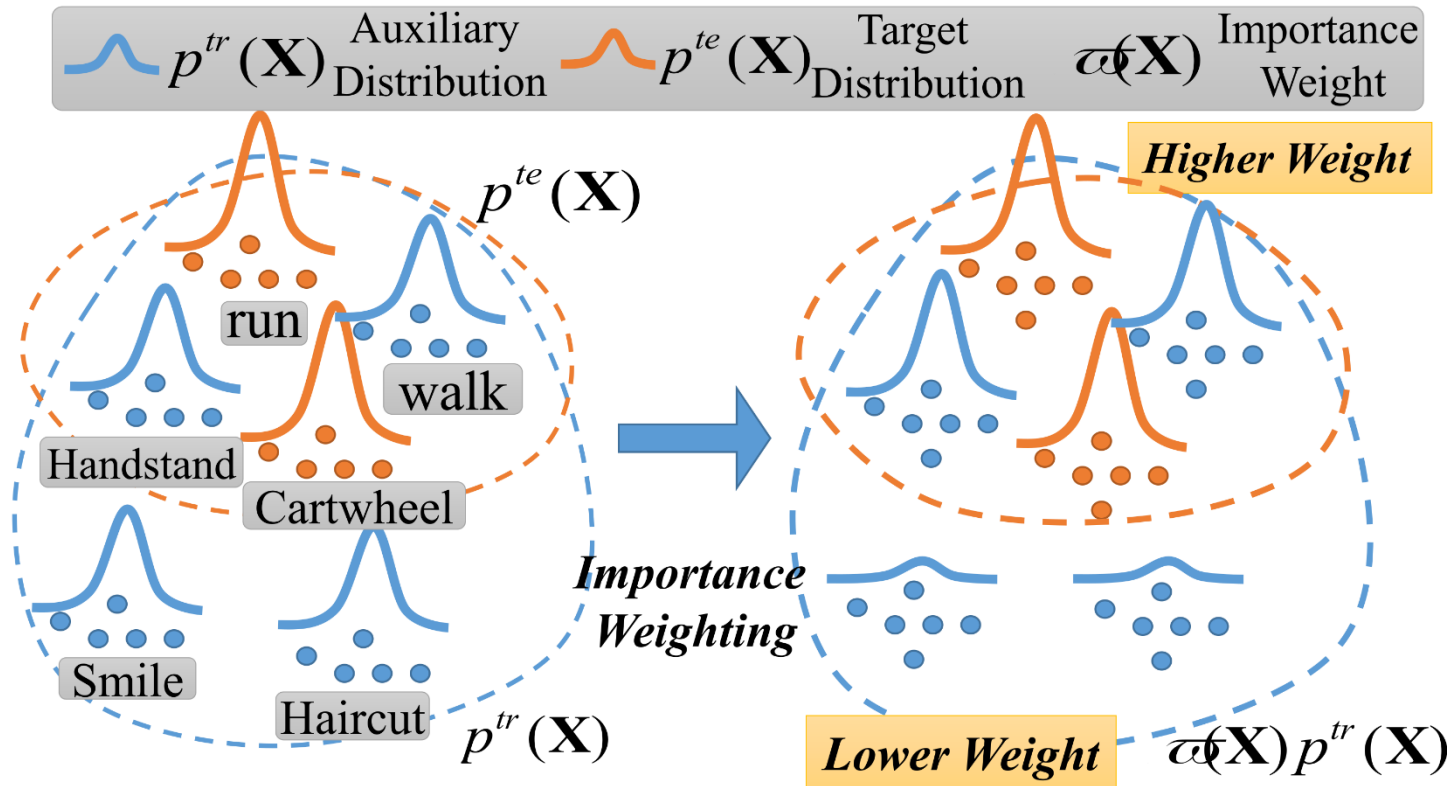
Multi-Task Embedding

- Lower dimension subspace embedding



$$\mathbf{z}^* = \underset{\mathbf{z}}{\operatorname{argmin}} \left\| \mathbf{A}\mathbf{x} - \mathbf{S}^{-1}\mathbf{z} \right\|_2^2$$

Importance Weighting for Domain Adaptation



$$\min_{\omega} D_{KL}(p^{te}(\mathbf{x}) \mid \omega(\mathbf{x}) p^{tr}(\mathbf{x})) = \int p^{te}(\mathbf{x}) \log \frac{p^{te}(\mathbf{x})}{\omega(\mathbf{x}) p^{tr}(\mathbf{x})} d\mathbf{x}$$

Revisit Visual-to-Semantic Mapping

- Uniform weight is given to all training examples

Uniform Model

$$\min_{\mathbf{W}} \sum_{i=1}^N \|\mathbf{z}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{W}\|_2^2$$

Weighted Model

$$\min_{\mathbf{W}} \sum_{i=1}^N \varpi_i \|\mathbf{z}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{W}\|_2^2$$

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

[1] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." *ICCV 2013*.

[2] Perronnin, Florent, Jorge Sánchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification." *ECCV 2010*

MTL v.s. STL

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

[1] Xu, X., et al. "Transductive Zero-Shot Action Recognition by Word-Vector Embedding." *IJCV* 2017

[2] Evgeniou, A., et al. "Regularized multi-task learning." *ACM SIGKDD* 2004

[3] Kumar, A., et al. "Learning Task Grouping and Overlap in Multi-task Learning." *ICML* 2012

Importance Weighting

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

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

	Method	Embed	Feat	TD	Aug	HMDB51	UCF101	Olympic Sports
Ours	MTE	W	FV	X	X	19.7±1.6	15.8±1.3	44.3±8.1
	MTE+Full KLIEP	W	FV	✓	✓	23.9±3.0	21.9±2.7	52.3±8.1
	MTE+Full KLIEP+PP	W	FV	✓	✓	24.8±2.2	22.9±3.3	56.6±7.7
	MTE	A	FV	X	X	N/A	18.3±1.7	55.6±11.3
State-of-the-art models	DAP [1] CVPR09	A	FV	X	X	N/A	15.9±1.2	45.4±12.8
	IAP [1] CVPR09	A	FV	X	X	N/A	16.7±1.1	42.3±12.5
	HAA [2] CVPR11	A	FV	X	X	N/A	14.9±0.8	46.1±12.4
	SVE [3] ICIP15	W	BoW	X	X	14.9±1.8	12.0±1.4	N/A
	SVE [3] ICIP15	W	BoW	✓	✓	22.8±2.6	18.4±1.4	N/A
	ESZSL [4] ICML15	W	FV	X	X	18.5±2.0	15.0±1.3	39.6±9.6
	ESZSL [4] ICML15	A	FV	X	X	N/A	17.1±1.2	53.9±10.8
	SJE [5] ICCV15	W	FV	X	X	13.3±2.4	9.9±1.4	28.6±4.9
	SJE [5] ICCV15	A	FV	X	X	N/A	12.0±1.2	47.5±14.8

[1] Lampert, C., et al. "Learning to detect unseen object classes by between-class attribute transfer." CVPR 2009

[2] Liu, J., et al. "Recognizing human actions by attributes." CVPR 2011

[3] Xu, X., et al. "Semantic embedding space for zero-shot action recognition." ICIP 2015

[4] Romera-paredes, B., Torr, P.H.S. "An embarrassingly simple approach to zero-shot learning." ICML 2015

[5] Akata, Z., et al. "Evaluation of Output Embeddings for Fine-Grained Image Classification." CVPR 2015

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

Violence Detection



Motivation

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



Only 124
positive violent
videos

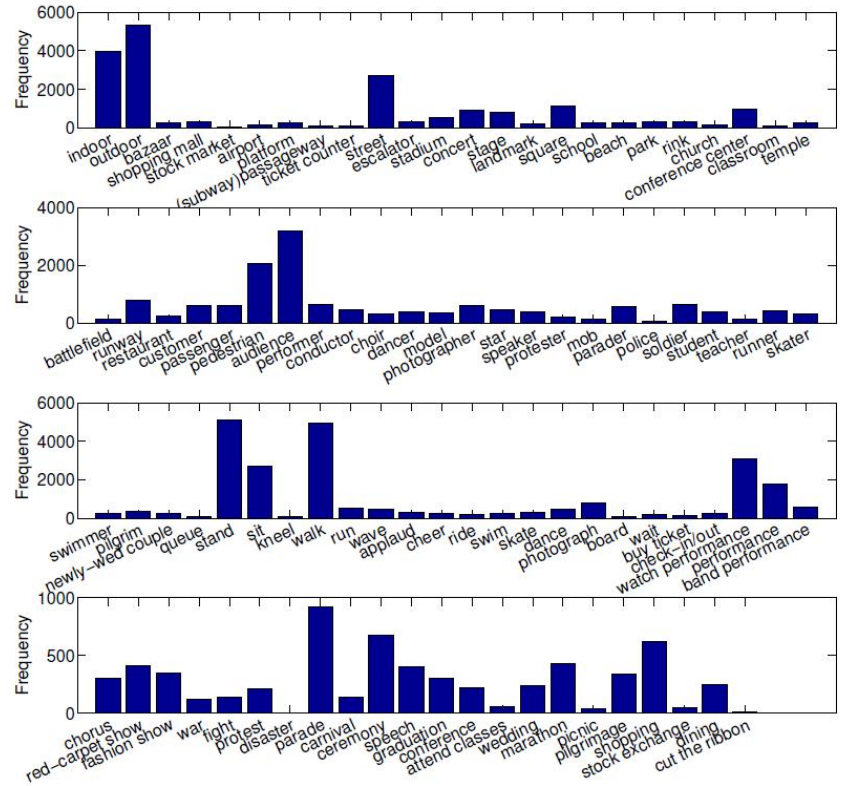
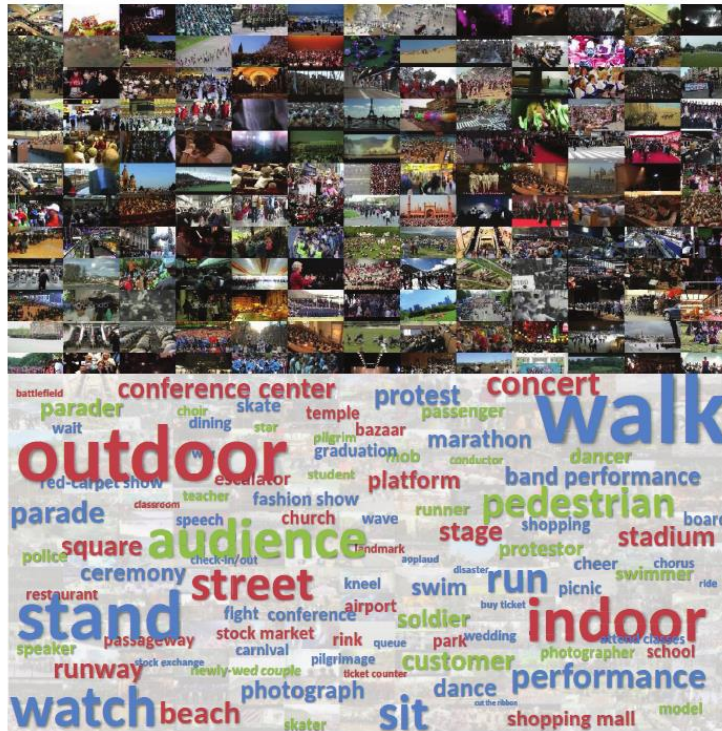
Violent Videos



Non-Violent Videos

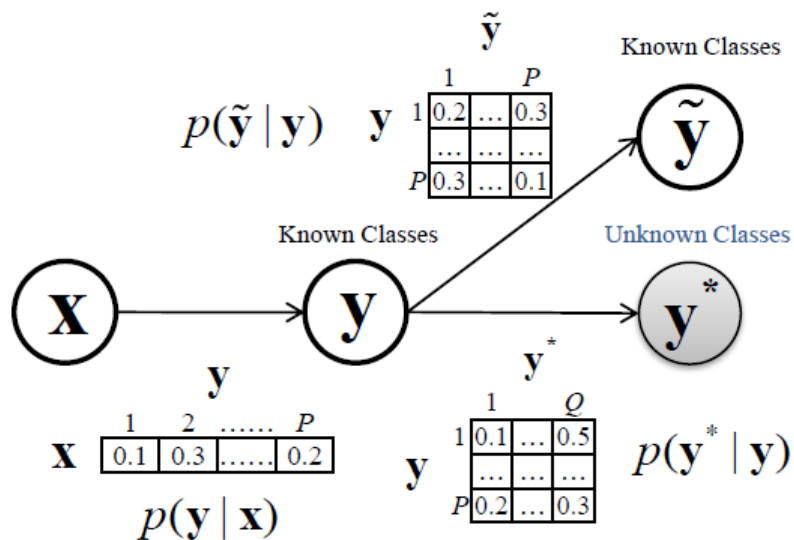
Motivation

- Exploit Existing Crowd Video Data, e.g. WWW Crowd dataset



Zero-Shot Predict Crowd Behaviour

- Predict “Violence” in Zero-Shot Manner



$$p(y_q^* | y_p) = \frac{\exp(\frac{1}{\gamma} \mathbf{v}_q^\top \mathbf{v}_p^S)}{\sum_{p=1}^P \exp(\frac{1}{\gamma} \mathbf{v}_q^\top \mathbf{v}_p^S)}$$

Challenges

- Semantic relatedness v.s. Visual relatedness

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

$$\text{vec}(\text{“Outdoor”})^T \text{vec}(\text{“Indoor”})=0.7104$$

But Visually Never Co-Occur!!

Solution

- Exploit co-occurrence of labels to improve ZSL

Context of Text Corpus

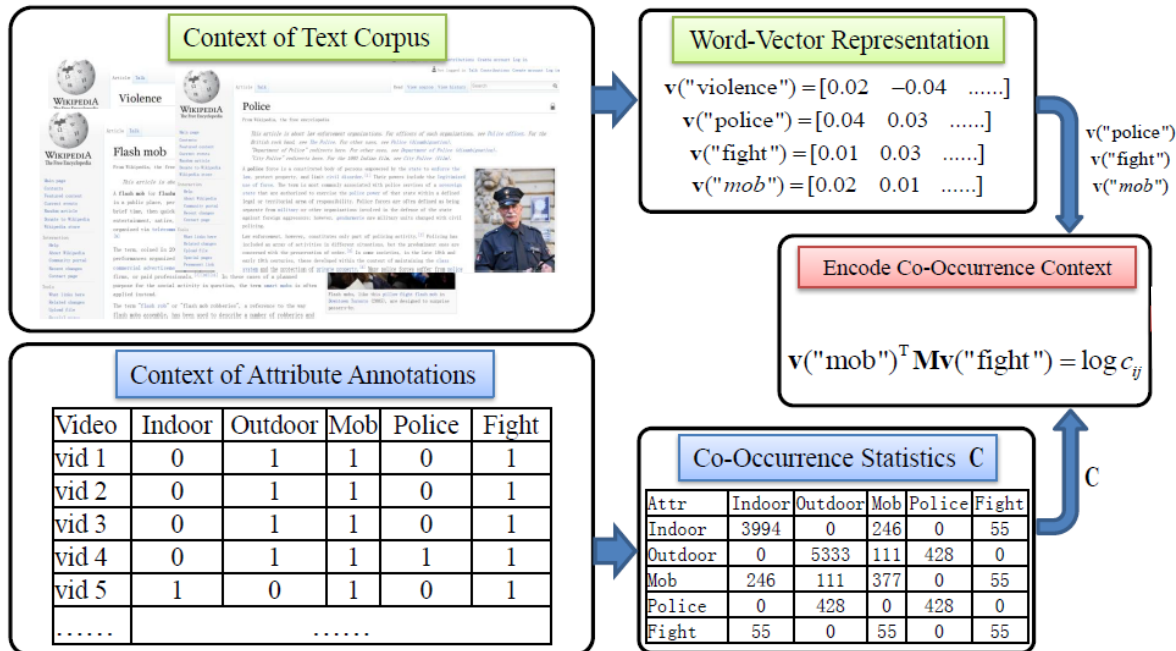
The image shows three Wikipedia article snippets. The 'Violence' snippet includes a table with columns for 'No name', 'Organized', 'Random acts', 'Human events', and 'Human actions'. The 'Flash mob' snippet includes a table with columns for 'Flash mob', 'Flash mob', 'Flash mob', 'Flash mob', and 'Flash mob'. The 'Police' snippet includes a table with columns for 'Police', 'Police', 'Police', 'Police', and 'Police'.

Context of Attribute Annotations

Video	Indoor	Outdoor	Mob	Police	Fight
vid 1	0	1	1	0	1
vid 2	0	1	1	0	1
vid 3	0	1	1	0	1
vid 4	0	1	1	1	1
vid 5	1	0	1	0	1
.....					

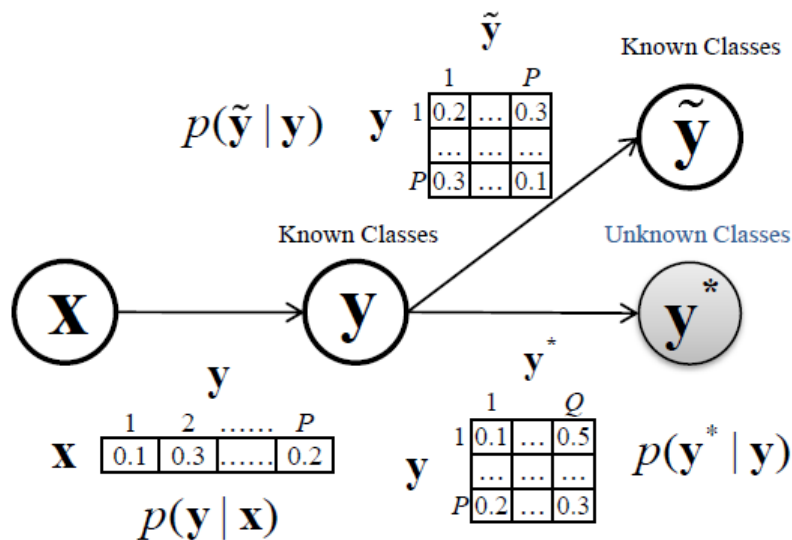
Solution

- Exploit co-occurrence of labels to improve ZSL



Zero-Shot Predict Crowd Behaviour

- Visual Context Aware ZSL



Text Only

$$p(y_q^* | y_p) = \frac{\exp(\frac{1}{\gamma} \mathbf{v}_q^\top \mathbf{v}_p^S)}{\sum_{p=1}^P \exp(\frac{1}{\gamma} \mathbf{v}_q^\top \mathbf{v}_p^S)}$$

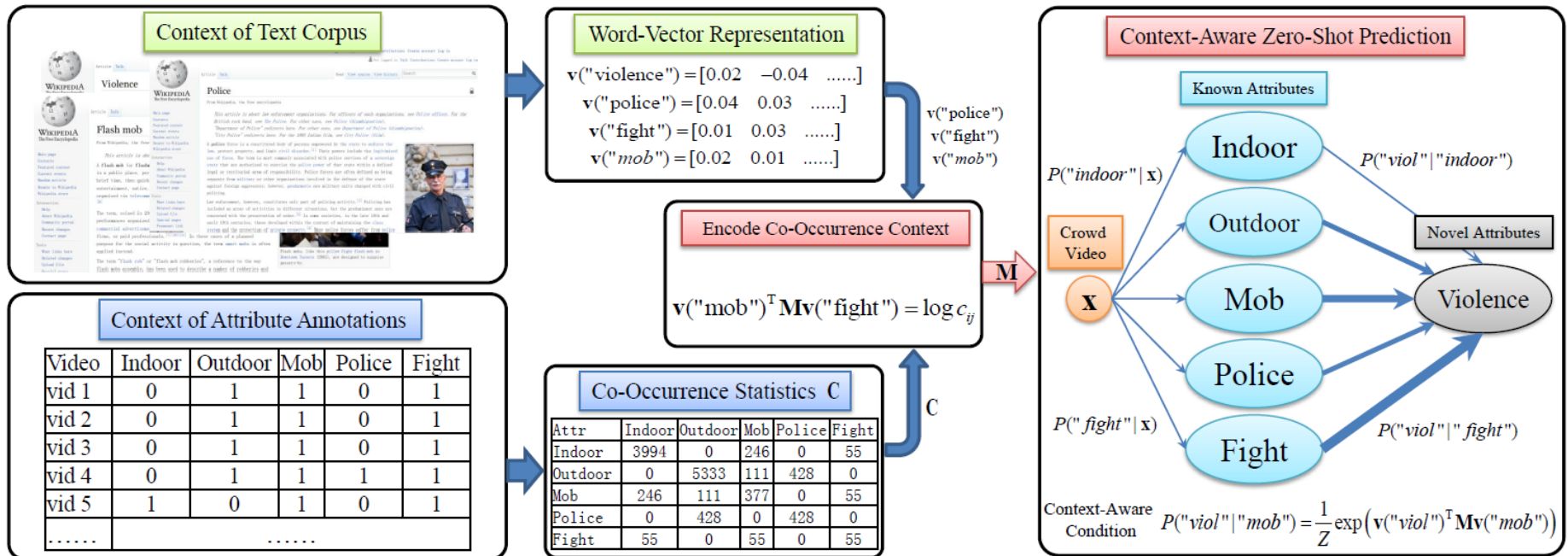


Visual Co-Occurrence

$$p(y_q^* | y_p) = \frac{\exp(\mathbf{v}_q^\top \mathbf{M} \mathbf{v}_p)}{\sum_p \exp(\mathbf{v}_q^\top \mathbf{M} \mathbf{v}_p)}$$

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

[1] Shao, J., et al. "Deeply learned attributes for crowded scene understanding." *CVPR 2015*

[2] Hassner, T., et al. "Violent flows: Real-time detection of violent crowd behavior." *CVPR 2012*

[3] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." *ICCV 2013*.

Performance

- Evaluation on Violence Detection Dataset

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

TexCAZSL uses $M=I$

CoCAZSL learns M from attribute co-occurrence

$$P("viol"|"mob") = \frac{1}{Z} \exp(\mathbf{v}("viol")^T \mathbf{M} \mathbf{v}("mob"))$$

[1] Shao, J., et al. "Deeply learned attributes for crowded scene understanding." CVPR 2015

[2] Romera-paredes, B., Torr, P.H.S. "An embarrassingly simple approach to zero-shot learning." ICML 2015

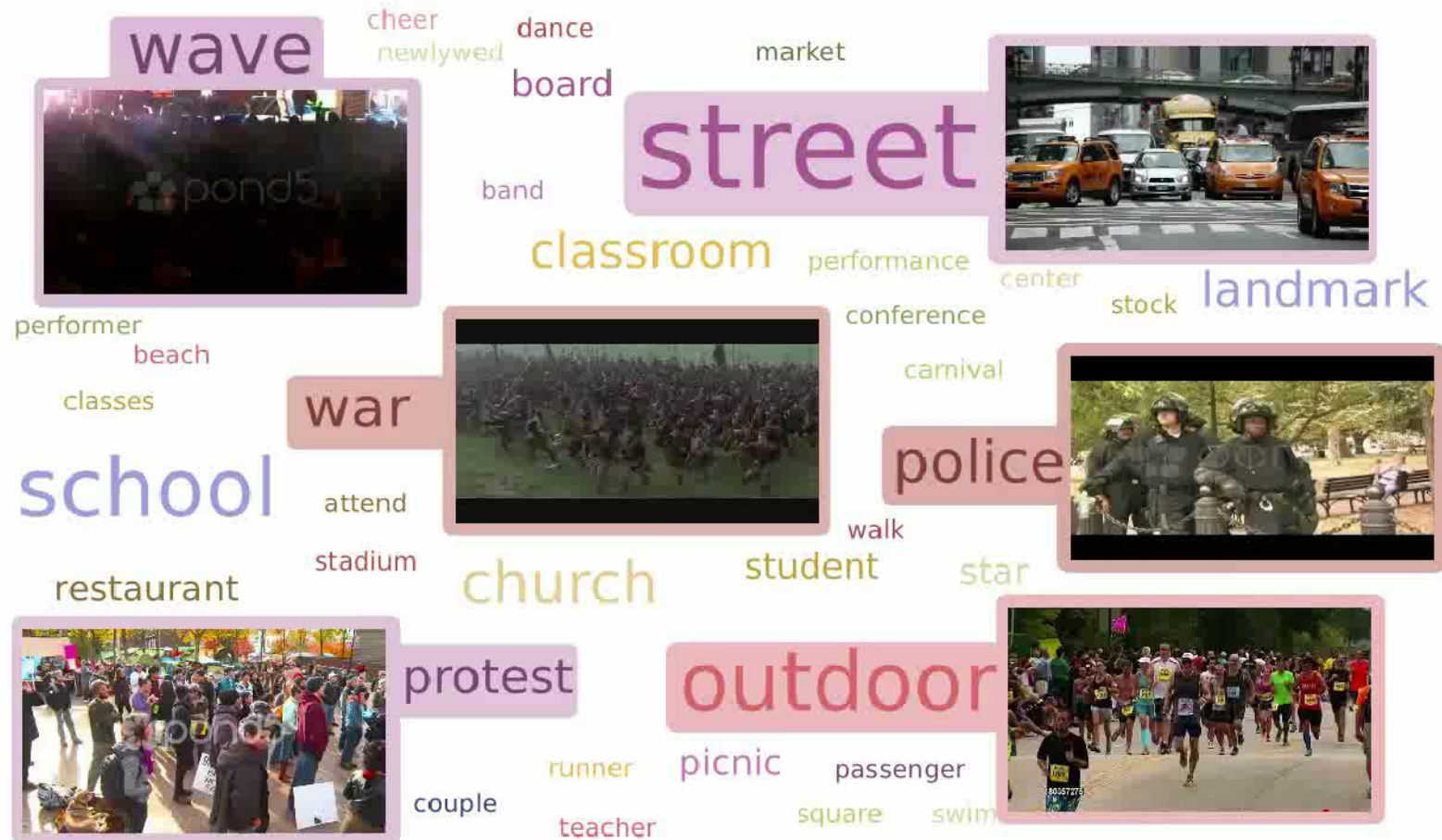
[3] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." ICCV 2013.

[4] Hassner, T., et al. "Violent flows: Real-time detection of violent crowd behavior." CVPR 2012

Qualitative Evaluation

- Relation to “Violence”

$$p(y_q^* | y_p) = \frac{\exp(\mathbf{v}_q^T \mathbf{M} \mathbf{v}_p)}{\sum_p \exp(\mathbf{v}_q^T \mathbf{M} \mathbf{v}_p)}$$



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

