Prof. Sergio A Velastin

(Professor of Applied Computer Vision, Senior Research Scientist, Cortexica)

From Objects to Actions

(Jorge Espinosa, National University Colombia
Fiza Murtaza, Saima Nazir, M.H. Yousaf, Tech Univ Taxila, Pakistan
Huy Hieu Pham, L. Kouhdour, A. Crouzil, U. Paul Sabatier, France)

ICPRS-19, Tours, France, 8^{th} – 10^{th} July 2019

sergio.velastin@ieee.org

CORTEXICA
AI SOLUTIONS FOR BUSINESS

Outline

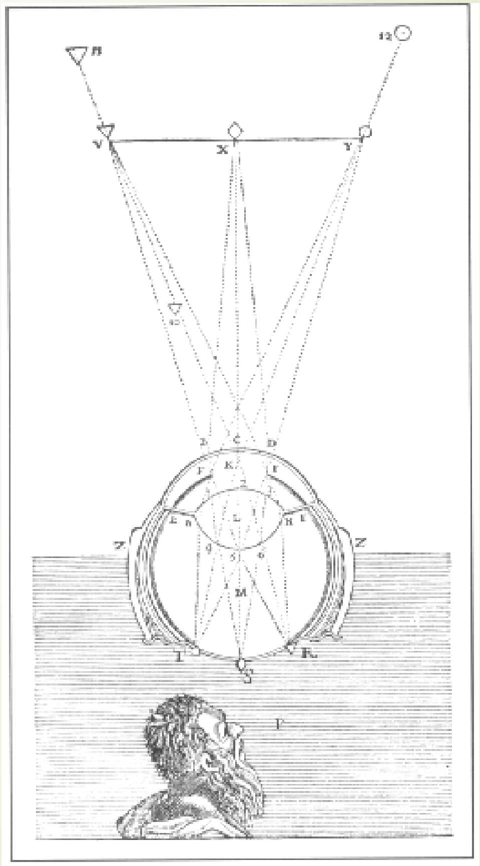
- Motivation
- Transport Applications
- Using RGB-D in semi-open spaces
- Human Action Recognition
- Where Next?

Introduction

- “It is by looking and seeing that we come to know ***what*** is ***where*** in the world”

and ***when*** …

David Marr (1945-1980)



Story telling

Turing's test? ....

CORTEXICA
AI SOLUTIONS FOR BUSINESS

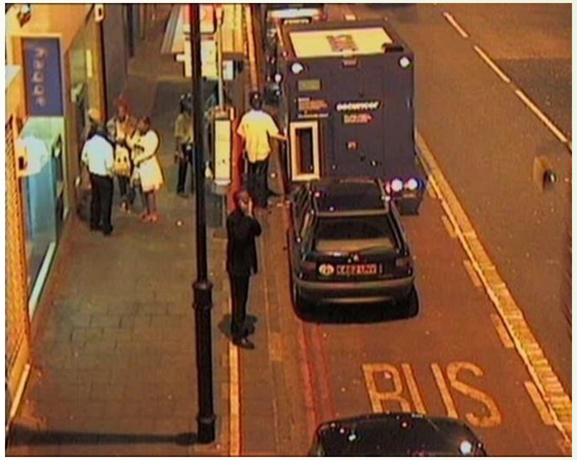
Many Applications …

What does a picture MEAN?

CORTEXICA
AI SOLUTIONS FOR BUSINESS

- Meaning implies **context** and experience (incl. non-visual).
- We are still not sure how to represent and manipulate these.
- Systems more successful when context is implicit/known (engineering?).
- But human activity is very rich!

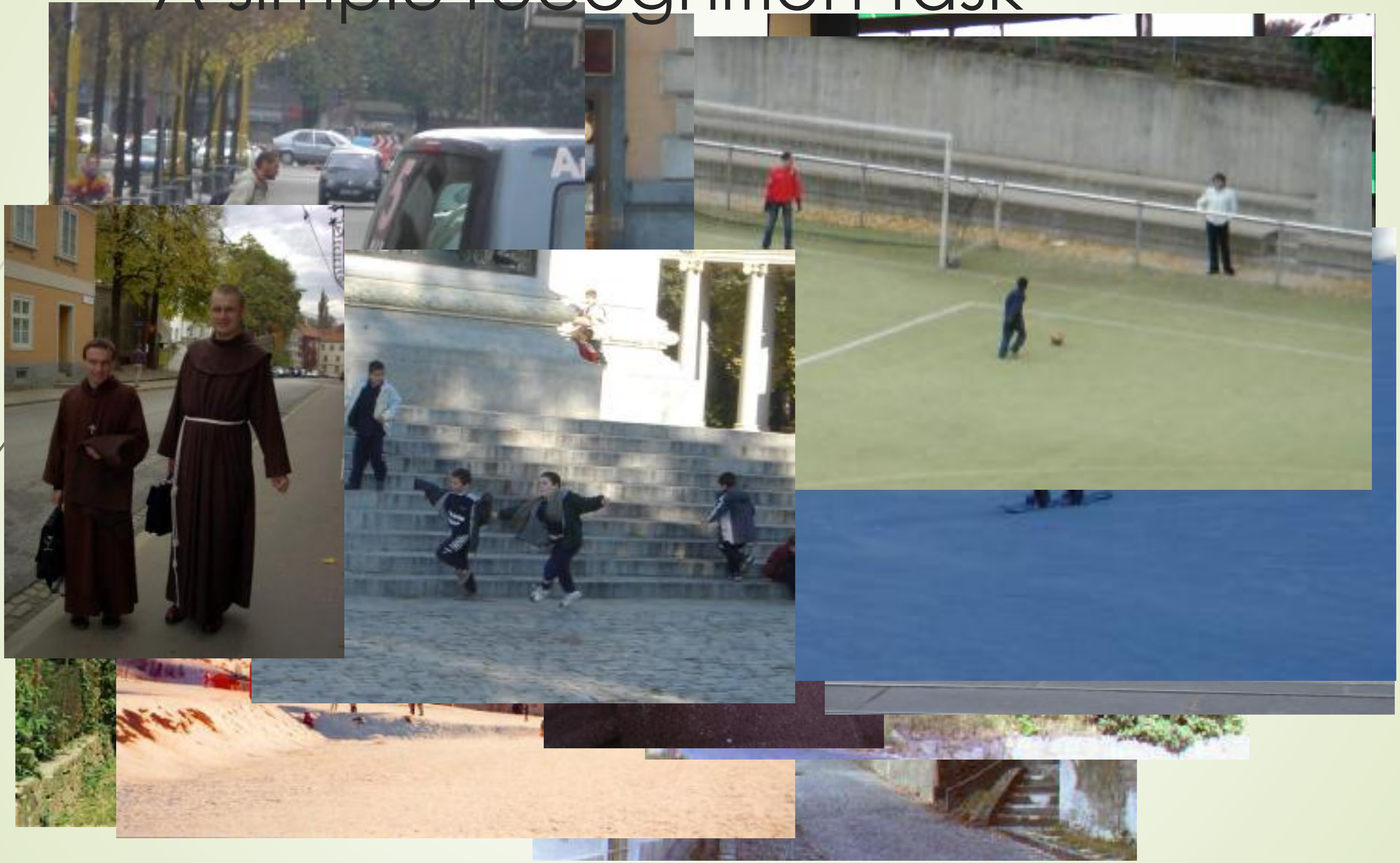
Is one picture worth 1000 words?



So we can think of computer vision as converting visual data to temporal/spatial **narratives**

Not quite there yet, unless we significantly constrain the environment

A simple recognition task



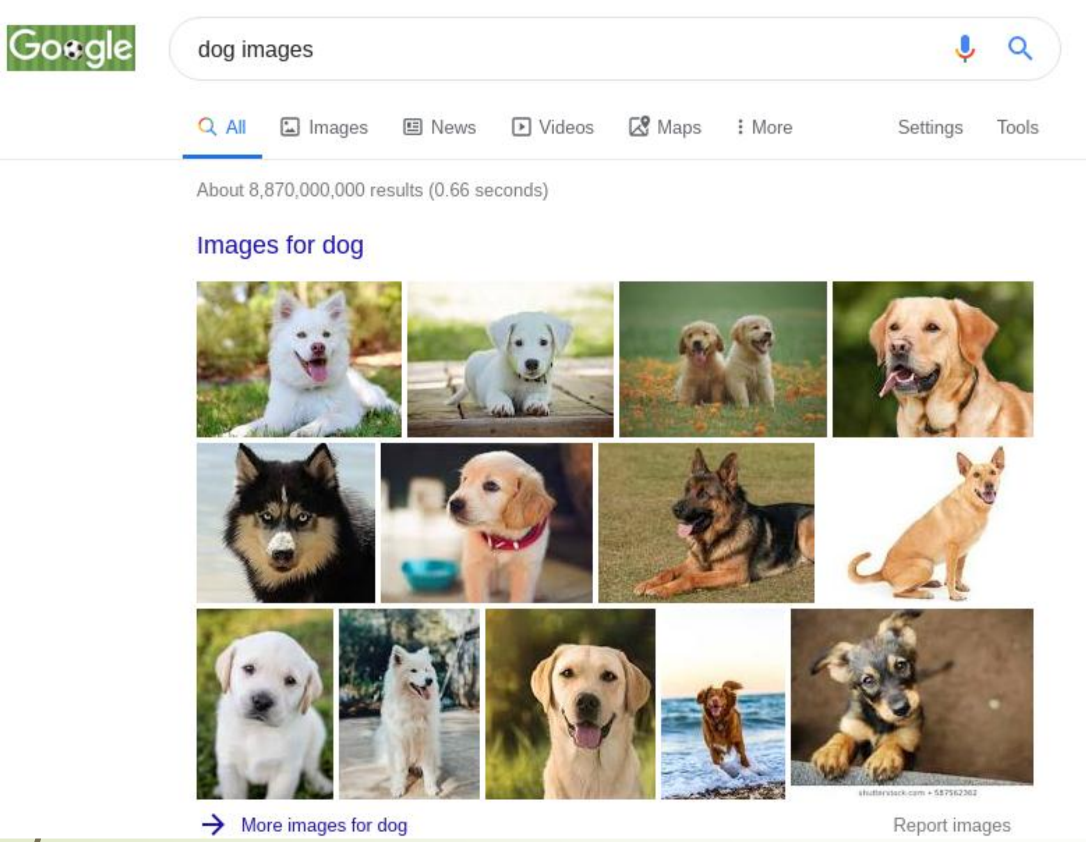
Detection and tracking of people

Oxford Dataset

RBK Dataset

- Multi-scale
- Occlusion

How to recognise “objects”?



- **Internet**: explosion of available *labelled* images/videos (eg. Google search “dog images”
- **Video Games: Very Powerful Graphics Cards (GPUs)** that can do many operations in parallel and very quickly
- **Neural networks**, in particular “Convolutional Neural Networks”:
  - Can reach good accuracy *if* trained with LOTS of *labelled* data
  - GPUs can implement “deep” networks (many layers) able to “generalise” from LOTS of data
- For photos like these, **deep nets** outperform humans

Back to the real world…
Objects and Actions

Environment/Transport

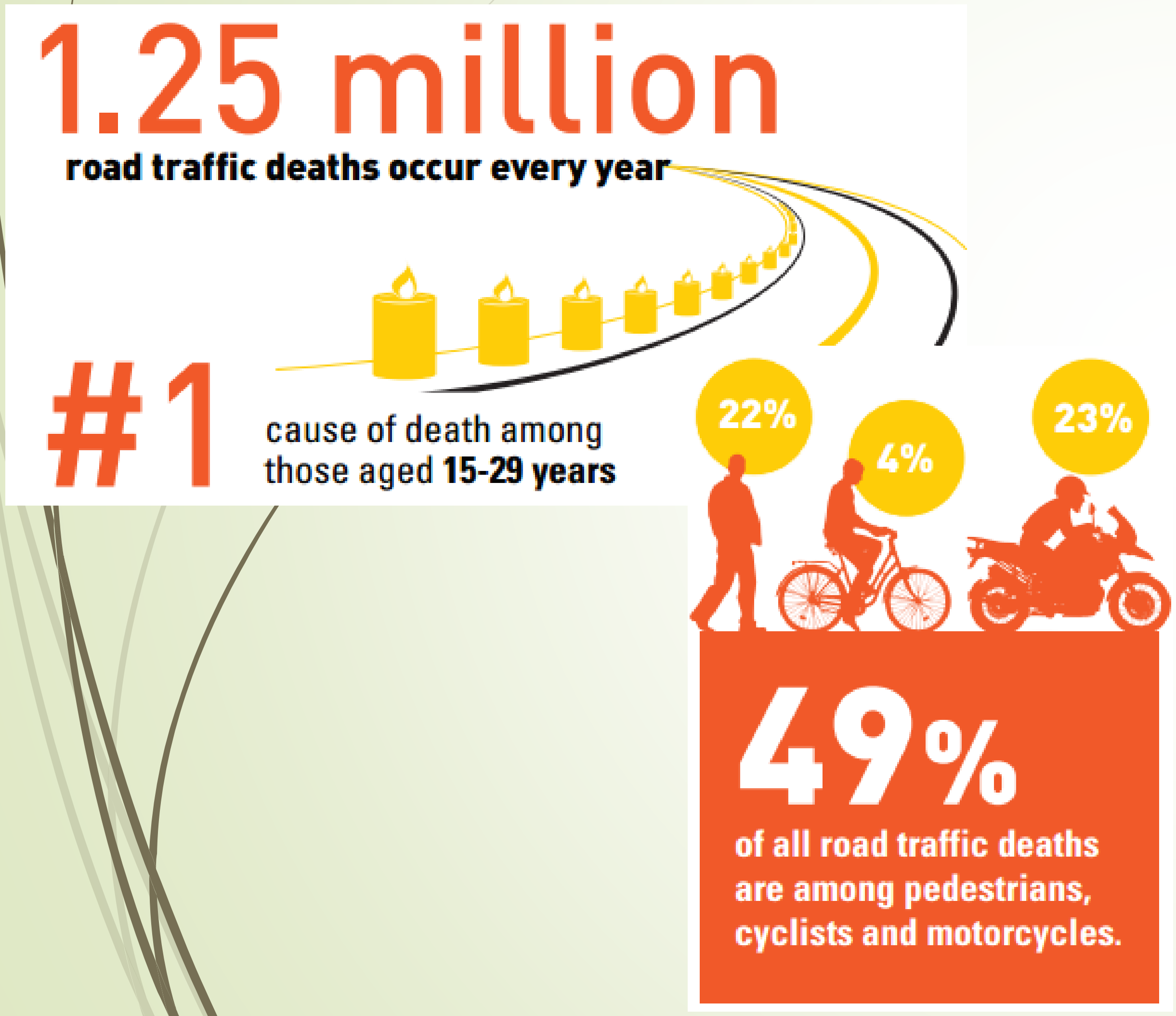
Evening Standard.

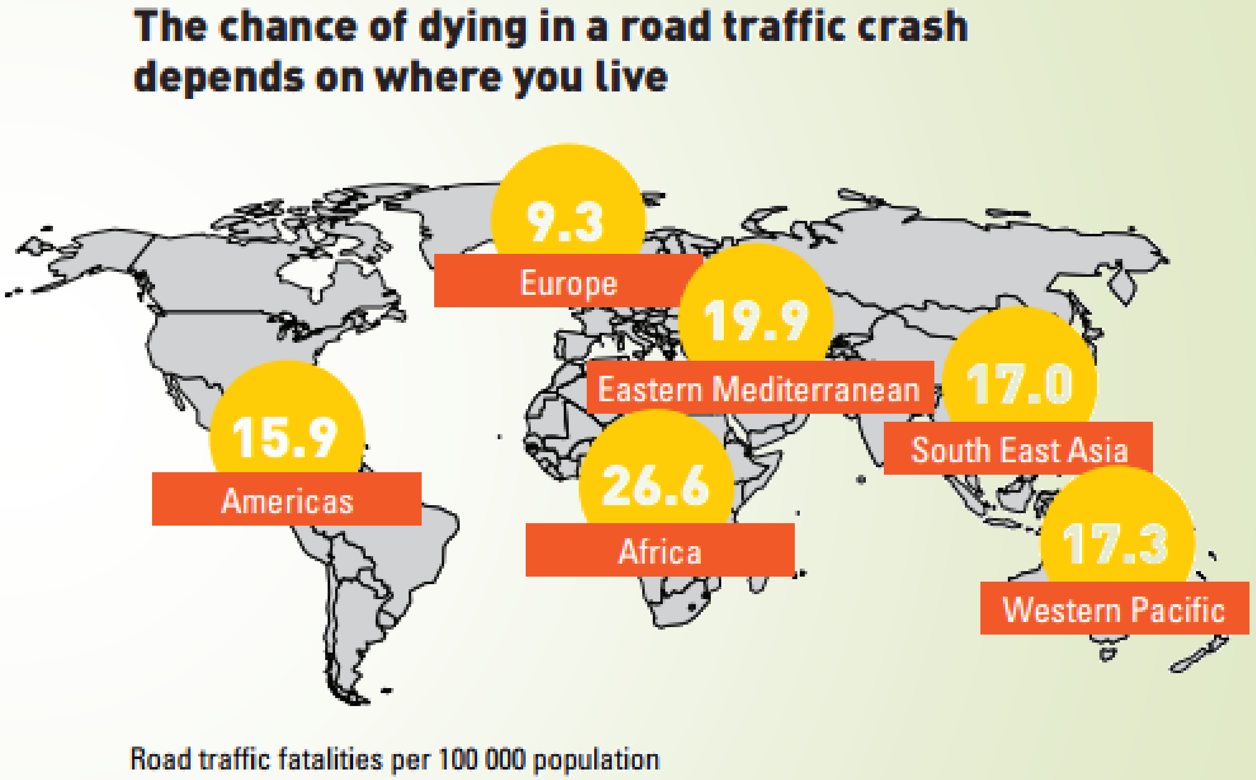
1 April 2019

Revealed: two million Londoners live in areas with illegal toxic air

"Pollution levels have been falling gradually for almost a decade due to the introduction of cleaner vehicle engines but experts are concerned that an increase in the number of **motorbikes and scooters** since 2010 is causing "hotspots" of roadside particulates."

Fatalities and Vulnerable Road Users





Malaria: 1 million per year

Objective

- To *detect* and *track* individual motorbikes even under occlusion. Use to increase safety and traffic enforcement
- Hypothesis: can use deep-learning object detection/classification
- Problem: virtually no large ground-truthed datasets of motorbike traffic

A public motorbike dataset (UMD)

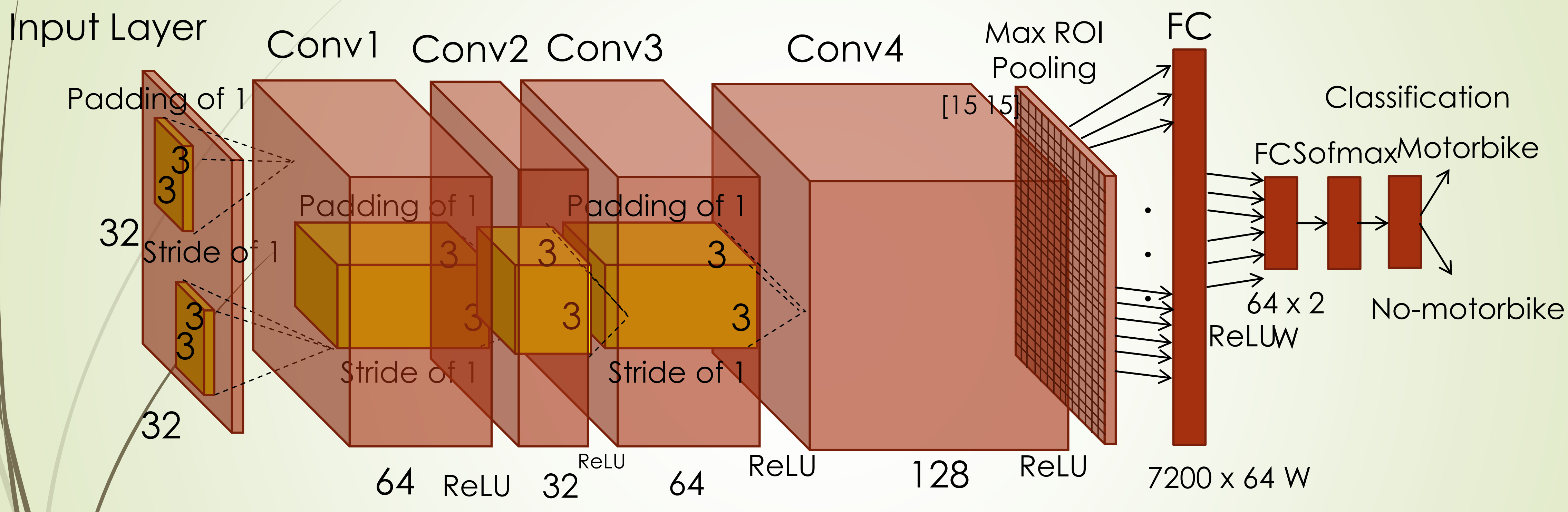
CORTEXICA
AI SOLUTIONS FOR BUSINESS

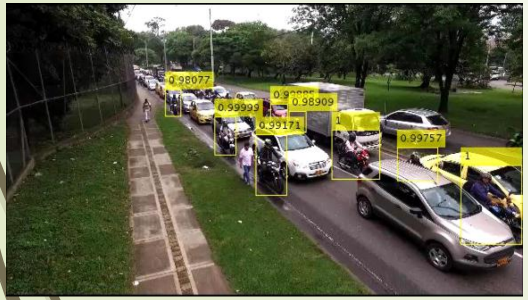
- 7,500/10,000 annotated images
- 220/317 motorcycles on urban traffic.
- 41,040/56,795 ROI annotated objects
- **60% Annotated object are occluded**

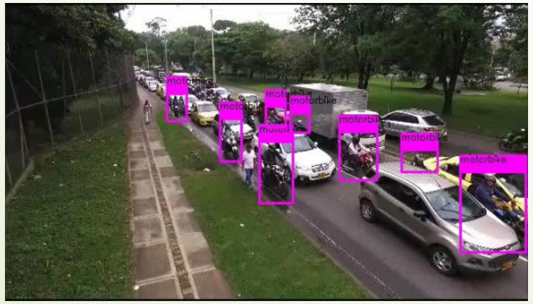
Available at: http://videodatasets.org

EspiNet4: Derived from Faster R-CNN

CORTEXICA
AI SOLUTIONS FOR BUSINESS



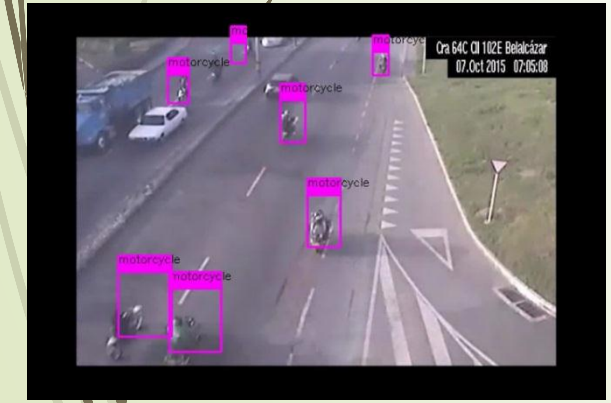




- Took 62 hours for training the dataset (90% Training – 5%Validating – 5%Testing)
- **AP=89,3% on UMD10K (2 layers=75%)**
- **YOLO AP=80%, Faster R-CNN=69%**

Under conventional CCTV conditions

EspiNet = 80%



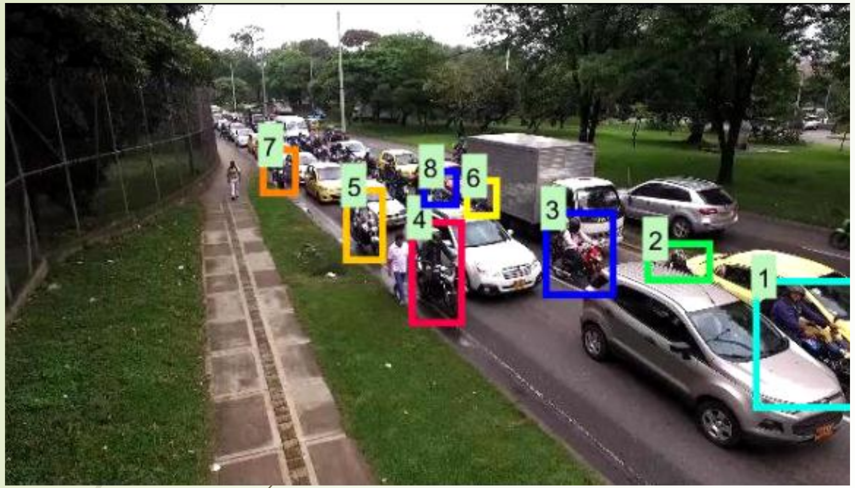
YOLO V3 AP = 77%

- 5000 annotated images (6 different cameras)
- 827 motorcycles tracks on urban traffic
- 704 x 480 (low resolution)
- **21,625 ROI** annotated motorbike objects
  Minimum H size 25 px
- **40 % Annotated object are occluded**

Available Soon at: http://videodatasets.org

CORTEXICA
AI SOLUTIONS FOR BUSINESS

Tracking by detection



Rcll	Prcn	FAR|	GT	MT	ML|	IDs	|	MOTA	MOTP
86.5	87.5	0.75|	128	126	2 |	128	|	**93.52**	**96.8**

Rcll	Prcn	FAR|	GT	MT	ML|	IDs	|	MOTA	MOTP
83,3	56,3	2,70|	816	411	81 |	503	|	16,3	67,2

Y. Xiang, A. Alahi, y S. Savarese, "Learning to track: Online multi-object tracking by decision making", en *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4705–4713.

Detection of People Boarding/Alighting a Metropolitan Train

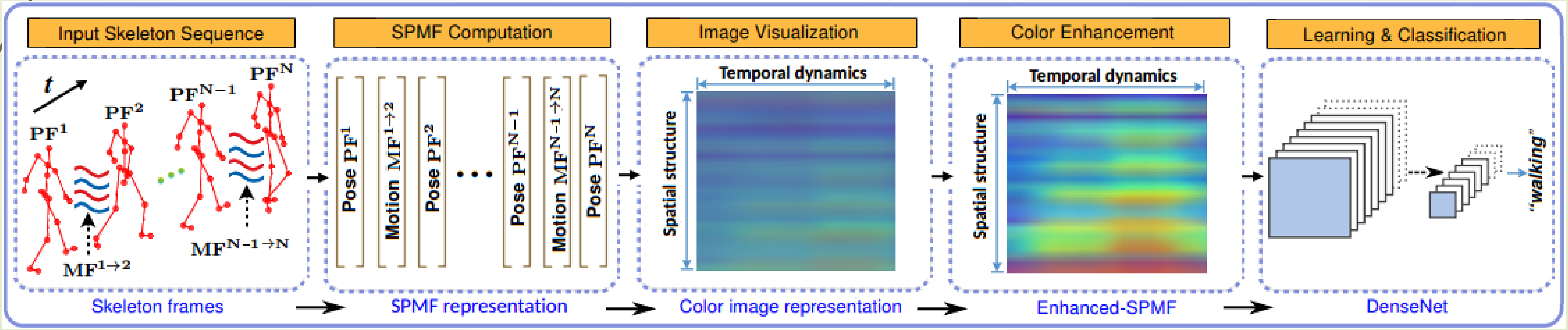
PAMELA-UANDES dataset (http://videodatasets.org)
EspiNet4 AP= 82%

Using RGB-D for human action recognition

CEREMA Metro Station Dataset (CEMEST)

Approach (using articulated data)

- NTU RGB+D dataset
- 3 Kinect-2 sensors
- Skeletons, RGB, depth
- 56K videos, 4M frames, 40 subjects, 60 classes



Experimented with ResNets and (latest) DenseNets (100, 190 and 250 deep)

Results

Method (protocol of [44])	**Year**	**Cross-Subject**	**Cross-View**
Lie Group Representation [28]	2014	50.10%	52.80%
Hierarchical RNN [42]	2016	59.07%	63.97%
Dynamic Skeletons [97]	2015	60.20%	65.20%
Two-Layer P-LSTM [44]	2016	62.93%	70.27%
ST-LSTM Trust Gates [45]	2016	69.20%	77.70%
Skeleton-based ResNet [2]	**2018**	**73.40%**	**80.40%**
Geometric Features [73]	2017	70.26%	82.39%
Two-Stream RNN [94]	2017	71.30%	79.50%
Enhanced Skeleton [98]	2017	75.97%	82.56%
Lie Group Skeleton+CNN [99]	2017	75.20%	83.10%
CNN Kernel Feature Map [96]	2018	75.35%	N/A
GCA-LSTM [95]	2018	76.10%	84.00%
SPMF Inception-ResNet-222 [1]	**2018**	**78.89%**	**86.15%**
Enhanced-SPMF DenseNet ($L = 100$, $k = 12$) (**ours**)	2018	**79.31%**	**86.64%**
Enhanced-SPMF DenseNet ($L = 250$, $k = 24$) (**ours**)	2018	**80.11%**	**86.82%**
Enhanced-SPMF DenseNet ($L = 190$, $k = 40$) (**ours**)	2018	**79.28%**	**86.68%**

Action Recognition (RGB-based)

- Recapping a bit:
  - Most of our daily life is about dealing with human activity
  - Driving
  - Working
  - Interacting with the city/people
  - Assisted living
  - Video search
  - Health & Safety
  - ….
- So, automating human action recognition can be a major technical and societal enabler

Simple

Complex

Some real-world challenges

- Camera movement
- Illumination changes
- View-point changes (including sudden changes as in cinema)
- Occlusion
- Diversity of subjects
- Visual similarity of different classes (difficult to train a classifier)
- When an action starts/end (temporal detection)?
- Where is the action (spatial localisation)?
- Many different subjects/actions at the same time
- Datasets (action “ImageNets”) e.g. Kinetics-600, activity.net, ….

Overview

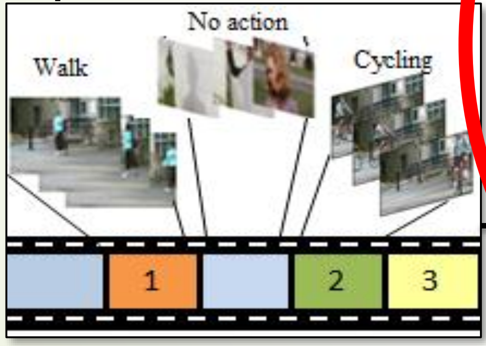
Remember this?

Input: Trimmed Videos

Output: What
Action Labels

Action Classification

Input: Untrimmed Videos



Temporal Proposals

Output: When + What
Start and end time + Action Labels

Temporal Action Detection

Input: Untrimmed Videos

Spatiotemporal Proposals

Output: Where + When + What
Bounding Box + Start and end time + Action Labels

Action Localization

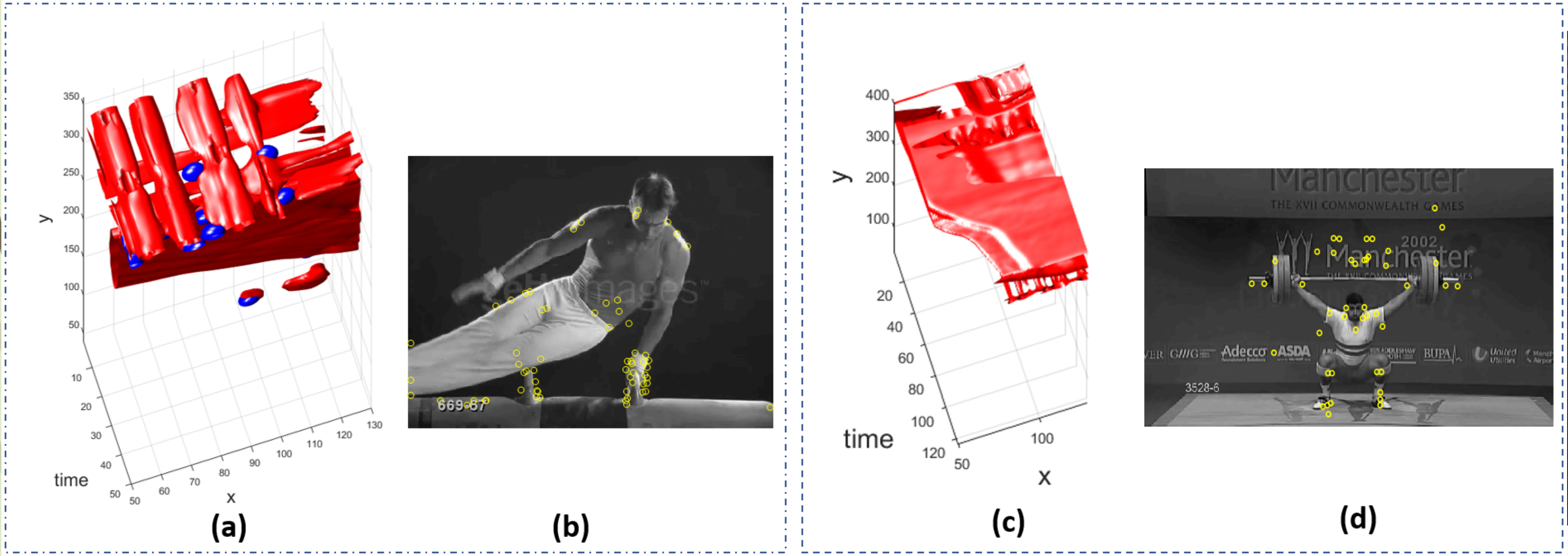
CORTEXICA
AI SOLUTIONS FOR BUSINESS

Some popular datasets

Dataset	No. of Actions	No. of Actors	No. of Videos	Camera Motion	Background clutter	Task	Evaluation Measure
KTH [35] (2004)	6	25	600	No	No	Recognition	Accuracy
Weizmann [36] (2005)	10	9	600	No	No	Recognition	Accuracy
CMU Crowded Videos [37] (2007)	5	6	53	No	Yes	Recognition	Accuracy
MSR Action I [37] (2009)	3	10	16	No	Yes	Spatiotemporal Detection	Recall, mAP
MSR Action II [38] (2010)	3	10+	54	No	Yes	Temporal Detection	Recall, mAP
MuHAVi-uncut [39] (2010)	17	7	8	N0	Yes	**Temporal Detection**	Recall, mAP
UCF11 (YouTube) [40] (2009)	11	R	1,600	Yes	Yes	Recognition	Accuracy
UCF50 [41] (2012)	50	R	6,681	Yes	Yes	Recognition	Accuracy
UCF101 [42] (2012)	101	R	12,320	Yes	Yes	Recognition	Accuracy
HMDB 51 [43] (2013)	51	R	6,766	Yes	Yes	Recognition	Accuracy
Thumos14 [44] (2014)	20	R	413	Yes	Yes	**Temporal Detection**	Recall, mAP
ActivityNet [45] (2015)	203	R	19,994	Yes	Yes	**Temporal Detection**	Recall, mAP

Features

3D Harris – STIP Detector

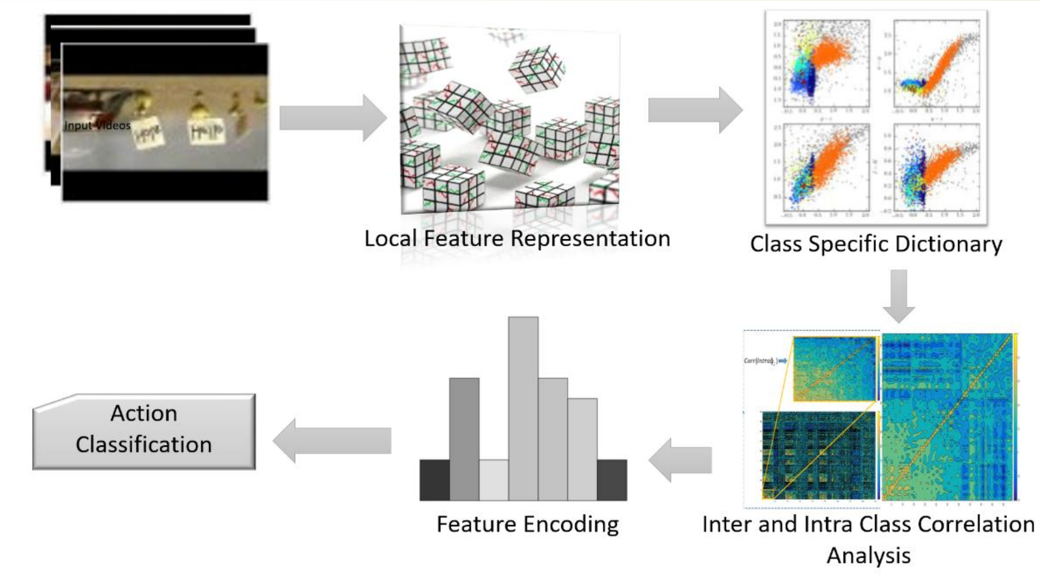


Inter and Intra Class Correlation Analysis (IIcCA)

CORTEXICA
AI SOLUTIONS FOR BUSINESS

- Optimise inter and intra class discrimination for a given training dataset
- Obtain highly *correlated* intra class visual words
- Obtian highly *uncorrelated* inter class visual words

Method	Accuracy
$II_C CA$	**98.9%**
CNN + Rank Pooling	87.2%
Dense Trajectories + MBH	88.0%
Spatio-temporal features using independent sub space analysis	86.5%

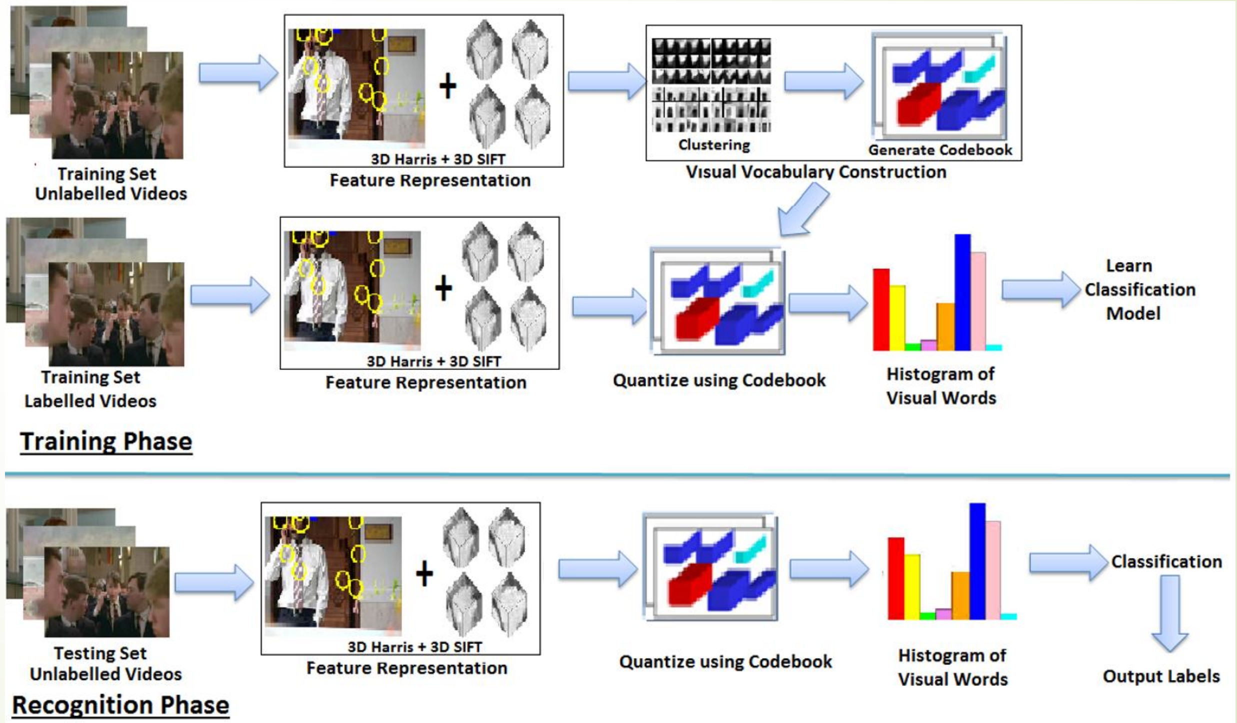


Nazir, S., *Yousaf. M.H., and Velastin. S.A., Inter and Intra Class Correlation Analysis (IIcCA) for Human Action Recognition in Realistic Scenarios. IET; International Conference of Pattern Recognition Systems, 2017.*

"Bag of Expressions"

Framework: Bag of Visual Words

- Spatio-Temporal Feature Representation
  - 3D Harris – Space Time Interest Point Detector
  - 3D SIFT – STIP Descriptor
  - C3D or R(2+1)D deep features
- Visual Vocabulary Construction
  - K-Mean Clustering
- Action Recognition
  - Histogram of Visual Word
  - Classification
    - Support Vector Machine
    - Naïve Bayes Classifier

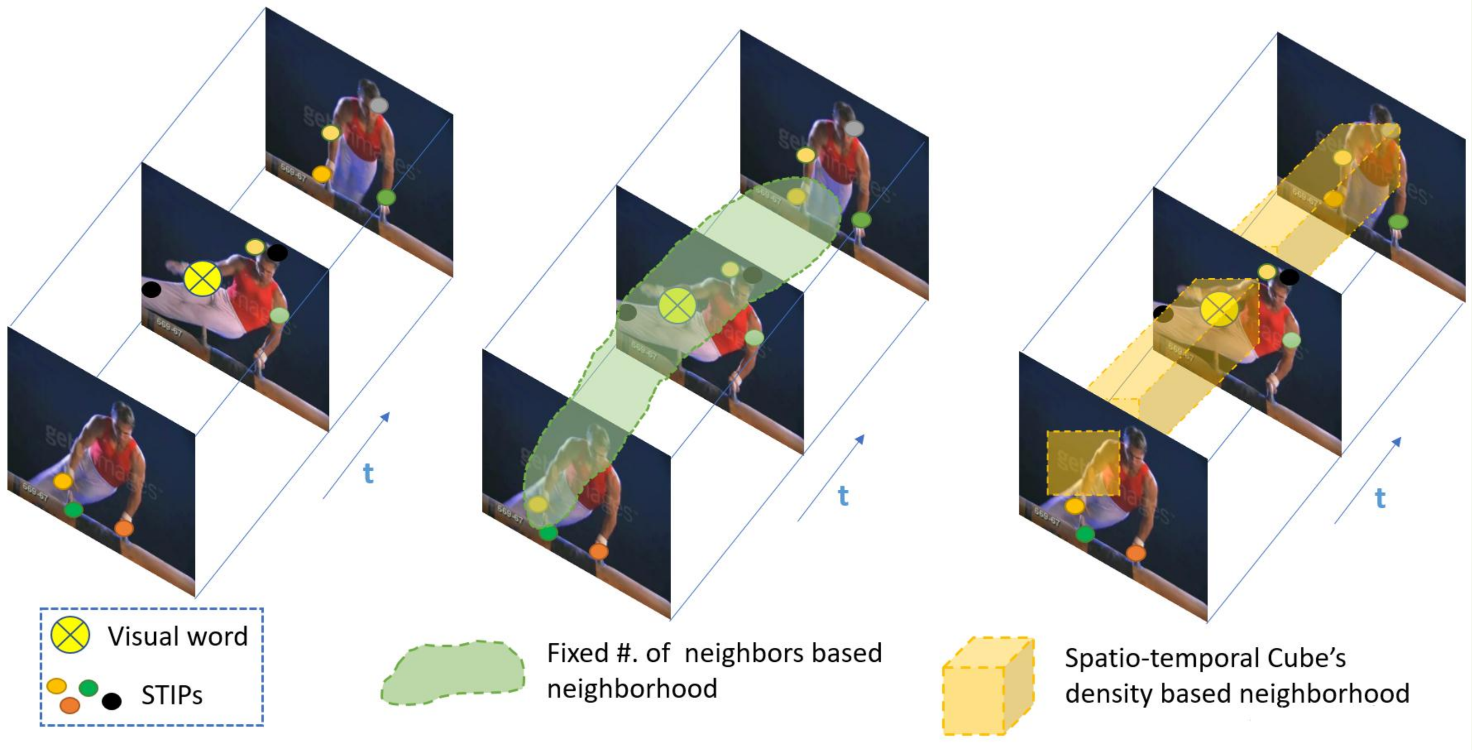


Nazir, S., Yousaf. M.H., and Velastin. S.A., Evaluating Bag of Visual Features (BoVF) Approach using Spatio Temporal Features for Action Recognition, Computers and Electrical Engineering, 2018.

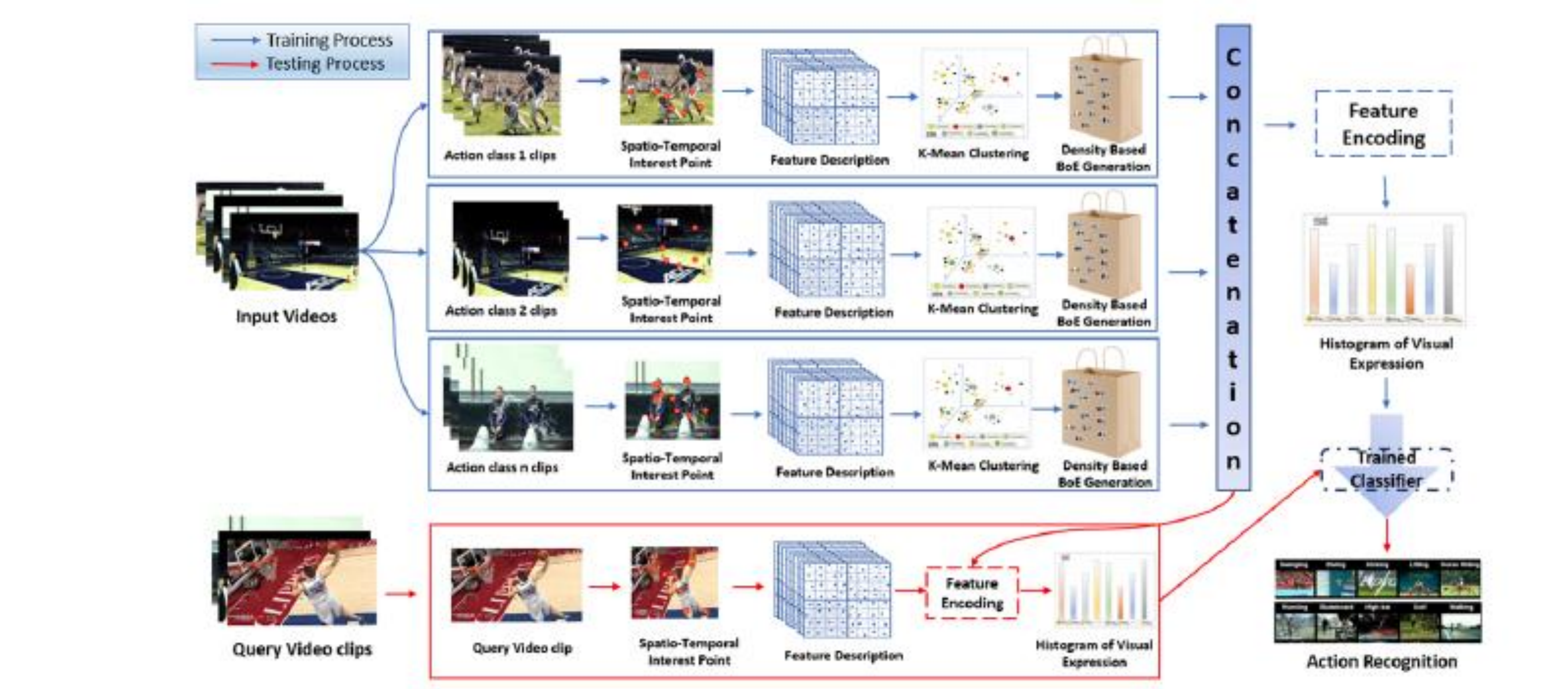
Results

HOLLYWOOD2		UCF Sports		KTH	
Ullah et al [13]	55.7%	Wang et al [2]	88.2%	Tsai et al [17]	100%
Wang et al [2]	58.3%	Yuan et al [20]	87.3%	Gilbert et al. [3]	94.5%
Jain et al [16]	66.4%	Zhu et al [23]	84.3%	Wang et al [2]	94.2%
Sun et al. [24]	48.1%	Sun et al. [24]	86.6%	Sun et al. [24]	93.1%
Ours	**68.1%**	**Our**	**94%**	**Our**	**91.82%**

Dynamic Neighbourhoods



Pipeline



Saima Nazir, Muhammad Haroon Yousaf, Jean-Christophe Nebel, Sergio A. Velastin. "Dynamic Spatio-Temporal Bag of Expression (D-STBoE) Model for Human Action Recognition", Sensors, https://www.mdpi.com/1424-8220/19/12/2790 DOI: https://doi.org/10.3390/s19122790 (2019)

Results

Author	Method	Results
Proposed	Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model	**94.10**
[71]	HMG + iDT Descriptor	93.00
[72]	Bag of Words and Fusion Methods	92.30
[5]	Dense Trajectories	91.70
[66]	Dense Trajectories and motion boundary descriptor	91.20

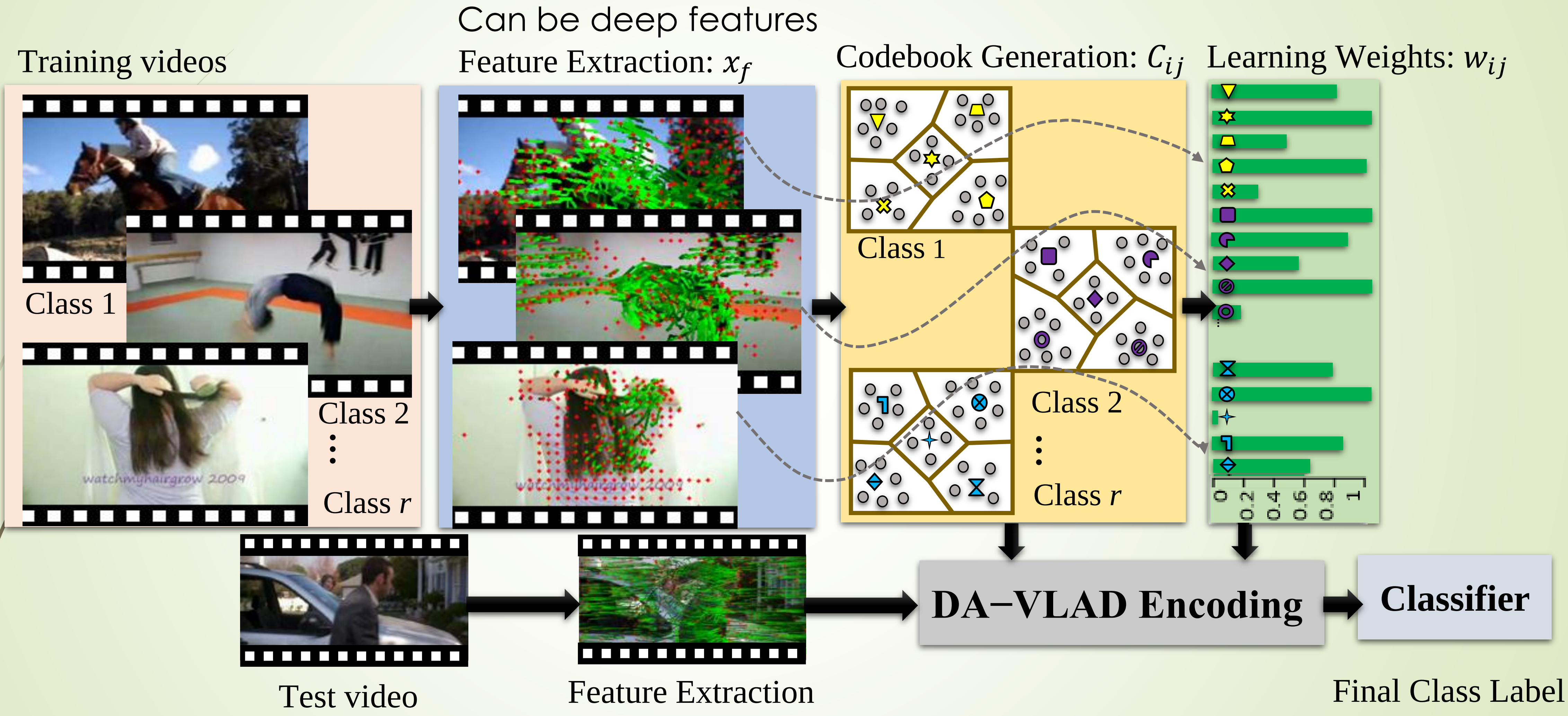
UCF-50

Author	Method	Results
Proposed	Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model	96.94
[43]	Spatio-temporal features with deep neural network	**98.76**
[59]	Universal multi-view dictionary	85.90
[55]	Foreground Trajectory extraction method	91.37
[70]	Graph-based multiple-instance learning	84.60
[65]	Local motion and group sparsity-based approach	86.10
[66]	Dense trajectories and motion boundary descriptors	84.10
[68]	Invariant spatio-temporal features with independent subspace analysis	75.80

UCF-11

Competitive with deep neural methods, and does not need large amounts of data

Combine features with BoW



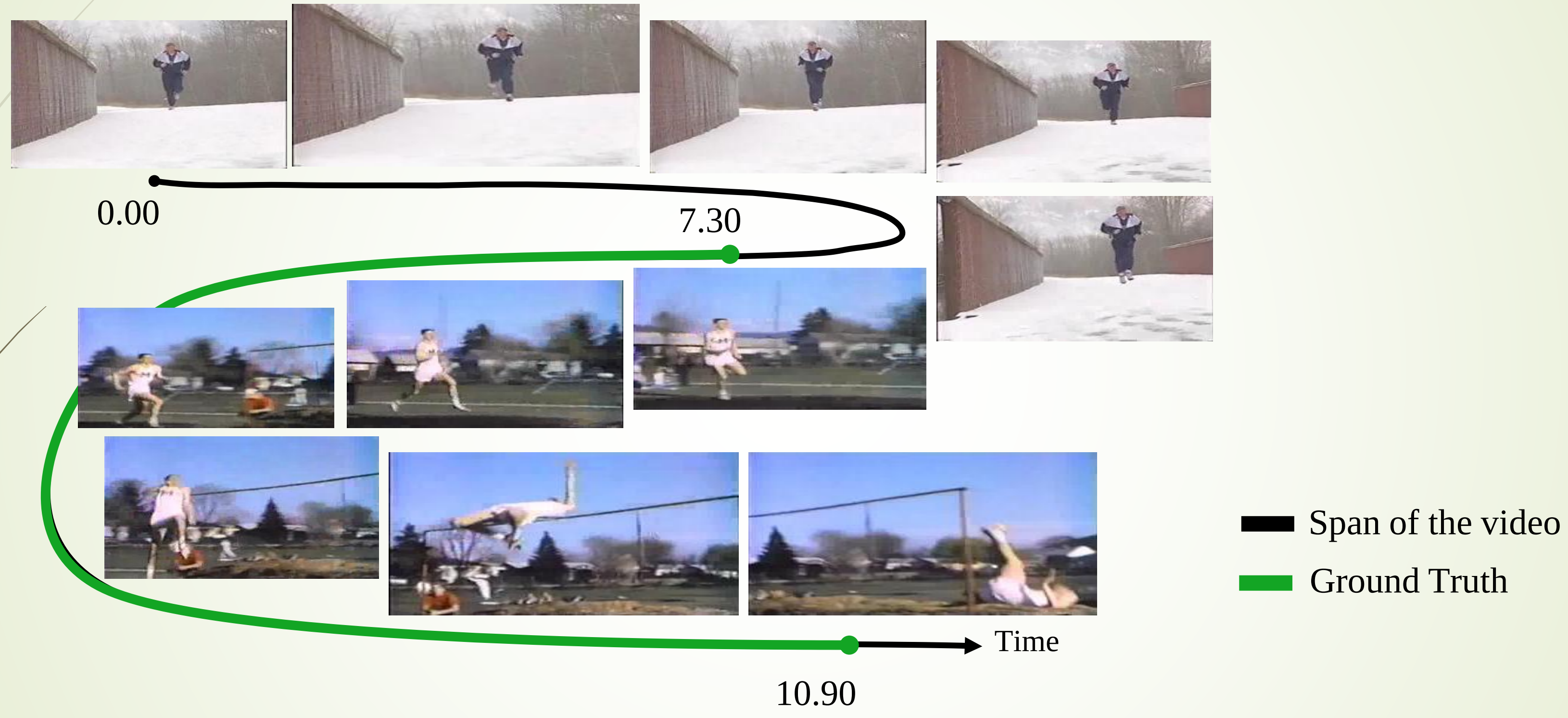
Fiza Murtaza, Muhammad Haroon Yousaf, Sergio A. Velastin. "DA-VLAD: DISCRIMINATIVE ACTION VECTOR OF LOCALLY AGGREGATED DESCRIPTORS FOR ACTION RECOGNITION", IEEE International Conference on Image Processing, ICIP-2018, October 7-10, Athens, Greece (2018)

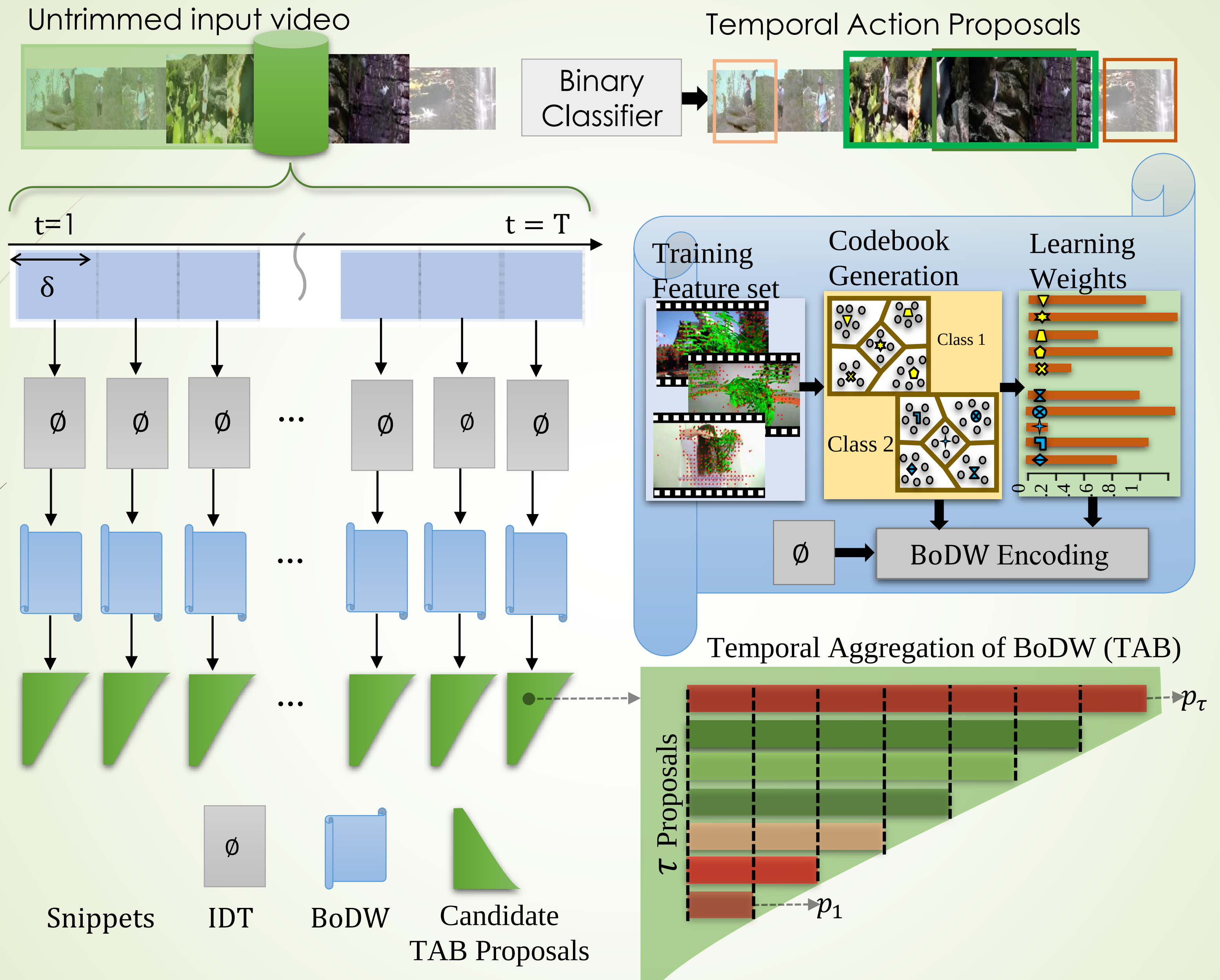
Comparison with other methods

Methods	UCF101	HMDB51
DT+MVSV [23]	83.5	55.9
iDT+iFV [24]	85.9	57.2
iDT+Hybrid [25]	87.9	61.1
iDT+MoFAP [26]	88.3	61.7
iDT+C3D [27]	90.4	-
iDT+C3D+ AdaScan [28]	93.2	66.9
iDT+GRP [29]	92.3	67.0
iDT+LTC [30]	92.7	67.2
iDT+ST-VLAD [31]	91.5	67.6
iDT+Two-Stream Fusion [33]	93.5	69.2
iDT+ActionVLAD(VGG-16) [33]	93.6	69.8
iDT+ST-VLMPF [34]	94.3	73.1
Our: iDT+DA-VLAD	**95.1**	**80.1**

~ recent "deep" with pre-training)

Temporal detection





Fiza Murtaza, Muhammad Haroon Yousaf and Sergio A Velastin. "TAB: Temporally Aggregated Bag-of-Discriminant-Words for Temporal Action Proposals", CVIU, 2019

CORTEXICA
AI SOLUTIONS FOR BUSINESS

mAP @ IoU threshold	0.3	0.4	0.5	0.6	0.7
Thumos14					
FG [12] (2016)	36.0	-	17.1	-	-
PSDF [13] (2016)	33.6	26.2	18.8	-	-
SCNN [6] (2016)	36.3	28.7	19.0	10.3	5.3
CDC [14] (2017)	40.1	29.4	23.3	13.1	7.9
TURN [15] (2017)	44.1	34.9	25.6	-	-
TPN [16] (2017)	44.1	37.1	28.2	20.6	12.7
TAG [17] (2017)	48.7	39.8	28.2	-	-
R-C3D [18] (2017)	44.9	35.6	28.9	-	-
SS-TAD [19] (2017)	45.7	-	29.2	-	9.6
SSN [20] (2017)	51.9	41.0	29.8	-	-
CBR [21] (2017)	50.1	41.3	31.0	19.1	9.9
ETP [22] (2018)	48.2	42.4	34.2	23.4	13.9
BoDS [Ours]	**54.9**	**47.2**	**41.5**	**37.5**	**31.6**
ActivityNet (Sports subset)					
[27]	-	-	33.2	-	-
FG [12] (2016)	-	-	36.7	-	-
TURN [15] (2017)	-	-	37.1	-	-
BoDS [Ours]	51.1	45.0	**38.1**	34.2	29.0

Where Next?

- **ViVIAN: Vulnerability via VIsual Analysis**
  - Vulnerable Road Users (pedestrians, 2-wheelers, mobilty impairment)
  - Active and Assisted Living
  - Inherently multi-disciplinary (computer vision, healthcare, transport engineering, …)
  - Intelligent roads: warn autonomous cars, optimise night lighting
- **FUE: Falls in Urban Environments**
  - USA, 30 deaths, 17,000 injuries in escalators/lifts in 2017
  - Over 65, one fall per year. Most common cause of injury in the over 75
  - 240,000 falls p.a. (2018) in hospitals, hip fractures 1.8M bed/days p.a., £1.9 billion. Fragility factures cost ~$4.4 billion, p.a. 25% social care

Story so far …

- Vision is a major sensing mode in us humans that allows us to interact with other humans and the environment, as such it makes heavy use of our brains!
- Machines would need to have similar capability to be able to interact well with us and the world
- Computer Vision in most cases is about converting visual data into *narratives* (language) that we are particularly good at processing
- The combination of big data, powerful GPUs and neural networks has given rise to impressive performance
- "Programs" become building learning models.
- But there is still much to be done particularly in "wild" environments

CORTEXICA
AI SOLUTIONS FOR BUSINESS

Thank you!

sergio.velastin@ieee.org