Comparison of motion-based approaches for multi-modal action and gesture recognition from RGB-D

Hugo Bertiche Argila
Supervisor: Sergio Escalera Guerrero
Co-supervisor: Maryam Asadi

Facultat d'Informàtica de Barcelona

31th May 2017

1 Introduction

2 Background

3 Methodologies

4 Results and Conclusion

Introduction

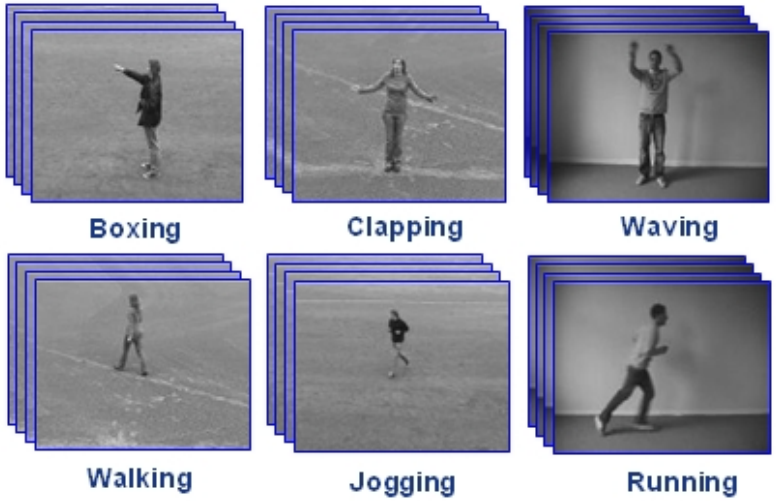
Introduction

- Problem Definition
- Application
- Temporal Dimension
- Multi-Modal Approaches
- Motivation and Goals

Problem definition

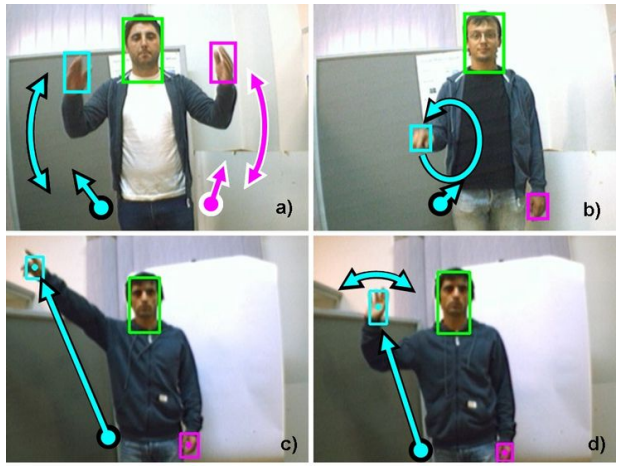
Action. Generic human activity.

- Walking
- Reading
- Jumping
- ...



Gesture. Movement of the hand, arm, body, head, or face that is expressive of an idea, opinion, emotion, etc.

- Waving
- Saluting
- Negating
- ...



Applications

- Human Computer Interaction.

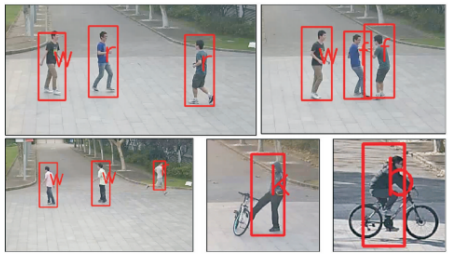
  Wilson et al. 2014.
- Ambient Assisted Living Systems.

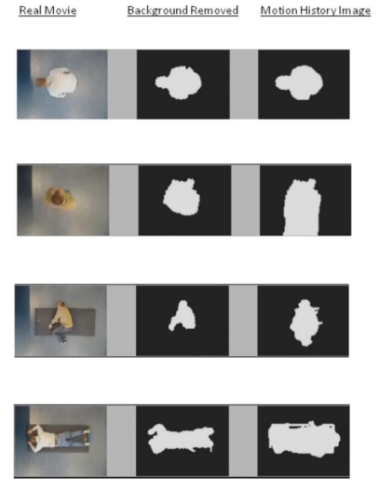
  Al-Shaqi et al. 2016.
- Healthcare monitoring systems.

  Osmani et al. 2008
- Security and surveillance.

  Danafar et al. 2007.
- Communications.

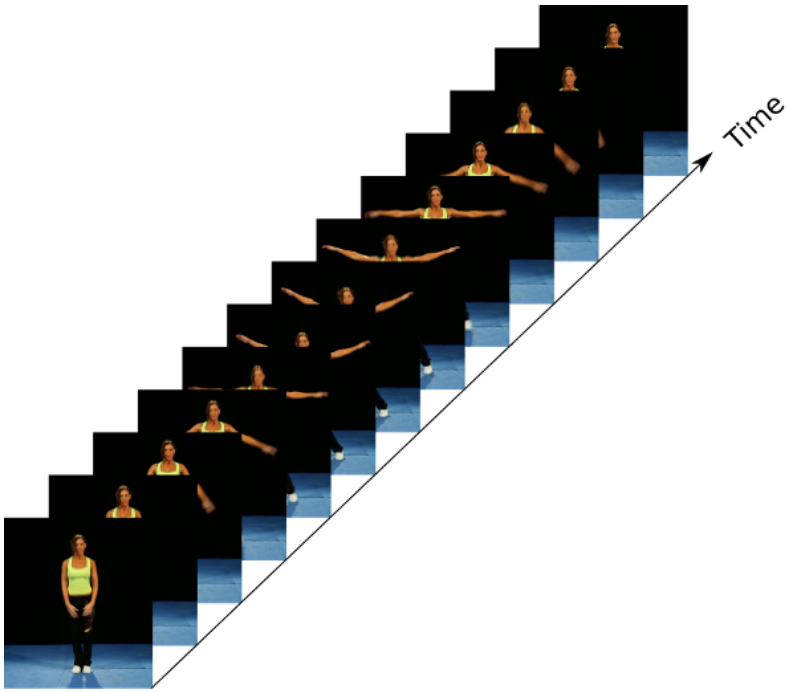
  Halim et al. 2015.
- Entertainment (Microsoft's Kinect).







Temporal dimension



Characteristics:

- Much more challenging task
- Larger volume of data to process
- Increase in model's complexity
- Different action speed

Approaches:

- Video summarization
- Local frame-level features aggregation
- Temporal sequence modeling
- ...

Multi-modal approaches

- Depth sensor availability
- Demonstrated benefits:

  Giles 2010.
  Shotton et al. 2013.
  Tang et al. 2012.
  Oreifej et al. 2013.
  ...
- Real world spatial units





Motivation and Goals

Motivation:

- Domain applications
- *Unexplored* field of research
- Multi-modal approaches

Goals:

- Exploration of the state-of-the-art techniques
- Application of a multi-modal approach (RGB-D)
- Assess future paths of research for the domain

1 Introduction

2 Background

3 Methodologies

4 Results and Conclusion

Background

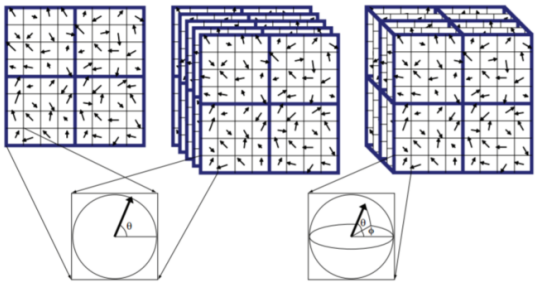
Background

- Hand-Crafted Features
- Motion-Based Features
- Deep Learning Models

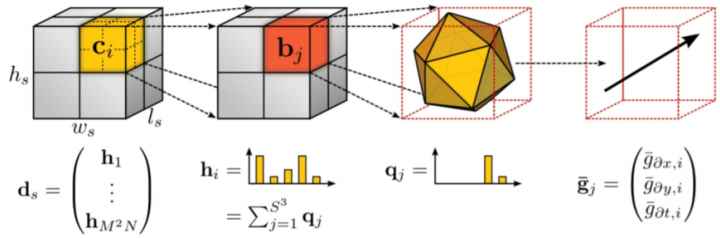
Hand-crafted features

Extended Spatio-Temporal Features

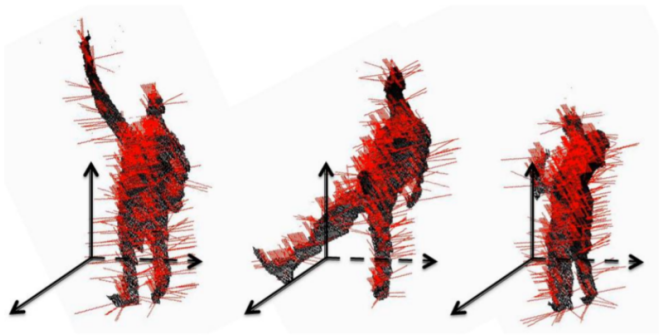
3D-SIFT



HOG-3D



HON4D



Motion-based Features

Optical Flow



For a given time t and pixel $(x, y)_t$:

$$(x, y)_{t+1} = (x, y)_t + d_t^{(x,y)} \quad (1)$$

Applications:

- Trajectory construction
- Descriptors: *HOF*, *MBH*
- CNN input

Motion-based Features

Scene Flow

For a given time t and point $(x, y, z)_t$:

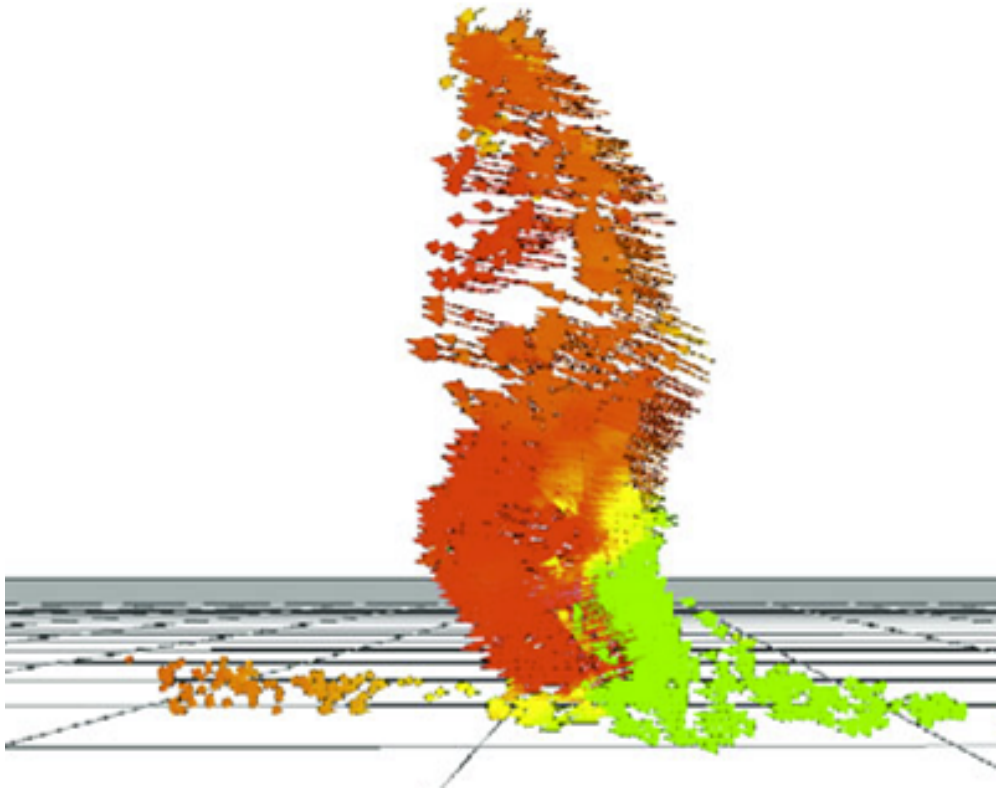
$$(x, y, z)_{t+1} = (x, y, z)_t + d_t^{(x,y,z)} \quad (2)$$

Applications:

- 3D trajectory construction
- Descriptors: $HOSF$, MBH_z
- CNN input

Advantages over optical flow:

- Real world motion units
- Z-axis motion



Deep Learning Models

Four non-mutually exclusive groups:

2D ConvNet:

- Recognition on one or more frames
- Scores merged

3D ConvNet

- 3D spatio-temporal kernels
- Larger number of parameters

Motion-based ConvNet

- Fed with optical/scene flow
- Explicitly models motion

Temporal Models: RNN and LSTM

- CNNs fed with video frames
- Modelling of temporal evolution

1 Introduction

2 Background

3 Methodologies

4 Results and Conclusion

Methodologies

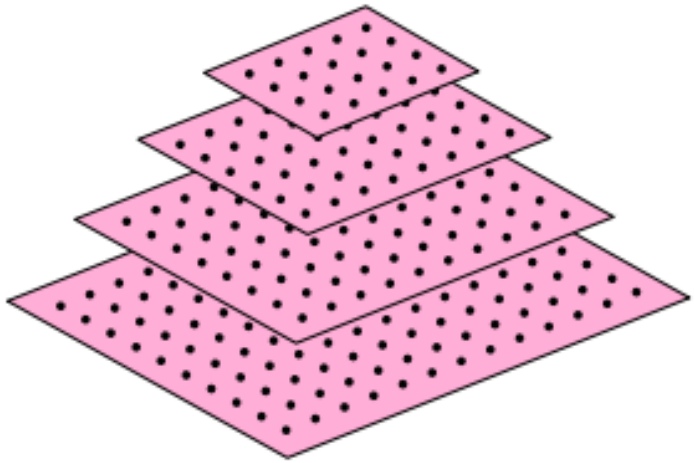
Methodologies

- Dense Trajectories
- Multi-Modal Dense Trajectories
- Two-Stream ConvNet

Dense Trajectories

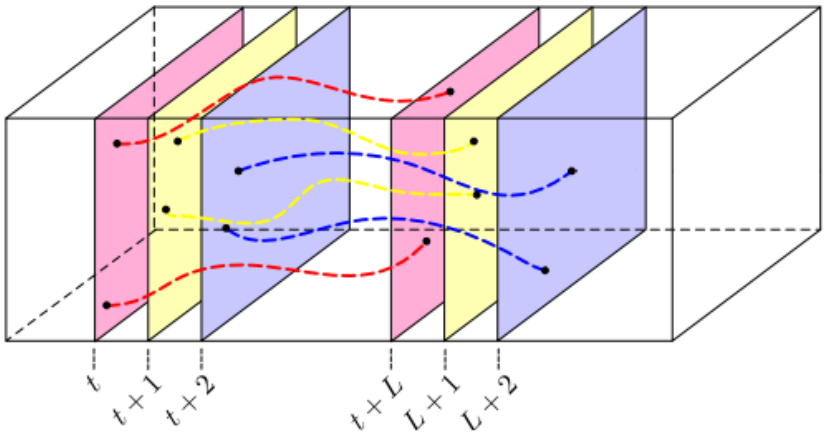
1^{st}: Multi-scale dense feature point sampling

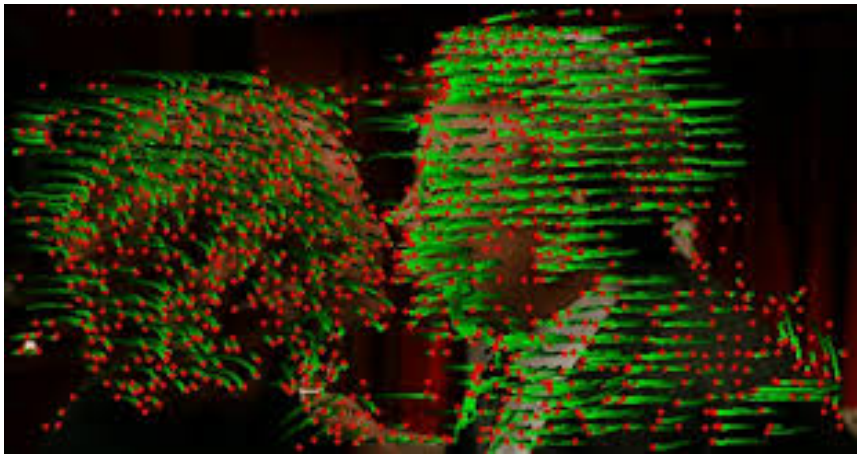




Dense Trajectories

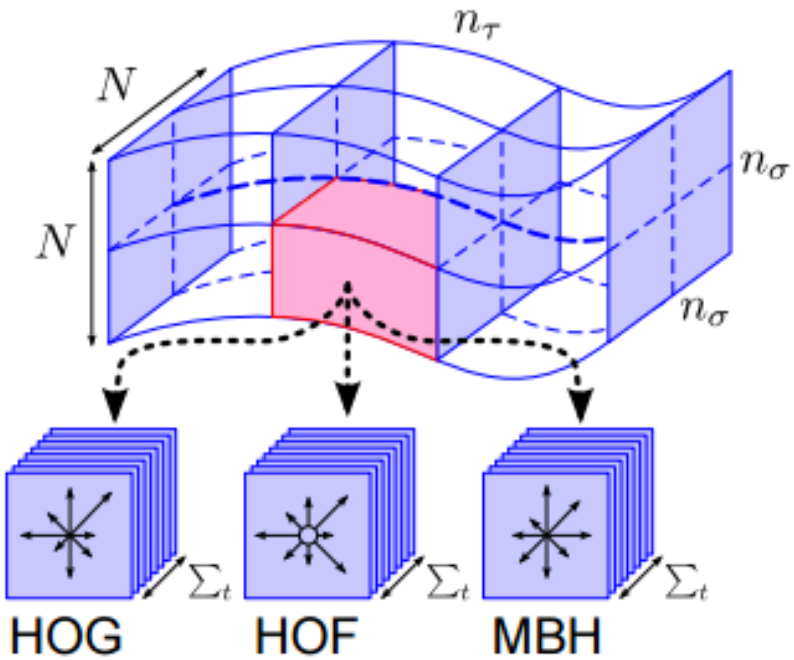
2^{nd}: Multi-scale trajectory construction





Dense Trajectories

3^{rd}: Trajectory description

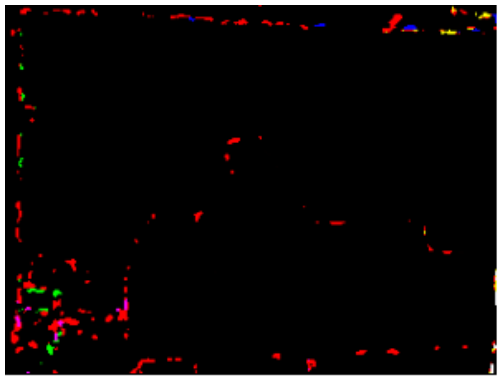


- $n_\sigma \times n\sigma \times n_T = 12$ spatio-temporal cells
- HOG: 96 dimensions
- HOF: 108 dimensions
- MBH: 192 dimensions
- Final descriptor: HOG + HOF + MBH

Multi-Modal Dense Trajectories

Trajectory construction

	Optical Flow	Scene Flow
Advantages	· Pixel tracking · Accurate	· 3D trajectories · Real world units
Disadvantages	· 2D trajectories · Camera distance dependant	· Noisy · Requires transformation for pixel tracking



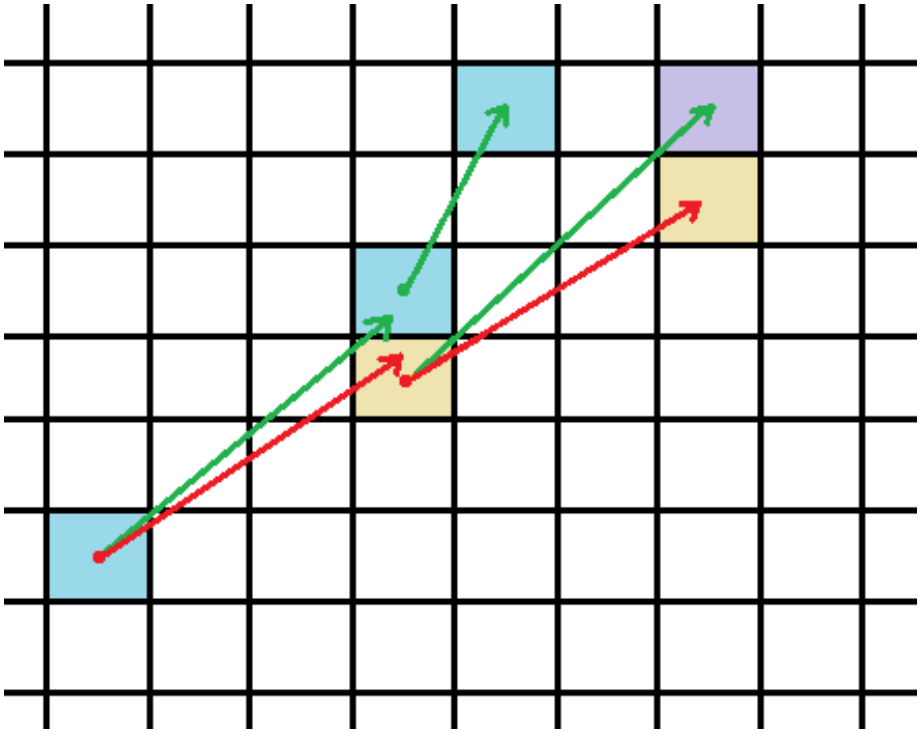
Scene flow



RGB frame

Multi-Modal Dense Trajectories

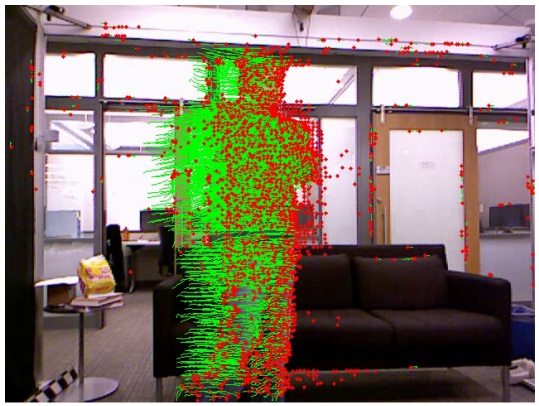
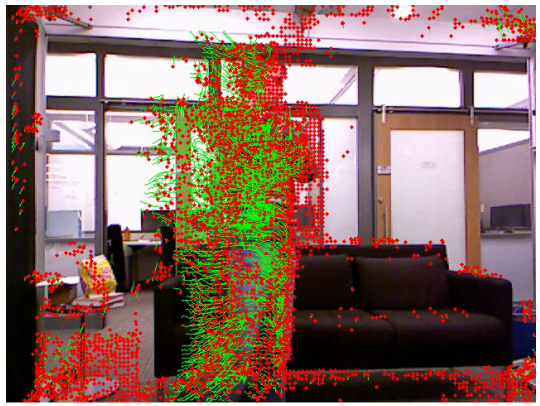
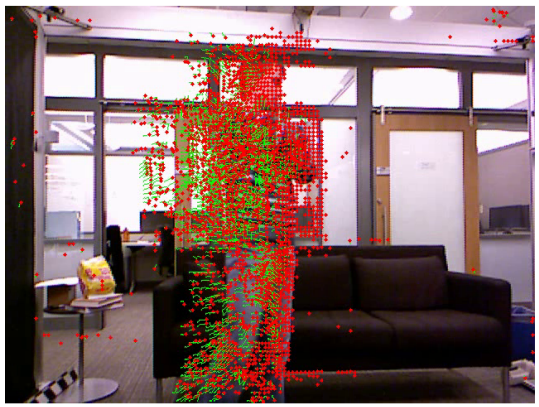
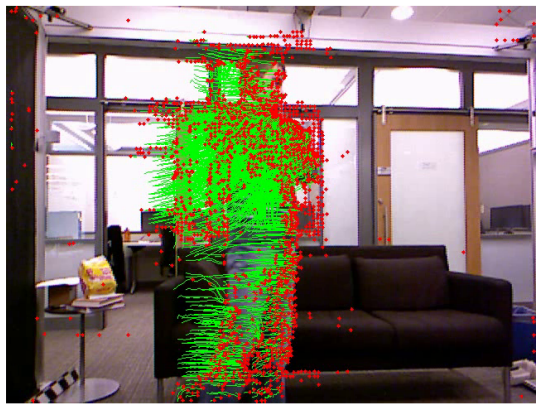
Pixel tracking with Scene Flow



- Requires camera parameters
- Sensitive to noise
- Error propagates

Multi-Modal Dense Trajectories

Trajectories configurations

Multi-Modal Dense Trajectories

Multi-modal Descriptors

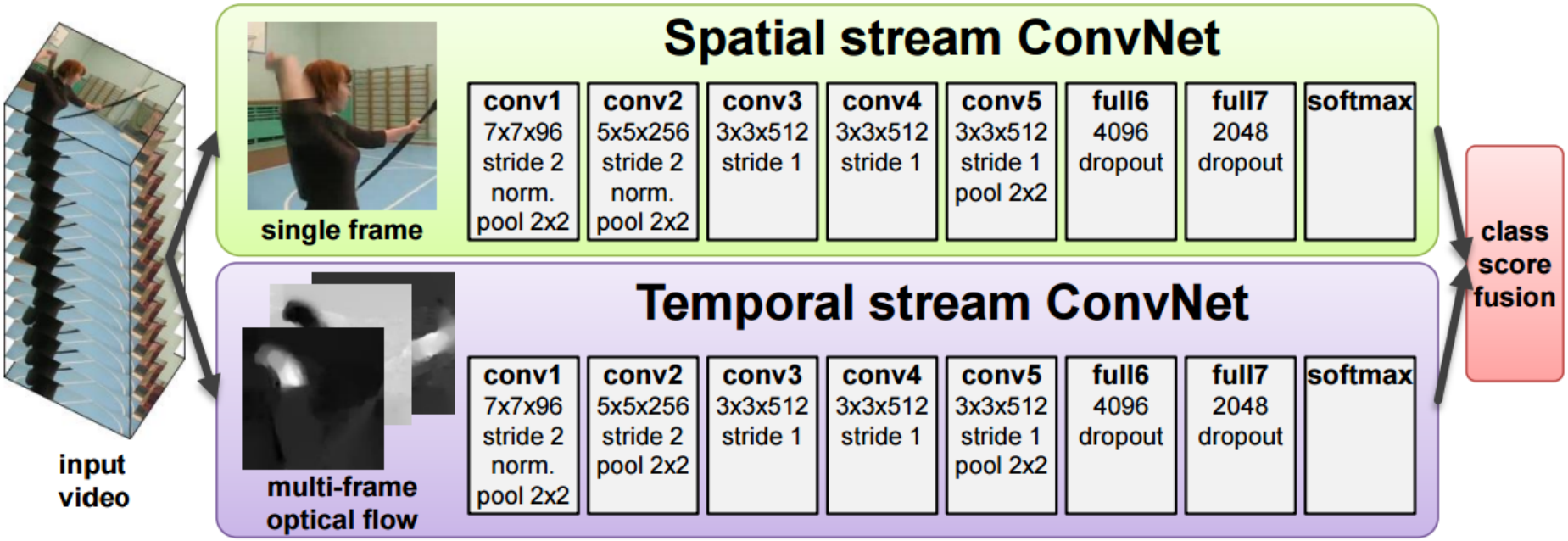
HON

- Depth based
- Appearance
- Object's 3D structure
- 300 dimensions

HOSF

- RGB-D based
- Motion
- 3D displacement
- 492 dimensions

Two-Stream ConvNet



1 Introduction

2 Background

3 Methodologies

4 Results and Conclusion

Results and Conclusion

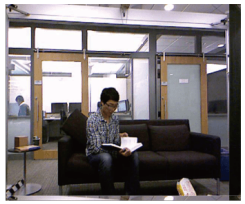
Results and Conclusion

- Datasets
- Experiments
- Results
- Conclusion and Future Work

Dataset: MSRDailyAct3D


Drink


Eat


Read book


write on a paper


Use laptop


play game


call cellphone


use vacuum cleaner

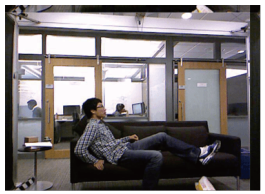

Cheer up


Sit Still

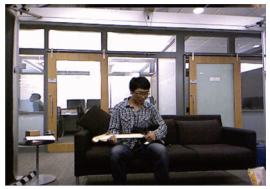

Walking


Sit down


Toss paper


lay down on sofa


Stand up


play guitar

Characteristics:

- Action recognition
- 16 classes
- 10 subjects
- 2 examples:

  Sitting down
  Standing up
- 320 samples

Evaluation:

- 50% Train
- 50% Test

Dataset: Montalbano II



Characteristics:

- Gesture detection and recognition
- 20 classes
- 940 samples
- $\approx$ 13.000 gestures

Evaluation:

- 1-470 Train
- 471-700 Validation
- 701-940 Test

Experiments

Dense Trajectories (Original/Multi-Modal)

- Run algorithm on each sample
- Treat each descriptor separately
- Fisher Vector encoding of each sample
- SVM classifier

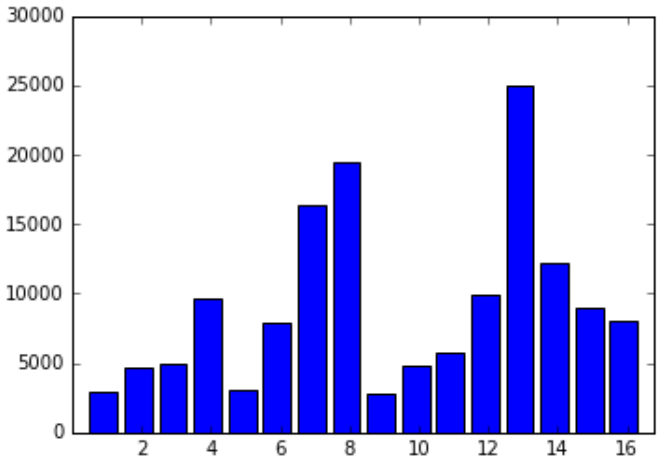
Two-Stream ConvNet

- Train each stream separately
- Pre-train on UCF101
- Fine-tuning
- Class scores fusion by SVM

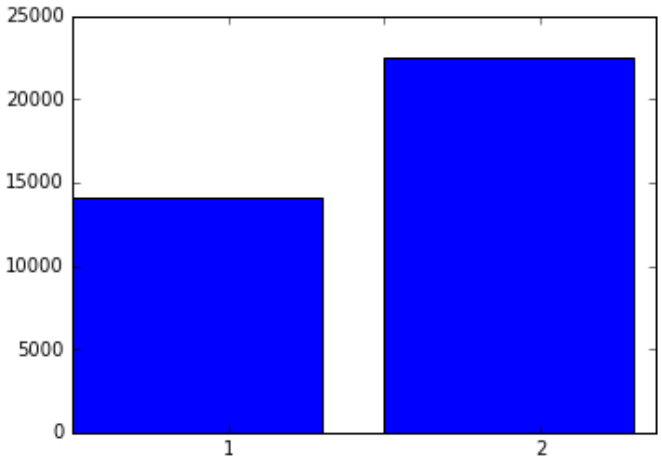
Dense Trajectories

MSRDailyAct3D

Trajectories per 'action'



Trajectories per 'example'



Dense Trajectories

MSRDailyAct3D

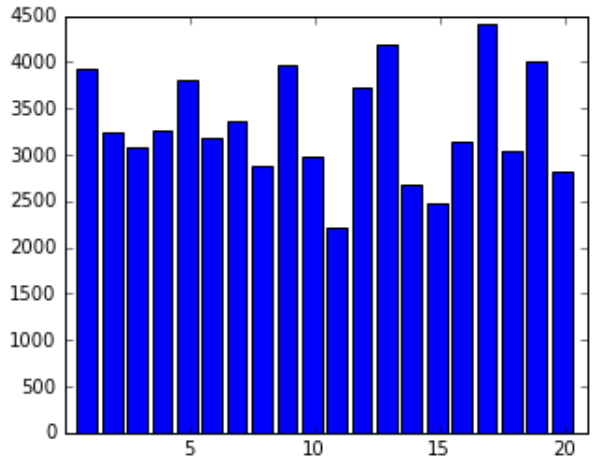
HOG	43.125%	HOG + HOF	58.125%
HOF	58.75%	HOG + MBH	58.75%
MBH	61.875%	HOF + MBH	62.5%
HOG + HOF + MBH		**63.125%**	

Figure 1: Accuracy on MSRDailyAct3D dataset

Dense Trajectories

Montalbano II

Trajectories per gesture



Dense Trajectories

Montalbano II

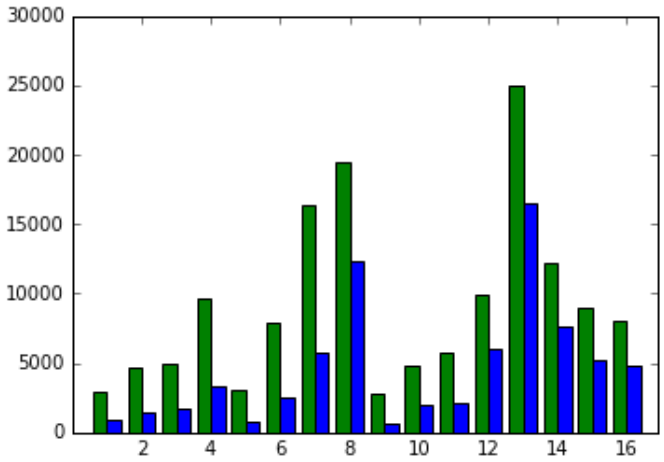
HOG	67.3%	HOG + HOF	79.4%
HOF	74.4%	HOG + MBH	81.6%
MBH	79.9%	HOF + MBH	82.0%
HOG + HOF + MBH		**83.5%**	

Figure 2: Accuracy on Montalbano II dataset

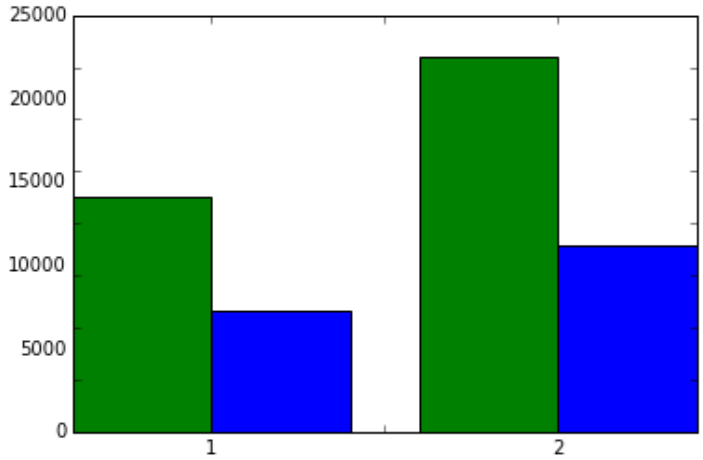
Multi-Modal Dense Trajectories

MSRDailyAct3D

Trajectories per action



Trajectories per 'example'



Multi-Modal Dense Trajectories

MSRDailyAct3D

HOG	45.625%	HOG + HOF	62.5%
HOF	60.625%	HOG + MBH	66.25%
MBH	69.375%	**HOF + MBH**	**70%**
HOG + HOF + MBH		65.625%	

Figure 3: Accuracy on MSRDailyAct3D dataset

HON	72.5%	HON + HOF + MBH	75%
HOSF	58.125%	HOG + HOSF + MBH	66.875%
HON + HOSF	69.375%	HON + HOSF + MBH	75%
HON + MBH	**78.125%**	HOG + HON + HOF + MBH	74.375%
HOSF + MBH	68.125%	HOG + HOF + HOSF + MBH	66.875%
HOG + HON + HOF + HOSF + MBH		73.75%	

Figure 4: Accuracy on MSRDailyAct3D dataset

Multi-Modal Dense Trajectories

Montalbano II

HON	77.67%	HON + HOG	79.57%
HON + HOF	82.81%	**HON + HOF + MBH**	**85.66%**
HON + MBH	85.10%	HON + HOG + HOF + MBH	85.66%

Figure 5: Accuracy on Montalbano II dataset

Two-Stream ConvNet

MSRDailyAct3D

RGB	48.75%
Optical Flow	47.50%
RGB + Optical Flow	**62.50%**

Figure 6: Accuracy on Montalbano II dataset

Montalbano II

RGB	94.34%
Optical Flow	53.82%
RGB + Optical Flow	**97.08%**

Figure 7: Accuracy on Montalbano II dataset

Results summary and Comparison

MSRDailyAct3D

Methods	Accuracy
Dynamic Temporal Warping*	54.0%
Actionlet Ensemble*	85.8%
HON4D*	85.0%
3D Trajectories	72.0%
Long-term Motion Dynamics	**86.9%**

Dense Trajectories	63.125%
Multi-modal Dense Trajectories	**78.125%**
2D CNN RGB	48.75%
2D CNN Optical Flow	47.50%
2D CNN RGB + Optical Flow	62.50%
MMDT + Two-Stream CNN	65%

Montalbano II

Methods	Jaccard Index / Accuracy
MRF, KNN, PCA, HoG*	0.827
AdaBoost, HoG*	0.834
Multi-Scale DNN	0.836
Multi-Scale DNN*	0.870
Temp Conv + LSTM	**0.906**

Dense Trajectories	83.5%
Multi-modal Dense Trajectories	85.66%
2D CNN RGB	94.34%
2D CNN Optical Flow	53.82%
2D CNN RGB + Optical Flow	97.08%
MMDT + Two-Stream CNN	**97.37%**

Conclusion and Future Work

Conclusions:

- Hand-Crafted vs. Deep Learning
- Action vs Gesture
- Multi-Modal benefits

Future work:

- Datasets: NTU & IsoGD
- Scene Flow
- Four-Stream ConvNet
- Hand-Crafted + Deep: fusion strategies

Thanks for your attention