Machine Learning Models for Action Recognition

Ceyhun Burak Akgül
Vistek ISRA Vision & Boğaziçi University / Istanbul
www.vistek-isravision.com
www.busim.ee.boun.edu.tr
www.cba-research.com
Georges Braque 1913
The “WHAT”

Action Recognition

*Given one or more images with one or more persons performing an action, we want to design a system recognizing the performed action.*

The “HOW” – *at least part of it*

Machine Learning Model

- Learning by example
- Any statistical approach, which involves training with un/labeled data
What’s in an Action?

The Lindy Hop is an American dance that evolved in Harlem, New York City in the 1920s and 1930s and originally evolved with the jazz music of that time.

The Lindy Hop combines elements of both partnered and solo dancing by using the movements and improvisation of black dances along with the formal eight-count structure of European partner dances.

This is most clearly illustrated in the Lindy's basic step, the swingout.
What’s in an Action?
What’s in an Action?

Atomic movement
described at the limb level

Multiple primitives
in sequence or in combination,
possibly cyclic whole-body movement

Multiple actions
in sequence or in combination,
possibly involving multiple
persons and/or objects

Series of multiple activity
instances involving different
people and objects in a specific context

Basic moves:
Basic step, rock step,
triple step, hold hand,
turn, …

Solo Swingout:
Rock step -> step -> triple step &
turn -> step -> step -> triple step

Swing out with a partner

The Lindy Hop Party!
The Blueprint Action Recognizer...

**INPUT**
- RGB video
- Depth sequence

**REPRESENTATION** // LOW-LEVEL
- Silhouette
- Gradients
- Optical Flow
- Local Space-Time Features

**DETECTION & TRACKING**
- Face, shape, body(-part) models
- Kinematic models

**DESCRIPTION** // MID-LEVEL
- Motion History Images
- Space-Time Motion Templates
- Space-Time Objects
- BoW-style Action Descriptors

**LEARNING AND INFERENCE** // HIGH-LEVEL
- Template Matching
- Manifold Learning
- Discriminative Classifiers
- State-Space Models

- PCA
- LLE
- SVM
- AdaBoost
- RDF
- NN-Scheme
Outline(*)

Challenges
Surveys
Datasets

A Parade of ML Models
  The Nearest Neighbor Scheme
  Manifold Learning
  Discriminative Classifiers
  State-Space Models

Variations on the Theme
  Mining Action Data
  Use of Context

Concluding Remarks

(*) The full set of slides can be downloaded from

Woman with a Guitar
Georges Braque 1913
why is it difficult?

CHALLENGES

- Class Definitions and Variability
- Environment and Recording Settings
- Spatio-Temporal Variability
- Real-Time Recognition
- On-the-Fly Recognition
- Training Data Collection and Labeling
- Evaluation and Benchmarking
Challenges – 1/4

Class Definitions and Variability

**Basic**
- Walking
- Jogging
- Running
- Boxing
- Hand waving
- Hand clapping
- ...

**Daily Living**
- Getting out of bed
- Watching TV
- Reading a book
- Using computer
- Eating meal
- Drinking
- ...

**Outdoor**
- Walking alone
- Meeting w/ others
- Window shopping
- Fighting
- Leaving luggage behind
- ...

**Action**

*increasingly more complex and variable...*
- anthropometric differences
- multiple objects and people
- contextual differences

**Activity**
Challenges – 2/4

Environment and Recording Settings

<table>
<thead>
<tr>
<th>Issues</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Static vs. Dynamic backgrounds</td>
<td>• Person detection and tracking</td>
</tr>
<tr>
<td>• Occlusions</td>
<td>• Action detection and segmentation</td>
</tr>
<tr>
<td>• Lighting conditions</td>
<td>• Level of detail for understanding</td>
</tr>
<tr>
<td>• Recording rate and resolution</td>
<td>• Choice of method</td>
</tr>
<tr>
<td>• Recording modality</td>
<td></td>
</tr>
</tbody>
</table>
Challenges – 3/4

Spatio-Temporal Variability

<table>
<thead>
<tr>
<th>Issues</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Pose differences</td>
<td>• View invariance required</td>
</tr>
<tr>
<td>• Moving camera</td>
<td>• Person detection and tracking</td>
</tr>
<tr>
<td>• Execution time and rate</td>
<td>• Action detection and segmentation</td>
</tr>
<tr>
<td></td>
<td>• Temporal effects: remove or take into account?</td>
</tr>
</tbody>
</table>
Challenges – 4/4

Other Challenges

• Real-Time Recognition
• On-the-Fly Recognition
• Training Data Collection and Labeling
  – Reliable and objective annotations required for learning
  – Large and varied training and test data for all classes required for generalization
• Evaluation and Benchmarking
  – Common realistic benchmarks required to compare methods
who has done what?

SURVEYS
Taxonomies, taxonomies ...

[Moeslund et al., 2006]
• 352 papers covered for the period 2000-2006
• Functional taxonomy: Initialization, tracking, pose estimation, tracking

[Turaga et al., 2008]
• 144 papers covered

Turaga et al.’s methodological taxonomy
**Taxonomies, taxonomies ...**

[Poppe, 2010]

- 180 papers covered
- Representation and classification aspects treated separately

[Weinland et al., 2011]

- 153 papers covered
- Focused on representational aspects (spatial vs. temporal) as well as action segmentation and view invariance
- Classification and Learning aspects not discussed
**Taxonomies, taxonomies ...**

[Aggarwal and Ryoo, 2011]

- 102 papers covered

*Aggarwal and Ryoo’s hierarchical approach-based taxonomy*
where to train and test?

**DATASETS**

- The Usual Suspects
- Surveillance Datasets
- The Wild Ones
- Datasets for ADL
- Rising Stars: RGBD Datasets
# The Usual Suspects

**KTH**
- 6 actions
- 25 subjects
- Simple background

**Weizmann**
- 10 actions
- Class variations
- Varied background

**INRIA IXMAS**
- 11 actions
- 12 subjects
- Controlled env.
- Gaming scenario
Surveillance Datasets

PETS
- **Performance Evaluation of Tracking and Surveillance Challenge** (since 2000)
- Focused on crowd surveillance characteristics/events within a real-world environment
- Person count and density estimation – People Tracking – Flow Analysis and Event Recognition

CAVIAR
- **CAVIAR** project video clips collected at public spaces (entrance lobby and shopping mall) using a wide angle lens
- Activities: people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place.

SDHA
- **Semantic Description of Human Activities**: Three Challenges in ICPR 2010
- **Interaction Challenge**: High-level interactions between two humans, e.g., hand-shake and push
- **Aerial View Challenge**: Simple one-person actions taken from a low-resolution far-away camera
- **Wide Area Challenge**: Monitor human activities with multiple cameras observing a wide area

ViSOR
- **Video Surveillance Online Repository**
- Diverse environments and settings: outdoor, indoor,
- Object-level and action/activity-level meta-data available
The Wild Ones – 1/4

Hollywood2

12 classes of human actions and 10 classes of scenes
3669 video clips from 69 movies
Approximately 20.1 hours of video
Comprehensive benchmark in realistic and challenging settings
The Wild Ones – 2/4

UCF101

101 action categories:
(extension of UCF50)
(1) Human-Object Interaction
(2) Body-Motion Only
(3) Human-Human Interaction
(4) Playing Musical Instruments
(5) Sports.

13320 videos from YouTube

Large diversity:
actions classes, large variations in camera motion, object appearance and pose,
object scale, viewpoint, cluttered background, illumination conditions, etc.

No actors
51 action categories:
(1) General facial actions
(2) Facial actions with object manipulation
(3) General body movements
(4) Body move’ts with object interaction
(5) Body move’ts for human interaction

6849 clips from the Prelinger archive, YouTube and Google videos
(minimum 101 clips per category)
The Wild Ones – 4/4

ActionBank

A combination of KTH, UCF Sports, UCF50, HMDB51
Datasets for ADL [Activities of Daily Living] – 1/2

ADLs differ from typical actions in that they can involve long-scale temporal structure (making tea can take a few minutes) and complex object interactions (a fridge looks different when its door is open).

UCI-ADL
1 million frames of dozens of people performing ADL

Annotated with activities, object tracks, hand positions, and interaction events.

TUM-Kitchen
Observations of several subjects setting a table in different ways.

Video data
Motion capture data
RFID tag readings
Magnetic sensor data
Detailed action labels

YouCook
88 YouTube cooking videos (various recipes) from third-person viewpoint

Frame-by-frame object and action labels
ADLs differ from typical actions in that they can involve long-scale temporal structure (making tea can take a few minutes) and complex object interactions (a fridge looks different when its door is open).

**UESTC Senior Home Monitoring Dataset**

12 types of human actions: drinking, eating meals, eating snacks, getting out of bed, going to bed, sleeping, smoking, walking, playing mahjong, washing face, washing feet and watching TV.

Performed by 6 seniors in their own rooms
4 month long data collection
10 days per recording for each senior
Approximately 1.8TB data
(25fps, 360x288 pixels, Xvid MPEG-4 Codec)
Berkeley MHAD

11 actions by 7 male and 5 female subjects (23-30 years except one elderly)
5 repetitions per subject per action
660 action sequences, 82 minutes total recording time

(1) Movements in both upper and lower extremities
(2) Actions with high dynamics in upper extremities
(3) Actions with high dynamics in lower extremities

Simultaneously captured by five different systems: optical motion capture system, four multi-view stereo vision camera arrays, two Microsoft Kinect cameras, six wireless accelerometers and four microphones.
Rising Stars: RGBD Datasets – 2/5

Microsoft Research (MSR) Datasets

**MSRGesture3D**
Depth sequences captured by Kinect
12 dynamic American Sign Language (ASL) gestures
10 people, 2-3 times per subject per gesture class, 336 depth sequences

**MSRDailyActivity3D**
Depth, RGB, and skeletal data sequences captured by Kinect (RGB and depth not synchronized)
16 activities: drink, eat, read book, call cellphone, write on a paper, use laptop, ...
10 subjects, 2 times per subject per activity (one in standing, the other in sitting position)

**MSRAction3D**
Depth and skeletal joint data sequences captured by Kinect-like device
20 general action classes
10 subjects, 2-3 times per subject per activity 567 depth sequences

**MSRC-12**
Depth sequences and skeletal data captured by Kinect
12 gesture classes from a 1st person shooter video game
30 people, 6244 gesture instances in 594 sequences (6hrs 40min)
Rising Stars: RGBD Datasets – 3/5

Cornell Activity Datasets

**CAD-60**
- 60 RGB-D videos and tracked skeletons
- 4 subjects: 2 male, 2 female (one left-handed)
- 5 different environments: office, kitchen, bedroom, bathroom, and living room
- 12 activities: rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, ...

**CAD-120**
- 120 RGB-D videos of long daily activities
- 4 subjects: 2 male, 2 female (one left-handed)
- 10 high-level activities: making cereal, taking medicine, un/stacking objects, microwaving ...
- 10 sub-activity (action) labels: reaching, moving, pouring, eating, drinking, ...
- 12 object affordance labels: reachable, movable, pourable, containable, ...
Rising Stars: RGBD Datasets – 4/5

LIRIS Human Activities Dataset

- RGB, grayscale and depth sequences
- RGB-D videos of various ADL: discussing, phone calls, giving an item, ...
- Fully annotated with spatial and temporal positions in video
- Originally shot for the ICPR-HARL 2012 competition
## WorkoutSU-10

### SL Balance with Hip Flexion
- Flex your hip of your non-weight bearing leg up to 90 degrees, bend your knee, and hold.
- Use your core & lower extremity muscles to control your center of mass to maintain your balance.

### SL Balance-Trunk Rotation
- Raise your arms to chest height and clasp your hands together.
- Slowly rotate your trunk to one side a comfortable distance, return to the starting position, and then rotate your trunk in the other direction.
- Use your core & lower extremity muscles to control your center of mass to maintain your balance.

### Lateral Stepping
- Slightly bend your knees and begin stepping to the side keeping your toes facing straight ahead.
- Use your core & lower extremity muscles to control your center of mass to maintain your balance.
- Perform this for a specific number of steps then return back in the other direction.

### Thoracic Rotation – Bar on shoulder
- Assume standing position with bar across shoulders.
- Rotate your trunk to one side.
- Hold 30 (s) at end range, then slowly release stretch.

### Hip Adductor Stretch
- Shift your weight over one leg by bending your knee and straighten the opposing leg to be stretched.
- You should feel a stretch on the inside aspect of your thigh and groin of the straight leg.
- Hold 30 (s) at end range, then slowly release the stretch.

---

### Depth sequences and skeletal data captured by Kinect

### Balance Exercises

### Stretching and Flexibility Exercises

### Strengthening Exercises

10 therapeutic action classes in 3 broad categories

15 participants

10 repetitions per subject per class, 1200 instances in total

Recorded in the context of the ViPSafe Project on elderly monitoring (Sabancı University and Vistek ISRA Vision)
MACHINE LEARNING MODELS

The Nearest Neighbor Scheme
Manifold Learning
Discriminative Classifiers
State-Space Models
The Bayes Classifier

\[ C^* = \text{argmax } P(C|D) \]

- **C**: action class
- **D**: description of the observed visual data
- **P(C|D)**: posterior probability of class **C** having observed description **D**

All machine learning models try to approximate this formula in one way or the other
The Machine Learning Pipeline...

Data

Unseen
Available

Model(s) to be specified

Training & Validation

Test data

Specified model(s)

Testing

Update Discard Combine/Fuse

Does it generalize?

Knowledge

Models

Generative? Discriminative?
The Nearest Neighbor Scheme
The Nearest Neighbor Scheme

Put a “ball” around the test instance
The Nearest Neighbor Scheme

Put a “ball” around the test instance

P(X|D) = 0.75
P(Y|D) = 0.25
P(Z|D) = 0.00

Assign D to class X
The Nearest Neighbor Scheme

Put a “ball” around the test instance

P(X|D) = 0.75  
P(Y|D) = 0.25  
P(Z|D) = 0.00

Assign D to class X

In which space should we put the ball? 

- Action description
- Matching measure
The Nearest Neighbor Scheme

Action Prototype Database

Class 1
View 1
View 2
... View K

Class 2
View 1
View 2
... View K

Class N
View 1
View 2
... View K

Test action descriptor

Matching

Class estimate

Distance Measure
Minkowski
Deformable
Dynamic Time Warping
...

Space-time shapelets [Batra et al. 2008]
Space-time shapes [Gorelick et al. 2007]
Motion History Images [Bobick and Davis 2001]
Action MACH [Rodriguez et al. 2008]
Shape Motion Prototype Trees [Lin et al. 2009]
The Nearest Neighbor Scheme

Dynamic Time Warping (DTW)

• DTW computes a nonlinear (space-)time normalization between a template and a test vector
• Vectors could be of different length
• Better capture intrapersonal variations in gait than linear warping
• Computation based on dynamic programming
• Can be speeded-up by using certain (spatio-) temporal consistency constraints.

Non-parametric matching of shape sequences [Veeraraghavan et al. 2005]
Function space of an activity [Veeraraghavan et al. 2006]
Deformable action templates [Yao and Zhu 2009]
Manifold Learning

In which space should we put the “ball”?

- Action description
- Matching measure
Manifold Learning

**In which space should we put the “ball”?**

- **Action description**
  - Can be very high-dimensional
  - Might be noisy
  - May lie on an intrinsically much lower dimensional space

- **Matching measure**
  - Can be adapted to the intrinsic structure of data
  - Can be learnt in a un/supervised manner
Manifold Learning

In which space should we put the “ball”?
Manifold Learning

In which space should we put the “ball”?
Manifold Learning

In which space should we put the “ball”? 
Manifold Learning

In which space should we put the “ball”? 

A

B
Manifold Learning

Apply the good old PCA

[Rosales 1998]
[Masoud and Papanikolopoulos 2003]

... or unravel a non-linear function between input and output spaces in an unsupervised way!

[Blackburn and Ribeiro 2007]
[Wang and Suter 2007]
[Wang and Suter 2008]

... or using some labeled data learn a metric between action instances discriminatively!

[Jia et al. 2008]
[Poppe and Poel 2008]
[Tran et al. 2008]
Given a pattern description, discriminative classifiers focus on separating two or more classes, rather than modeling the class-conditionals.

They constitute proxies to estimate the posterior probability.

Many off-the-shelf implementations available: SVM, AdaBoost and variants, Random Forests

**SVM**
- Directly minimize a regularized upper bound on empirical classification error: Exact solution (QP)
- Generalizes well provided enough data
- Good with fixed vectorial description

**Boosting**
- Combine several weak classifiers into a strong one
- Ability to choose features
- Generalizes well provided enough data
- Blueprint algorithm: works with any weak learner/feature

**Random Forests**
- Randomized extension of combined trees
- Ability to choose features
- Can seamlessly employ different types of features
- “A la mode”

[Smith et al. 2005]
[Jhuang et al. 2007]
[Laptev et al. 2007]
[Nowozin et al/ 2007]
[Fathi et al. 2008]
Discriminative Classifiers

STIPs + BoW-based Action Recognition Framework

Detection of feature / interest points

Space-time patches

Patch representation as feature vector \( v = (v_1, v_2, ..., v_n) \)

Description of space-time patches

NN, SVM, Adaboost or RDF classifier

Feature Detectors
Harris3D, Hessian, Cuboid, ...
Dense

Feature Descriptors
HOG/HOF, HOG3D, Cuboid, Ext. SURF, ...

[Schuldt et al. 2004]
[Laptev 2005]
[Dollár et al. 2005]
[Oikonomopoulos et al. 2009]
An action class and its observations can be described as a sequential probabilistic graphical model.

**Generative**

\[ P(C,D) \text{ then } P(C|D) \propto P(D|C)P(C) \]

Hidden Markov Models (HMM) generate states and observations.

**Observations**
Sequence of visual descriptions of an action instance

**States**
Sequence of phases that an action instance undergoes

**Discriminative**

\[ P(C|D) \text{ directly} \]

Conditional Random Fields (CRF) focus on the posterior without generating the states.
State-Space Models

Hidden Markov Models

\[ X_i \]: hidden states
\[ y_i \]: observations
\[ a_{ij} \]: state transition probabilities
\[ b_{ij} \]: output probabilities

[Feng and Perona 2002]
[Ikizler and Forsyth 2008]
[Lv and Nevatia 2006]
[Ramanan and Forsyth 2003]
[Yamato et al. 1992]
State-Space Models

Conditional Random Fields (CRF)

- Advantages over HMM -

• CRFs specify the probabilities of possible label sequences given an observation sequence:
  ➔ No modeling effort on the observations

• The conditional probability of the label sequence can depend on arbitrary features of the observation sequence without requiring to account for the extra distributions:
  ➔ Can incorporate more information without extra effort
  ➔ Independence assumptions not as strict as in HMMs

[Ning et al. 2008]
[Shi et al. 2008]
[Sminchisescu et al. 2006]
[Wang and Suter 2007]

[Zhang and Gong 2010]
[Natarajan and Nevatia 2008]
[Mendoza and Blanca 2008]
there is more to the story...

VARIATIONS ON THE THEME

Mining Action Data Using Context
Mining Action Data

- Feature Extraction
- Skeleton feature
- LOP feature
- Feature at all the joints
- Fourier Temporal Pyramid
- Actionlets
- Multiple Kernel Learning
- Action Labels

266 feature time-series
13300 unique features

Discriminatively select features by RDF

SVM learning

Action classifiers on selected features

[Wang et al. 2012]  
[Negin et al. 2013]
Using Context – 1/4

Slide credit: Hedvig Kjellström
Using Context – 2/4

- **Object Context**
  - The objects involved in the activity
  - Object state changes

- **Scene Context**
  - Scene category
  - Scene topology, metrics

- **Semantic Context**
  - Grammars, temporally close actions
  - Speech, captions, storyline
  - Expert and domain knowledge

- **Photogrammetric Context**
  - Image statistics, sensor info

Slide credit: Hedvig Kjellström
Using Context – 3/4

Object context
[Yao and Fei-Fei 2010]

Slide credit: Hedvig Kjellström
Using Context – 4/4

Semantic context
[Gupta et al. 2009]

Slide credit: Hedvig Kjellström
what else?

CONCLUDING REMARKS
Challenges are still there...

**CAN’T DO MUCH FOR THESE!**

- Class Definitions and Variability
- Environment and Recording Settings
- Spatio-Temporal Variability

**CAN AND SHOULD DO A LOT MORE HERE!**

- Real-Time Recognition
- On-the-Fly Recognition
- Training Data Collection and Labeling
- Evaluation and Benchmarking
But the biggest
(and most rewarding)
ones are how to...

ADAPT DOMAINS
GO LARGE SCALE!
no need to thank 😊

REFERENCES
Dataset Links

The Usual Suspects
- KTH: [http://www.nada.kth.se/cvap/actions/](http://www.nada.kth.se/cvap/actions/)
- Weizmann: [http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html](http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html)
- INRIA IXMAS: [http://shivvitaladevuni.com/action_rec/ixmas_example.htm](http://shivvitaladevuni.com/action_rec/ixmas_example.htm)

The Wild Ones
- HMDB: [http://serre-lab.clps.brown.edu/resources/HMDB/](http://serre-lab.clps.brown.edu/resources/HMDB/)
- ActionBank: [http://www.cse.buffalo.edu/~jcorso/r/actionbank/](http://www.cse.buffalo.edu/~jcorso/r/actionbank/)

Surveillance Datasets
- Visor: [http://imagelab.ing.unimore.it/visor/](http://imagelab.ing.unimore.it/visor/)

ADL Datasets
- UCI ADL Dataset: [http://deepthought.ics.uci.edu/ADLdataset/adl.html](http://deepthought.ics.uci.edu/ADLdataset/adl.html)
- YouCook: [http://www.cse.buffalo.edu/~jcorso/r/youcook/](http://www.cse.buffalo.edu/~jcorso/r/youcook/)
- UESTC Senior Home Monitoring: [http://www.uestcrobot.net/senioractivity/](http://www.uestcrobot.net/senioractivity/)

RGBD Datasets
- MSR Datasets: [https://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm](https://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm)
- Cornell Activity Dataset: [http://pr.cs.cornell.edu/humanactivities/data.php#format](http://pr.cs.cornell.edu/humanactivities/data.php#format)
- LIRIS: [http://liris.cnrs.fr/voir/activities-dataset/](http://liris.cnrs.fr/voir/activities-dataset/)
- WorkoutSU-10 Exercise: [http://vpa2.sabanciuniv.edu/databases/WorkoutSU-10/](http://vpa2.sabanciuniv.edu/databases/WorkoutSU-10/)
Survey References


Work References – 1/5

Nearest Neighbor Scheme


Dynamic Time Warping


Work References – 2/5

Principal Component Analysis


Manifold Learning: Non-linear Methods


Metric Learning


Discriminative Classifier Methods


Space-Time Interest Points Based Framework


Work References – 4/5

State-Space Methods: Generative


State-Space Methods: Discriminative


State-Space Methods: Discriminative – cont’d


Mining Action Data


Using Context
Yao, Fei-Fei, “Modeling mutual context of object and human pose in human-object interaction activities”, IEEE Conference on Computer Vision and Pattern Recognition 2010