Emerging Topics in Human Activity Recognition

Michael Ryoo  NASA Jet Propulsion Laboratory
Ivan Laptev  INRIA
Greg Mori  Simon Fraser University
Sangmin Oh  Kitware

CVPR tutorial on 2014/06/23
Aaction Recognition with Bag-of-Features and Beyond

Ivan Laptev
INRIA Paris

CVPR tutorial on 2014/06/23
Computer vision grand challenge: Video understanding
Computer vision grand challenge: Dynamic scene understanding

Objects: cars, glasses, people, etc...

Actions: drinking, running, door exit, car enter, etc...

Scene categories: indoors, outdoors, street scene, etc...

Geometry: Street, wall, field, stair, etc...

Constraints
What are the challenges?

- Large variations in appearance: occlusions, non-rigid motion, viewpoint changes, clothing...

- Manual collection of training samples is prohibitive: many action classes, rare occurrence

- Action vocabulary is not well-defined
What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, viewpoint changes, clothing...

**Part I**
- Early methods
- Bag-of-features action classification

**Part II**
- Mid-level representations
- Temporal models of action recognition
- Action localization

Action Hugging:
Activities are characterized by pose
Pose estimation is difficult

Finding People by Sampling Ioffe & Forsyth, ICCV 1999

Learning to Parse Pictures of People Ronfard, Schmid & Triggs, ECCV 2002


Pictorial Structure Models for Object Recognition Felzenszwalb & Huttenlocher, 2000

Articulated pose estimation with flexible mixtures-of-parts. Y. Yang and D. Ramanan. CVPR 2011

Actions are more than poses
Appearance methods: Shape

Idea: summarize motion in video in a *Motion History Image (MHI)*:

Appearance methods: Shape

Pros:
- Simple and fast
- Works in controlled settings

Cons:
- Prone to errors of background subtraction
- Does not capture interior Structure and motion

Variations in light, shadows, clothing…
What is the background here?

Silhouette tells little about actions
Appearance methods: Motion

Recognizing action at a distance

Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, 1997
Local feature methods

- No segmentation needed
- No object detection/tracking needed
- Loss of global structure
Local feature methods: Why working?

- Find similar events in pairs of video sequences
Bag-of-Features action recognition

Extraction of Local features

Occurrence histogram of visual words

Non-linear SVM with $\chi^2$ kernel

K-means clustering ($k=4000$)

Feature quantization

Feature description

[Laptev, Marszałek, Schmid, Rozenfeld 2008]
Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”, “Indiana Jones and the Last Crusade”
Alternative local video descriptors


- P. Matikainen, R. Sukthankar and M. Hebert "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features" ICCV VOEC Workshop 2009,

Dense trajectory descriptors

[Wang et al. CVPR 2011]
Improved trajectory descriptors

[Wang and Schmid ICCV 2013]

- Removes camera motion with Background motion estimation and person detection
- Uses Fisher Vector Encoding

<table>
<thead>
<tr>
<th>Hollywood2</th>
<th>HMDB51</th>
<th>Olympic Sports</th>
<th>UCF50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jain et al. [14]</td>
<td>Jain et al. [14]</td>
<td>Jain et al. [14]</td>
<td>Shi et al. [34]</td>
</tr>
<tr>
<td>Without HD</td>
<td>Without HD</td>
<td>Without HD</td>
<td>Without HD</td>
</tr>
<tr>
<td>63.0%</td>
<td>55.9%</td>
<td>90.2%</td>
<td>90.5%</td>
</tr>
<tr>
<td>With HD</td>
<td>With HD</td>
<td>With HD</td>
<td>With HD</td>
</tr>
<tr>
<td>64.3%</td>
<td>57.2%</td>
<td>91.1%</td>
<td>91.2%</td>
</tr>
</tbody>
</table>
Efficient features for action recognition

[Kantorov and Laptev CVPR 2014]

- Use sparse motion vectors from video compression.
- >100x speed-up of video feature extraction.

<table>
<thead>
<tr>
<th>Hollywood2</th>
<th>Acc.</th>
<th>Feat. (fps)</th>
<th>Quant. (fps)</th>
<th>Total (fps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF FLANN(4-32)</td>
<td>55.8%</td>
<td>52.4</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>MF VLAD(4)</td>
<td>56.7%</td>
<td>167.5</td>
<td>84.0</td>
<td></td>
</tr>
<tr>
<td>MF FV(32)</td>
<td>58.2%</td>
<td>40.9</td>
<td>32.9</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>59.9%</td>
<td>1.2</td>
<td>5.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UCF 50</th>
<th>Acc.</th>
<th>Feat. (fps)</th>
<th>Quant. (fps)</th>
<th>Total (fps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF FLANN(4-32)</td>
<td>81.6%</td>
<td>52.4</td>
<td>48.1</td>
<td></td>
</tr>
<tr>
<td>MF VLAD(4)</td>
<td>80.6%</td>
<td>671.4</td>
<td>314.6</td>
<td></td>
</tr>
<tr>
<td>MF FV(32)</td>
<td>82.2%</td>
<td>171.3</td>
<td>132.8</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>85.6%</td>
<td>5.1</td>
<td>1.8</td>
<td></td>
</tr>
</tbody>
</table>
What are the challenges?

- Large variations in appearance: occlusions, non-rigid motion, viewpoint changes, clothing...

Action Hugging:

Part I
- Early methods
- Bag-of-features action classification

Part II
- Mid-level representations
- Temporal models of action recognition
- Action localization
Mid-level action representations I

Discovering discriminative action parts from mid-level video representations
Raptis, Kokkinos, and Soatto, CVPR 2012.

• Groups point trajectories into tentative action parts by similarity in motion and appearance.

• Learns discriminative model with latent assignment of action parts.

- Improved classification
- Localization of discriminative action parts
Mid-level action representations II

Representing Videos using Mid-level Discriminative Patches
Jain, Gupta, Rodriguez and Davis, CVPR 2013

- Mines discriminative and representative patches from action videos
- Aligns patches and transfers additional annotation to test samples

High-ranked discriminative spatio-temporal patches

Discriminative patches for “Gangnam” style

Object Localization

2D Pose

3D Pose

Annotated selected e-SVM patch for label transfer

Action Localization
Mid-level action representations

Conclusions:

- Enable part-based action analysis
  - localization of action parts
  - annotation transfer

- Current classification performance is inferior to the state of the art obtained with Fisher Vector + Improved dense trajectories.

- CNNs will take over soon?
  Probably Yes, but:
  - What’s a good network architecture?
  - Where to get sufficient training data?

Some work is on the way:
Large-scale Video Classification using Convolutional Neural Networks Karpathy, Shetty, Toderici, Sukthankar, Leung and Fei-Fei, CVPR 2014 (Wednesday)
**Temporal structure of actions I**

Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification, J.C. Niebles, C.-W. Chen and L. Fei-Fei, ECCV 2010

- Latent SVM model with temporal action parts.
- Enables temporal localization of action parts
Learning Latent Temporal Structure for Complex Event Detection.
Kevin Tang, Li Fei-Fei and Daphne Koller, CVPR 2012

- Modeling of longer events such as Grooming an animal
- Discriminatively-trained Markov model
- Aims to infer and learn latent temporal structure of actions
Temporal structure of actions III

Poselet Key-framing: A Model for Human Activity Recognition.
Raptis and Sigal, In CVPR 2013

- Models actions as a sparse sequence of spatio-temporally localized key-frames
Scenario-based video event recognition by constraint flow.
Kwak, Han and Han, In CVPR 2011

- Models event by temporal logical composition of simple actions

<table>
<thead>
<tr>
<th>Composite event</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroke[A]</td>
<td>(Meet[A, ball] $&lt;$ Part[A, ball]) $\land$ Swing[A]</td>
</tr>
<tr>
<td>TennisPlay$^+$</td>
<td>Service[ser] $&lt;$ (#</td>
</tr>
</tbody>
</table>

- First service
- Second service (may not happen)
- Unknown number of strokes
Temporal structure of actions IV

Scenario-based video event recognition by constraint flow.
Kwak, Han and Han, In CVPR 2011
Scenario-based video event recognition by constraint flow. Kwak, Han and Han, In CVPR 2011

Temporal structure of actions IV

Transaction
TakeItem [customer]
- With[customer,item]
- Move[customer,out]
- Move[customer,desk]

TakeMoney [customer]
- Move[customer,out]
- With[customer,money]
- Without[customer,money]

BringMoney [cashier]
- Without[cashier,money]
- Move[cashier,desk]
- With[cashier,money]

TakeMoney [cashier]
- Move[cashier,desk]
- With[cashier,money]
- Without[cashier,money]

BringMoney [customer]
- Without[customer,money]
- Move[customer,desk]
- With[customer,money]

ScanItem [cashier]
- Move[cashier,scanner]
- With[cashier,item]

BringItem [customer]
- Move[customer,out]
- Move[customer,desk]
- With[customer,item]

"Missing"

"No pay"
Temporal activity models

Conclusions:

- Enable temporal analysis of sub-actions
- Enable modeling of long-term activities with variable structure of actions
- Current classification performance is inferior to the state of the art
- Open problems:
  - How to automatically discover and learn scenarios?
  - What are the meaningful action units for a target activity?
- Literature on action localization is relatively sparse
  => good area to make impact!
  See upcoming action recognition and localization challenge:
# References to some related work

## Local features and bag-of-features representations
- Laptev and Lindeberg “Space-time interest points”, *In ICCV’03.*
- Schüldt et al., "Recognizing human actions: A local SVM approach", *In ICPR’04.*
- Niebles et al., "Unsupervised learning of human action categories using spatial-temporal words", *In BMVC’06.*
- Laptev et al., “Learning realistic human actions from movies”, *In CVPR’08.*
- Jain et al., "Better exploiting motion for better action recognition", *In CVPR’13.*
- Kantorov and Laptev, "Efficient feature extraction, encoding and classification for action recognition “, *In CVPR’14.*

## Mid-level representations
- Liu et al., "Recognizing Human Actions by Attributes", *In CVPR’11.*
- Raptis et al., Discovering discriminative action parts from mid-level video representations." *In CVPR’12.*
- Jain et al., Representing Videos using Mid-level Discriminative Patches. *In CVPR’13.*

## Temporal action models
- Ryoo and Agarwal, Recognition of composite human activities through context-free grammar based representation, *CVPR’06*
- Laxton et al., Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video, *CVPR’07.*
- Niebles et al., Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. *In ECCV 2010.*
- Khamis et al., Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition, *ECCV 2012.*
- Raptis and Sigal. Poselet Key-framing: A Model for Human Activity Recognition. *In CVPR 2013*
- Amer et al., Monte Carlo Tree Search for Scheduling Activity Recognition, *In ICCV 2013.*
What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, viewpoint changes, clothing...

  Action *Hugging*:

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence

  Action *Open*:
What are the challenges?

- Large variations in appearance: occlusions, non-rigid motion, viewpoint changes, clothing...

- Manual collection of training samples is prohibitive: many action classes, rare occurrence

- Action vocabulary is not well-defined
Where to get training data?

- Shoot actions in the lab
  - KTH dataset
  - Weizman dataset,…
  - Limited variability
  - Unrealistic

- Manually annotate existing content
  - HMDB, Olympic Sports,
    - Very time-consuming
  - UCF50, UCF101, …

- Use readily-available video scripts
  - Scripts are available for 1000’s of hours of movies and TV-series
  - Scripts describe dynamic and static content of videos
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...
Scripts as weak supervision

Challenges:

- Imprecise temporal localization
- No explicit spatial localization
- NLP problems, scripts ≠ training labels

“… Will gets out of the Chevrolet. …” vs. Get-out-car
“… Erin exits her new truck…”
Previous work

Sivic, Everingham, and Zisserman, "Who are you?" -- Learning Person Specific Classifiers from Video, In CVPR 2009.

Buehler, Everingham, and Zisserman "Learning sign language by watching TV (using weakly aligned subtitles)", In CVPR 2009.

Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]

Rick walks up behind Ilse
Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]
Formulation: Cost function

\[
\frac{1}{N} \sum Z - \phi(X)w - b^2_F + \lambda_1 \text{Tr}(w^T w)
\]

Actor labels:
- Rick
- Ilsa
- Sam

Actor image features

Actor classifier
Weak supervision from scripts:

Person $p$ appears at least once in clip $N$:

$$\sum_{n \in N_i} z_{np} \geq 1$$
Formulation: Cost function

\[ \frac{1}{N} \| Z - \phi(X)w - b \|_F^2 + \lambda_1 \text{Tr}(w^T w) \]

\[ + \frac{1}{N} \| T - \psi(X)v - c \|_F^2 + \lambda_2 \text{Tr}(v^T v) \]

Weak supervision from scripts:

Action \( a \) appears at least once in clip \( N \):

\[ \sum_{n \in N_i} t_{na} \geq 1 \]

\( t_{11} \ldots t_{1A} \)
\( \vdots \)
\( t_{n1} \ldots t_{nA} \)
\( t_{n2} \ldots t_{nA} \)
\( t_{n3} \ldots t_{nA} \)
\( \vdots \)
\( t_{N1} \ldots t_{NA} \)

\( a = \text{Walk} \)
**Formulation: Cost function**

$$\min_{Z,T,w,b,v,c} \frac{1}{N} \| Z - \phi(X) w - b \|_F^2 + \lambda_1 \, Tr(w^T w)$$

$$+ \frac{1}{N} \| T - \psi(X) v - c \|_F^2 + \lambda_2 \, Tr(v^T v)$$

**Weak supervision from scripts:**

Person p appears in clip N:

$$\sum_{n \in N_i} z_{np} \geq 1$$

Action a appears in clip N:

$$\sum_{n \in N_i} t_{na} \geq 1$$

Person p and Action a appear in clip N:

$$\sum_{n \in N_i} z_{np} t_{na} \geq 1$$
Results for Person Labelling

Casablanca (17 character names)

American beauty (11 character names)
Results for Person + Action Labelling

Casablanca, Walking
Finding Actions and Actors in Movies

[Bojanowski, Bach, Laptev, Ponce, Sivic, Schmid, 2013]
References to some related work

Video and language
- Duchenne et al., Automatic Annotation of Human Actions in Video, ICCV 2009.
- Buehler et al., Learning sign language by watching TV (using weakly aligned subtitles), CVPR 2009.
- Das et al., A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching, CVPR’13.
- Ramanathan et al., Video Event Understanding using Natural Language Descriptions, ICCV 2013.
- Bojanowski et al., Finding Actors and Actions in Movies, ICCV 2013
- Ramanathan et al., Joint person naming in videos and coreference resolution in text, ECCV 2014
- Bojanowski et al., Weakly supervised action labeling in videos under ordering constraints, ECCV 2014
What are the challenges?

- Large variations in appearance: occlusions, non-rigid motion, viewpoint changes, clothing...

- Manual collection of training samples is prohibitive: many action classes, rare occurrence

- Action vocabulary is not well-defined
What are the challenges?

- Large variations in appearance: occlusions, non-rigid motion, viewpoint changes, clothing...

  Action Hugging:

- Manual collection of training samples is prohibitive: many action classes, rare occurrence

  Action Open:

- Action vocabulary is not well-defined
Where is computer vision going next?
Is image/video classification the right problem?

- Is action vocabulary well-defined?

Examples of “Open” action:

- What granularity of action vocabulary shall we consider?
Do we want to learn *person-throws-cat-into-trash-bin* classifier?

Source: http://www.youtube.com/watch?v=eYdUZdan5i8
Can object recognition help?
Limitations of Current Methods
Limitations of Current Methods

What is unusual in this scene?

Is this scene dangerous?

What is intention of this person?
Next challenge

Shift the focus of computer vision

Object, scene and action recognition

Is this a picture of a dog? Is the person running in this video?

Recognition of objects’ function and people’s intentions

What people do with objects? How they do it? For what purpose?

Enable new applications
Motivation

• Exploit the link between human pose, action and object function.

• Use human actors as active sensors to reason about the surrounding scene.
Scene semantics from long-term observation of people

V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, A. Efros
Goal

Recognize objects by the way people interact with them.

Lots of person-object interactions, many scenes on YouTube.

Time-lapse “Party & Cleaning” videos

Semantic object segmentation

- Sofa
- Shelf
- Floor
- Table
- Tree
- Wall
New “Party & Cleaning” dataset
Goal

Recognize objects by the way people interact with them.

Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

Semantic object segmentation

- Sofa
- Shelf
- Floor
- Table
- Tree
- Wall
Pose histogram
Some qualitative results
Using our model as pose prior

Given a bounding box and the ground truth segmentation, we fit the pose clusters in the box and score them by summing the joint’s weight of the underlying objects.
Using our model as pose prior