Action class detection and recognition in realistic video

[ICCV07]

Ivan Laptev
IRISA/INRIA, Rennes, France

http://www.irisa.fr/vista/Equipe/People/Ivan.Laptev.html
E-team: Visual Saliency topics overview

- Object classes learning and recognition: 7 GRAZ, UFR, VISTA, KTH, Cambridge, IMEDIA, UvA
- Interest points, Local features, patches: 6 GRAZ, UFR, IMEDIA, VISTA, ENST, Cambridge, UvA
  - Relations between local features: 5 UFR, IMEDIA, UvA, KTH, VISTA
- Spatio-temporal salient regions: 4 SZTAKI, UCL, VISTA, IMEDIA
  - Texture: 3 SZTAKI, KTH, ENST
  - Relations to Human visual attention: 3 KTH, UCL, SZTAKI
    - Copy detection: 3 VISTA, IMEDIA, UCL
  - Eyetracking, perceptual interface: 1 UCL
Human actions: Motivation

- Huge amount of video is available and growing
- Human actions are major events in movies, TV news, personal video …

Action recognition useful for:
- Content-based browsing
  e.g. fast-forward to the next goal scoring scene
- Video recycling
  e.g. find “Bush shaking hands with Putin”
- Human scientists
  influence of smoking in movies on adolescent smoking
What are human actions?

**Definition 1:**
- **Physical body motion**
  

**Definition 2:**
- **Interaction with environment on specific purpose**
  
  same physical motion -- different actions depending on the context
Context defines actions
Challenges in action recognition

- Similar problems to static object recognition:
  variations in views, lightning, background, appearance, …

- Additional problems: variations in individual motion; camera motion

Example:

Drinking

Difference in shape

Difference in motion

Smoking

Both actions are similar in overall shape (human posture) and motion (hand motion)

Data variation for actions might be higher than for objects

But: Motion provides an additional discriminative cue
Action dataset and annotation

• No datasets with realistic action classes are available
• This work: first attempt to approach action detection and recognition in real movies: “Coffee and Cigarettes”; “Sea of Love”

“Drinking”: 159 annotated samples
“Smoking”: 149 annotated samples

Spatial annotation

Temporal annotation
“Drinking” action samples

training samples  test samples
Actions == space-time objects?

“stable-view” objects

“atomic” actions

car exit  phoning  smoking  hand shaking  drinking

Objective: take advantage of space-time shape
Action features

- HOG features
- HOF features

Block-histogram features:
- $f = \frac{H}{\delta x}$
- $f = (H_1, H_2)$
- $f = (H_1, H_2, H_3, H_4)$

$\begin{pmatrix} X \\ Y \\ T \end{pmatrix}$

$\Delta T$

$\Delta Y$

$\Delta X$

First frame

Key-frame

Last frame
Histogram features

HOG: histograms of oriented gradient

HOF: histograms of optic flow

\~10^7 cuboid features
Choosing 10^3 randomly

4 grad. orientation bins

4 OF direction bins + 1 bin for no motion
Action learning

- Efficient discriminative classifier [Freund&Schapire’97]
- Good performance for face detection [Viola&Jones’01]
Action classification test

Random motion patterns

- Additional shape information does not seem to improve the space-time classifier
- Space-time classifier and static key-frame classifier might have complementary properties
Classifier properties

Compare selected features by

- Space-time action classifier (HOF features)
- Static key-frame classifier (HOG features)

Training output: Accumulated feature maps
Keyframe priming

Training

Positive training sample

Negative training samples

Test
Action detection

Test set:
- 25min from “Coffee and Cigarettes” with GT 38 drinking actions
- No overlap with the training set in subjects or scenes

Detection:
- search over all space-time locations and spatio-temporal extents

Keyframe priming

Similar approach to Ke, Sukthankar and Hebert, ICCV05
Test episode
20 most confident detections
Summary

- First attempt to address human action in real movies
- Action detection/recognition seems possible under hard realistic conditions (variations across views, subjects, scenes, etc...)
- Separate learning of shape/motion information results in a large improvement (overfitting?)

Future

- Need realistic data for 100’s of action classes:
  -> (semi-) automatic action annotation from movie scripts
  [M.Everingham, J.Sivic and A.Zisserman BMVC06]
- Explicit handling of actions under multiple views
- Combining action classification with text