Grenoble INP ENSIMAG

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA | centre de recherche GRENOBLE - RHÔNE-ALPES

Human Action Description

Master Student:

Javier Montoya: INPG, INRIA Rhône-Alpes

Supervisors:

Alexander Kläser: INRIA Rhône-Alpes

Cordelia Schmid: INRIA Rhône-Alpes

Sep 7th, 2010

Outline

- **Context and Goal.**
- Human Action Description.
- Experimental Results:
  - Action Classification.
  - Action Localization.
- Conclusions.

Context and Goal

- **Goal**:
  - Human action recognition in realistic video sequences.
- **Action Recognition:**
  - *Action Classification*: assign an action class label to sequences with *pre-defined* temporal extent.
  - *Action Localization:* localize human actions in space (the 2D image region) and in time (temporal extent).

Human Actions

- Variations in motion, view-point, illumination, camera ego-motion, and partial occlusions.



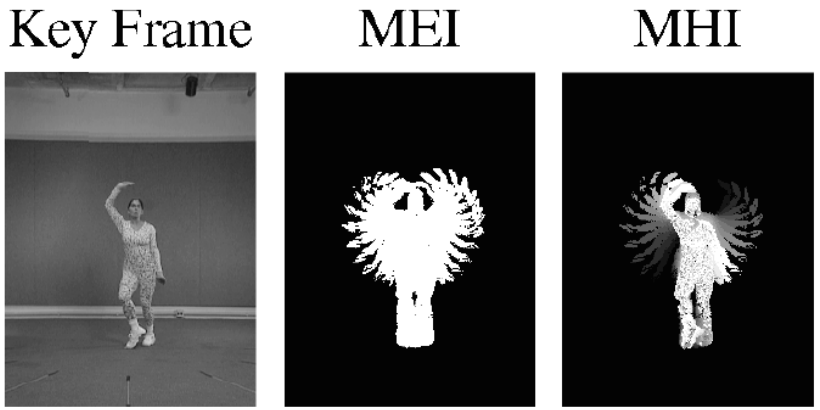
- Large intra-class variabilities due to the anthropometric differences across humans.



- Large volume of video data (an exhaustive search to localize an action is computationally expensive).

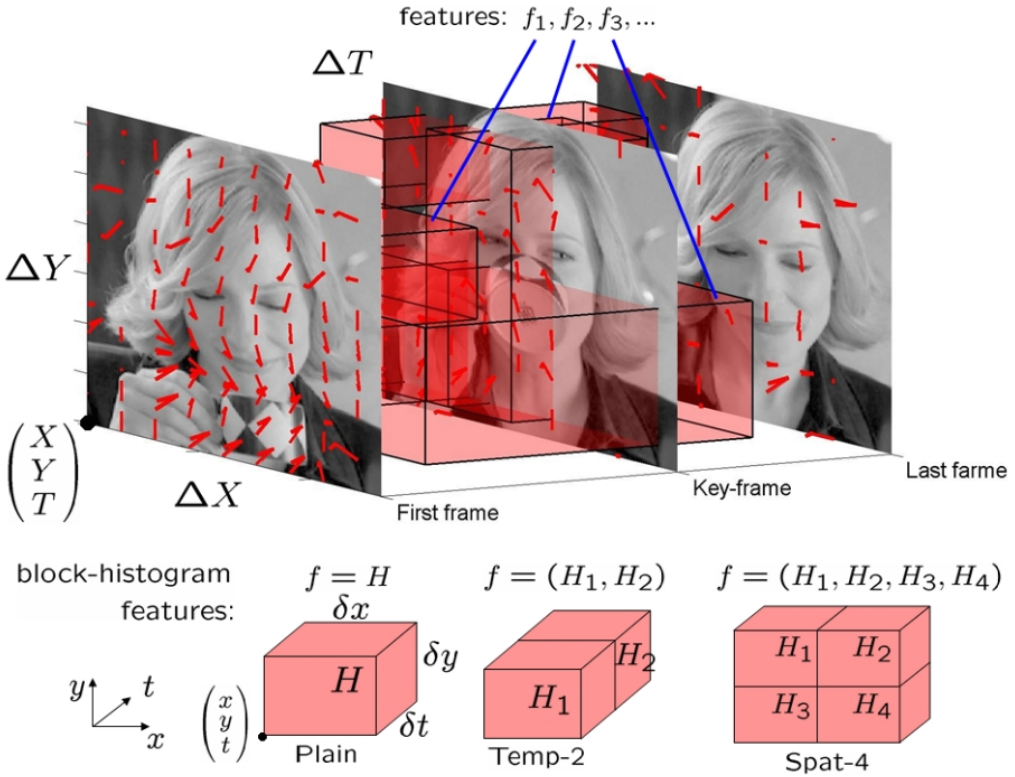
Related Work [Poppe 2010]

- Global Representations:
  - The human performing the action is localized.
  - The entire action region is encoded by a descriptor.



[Bobick et al., 2001]

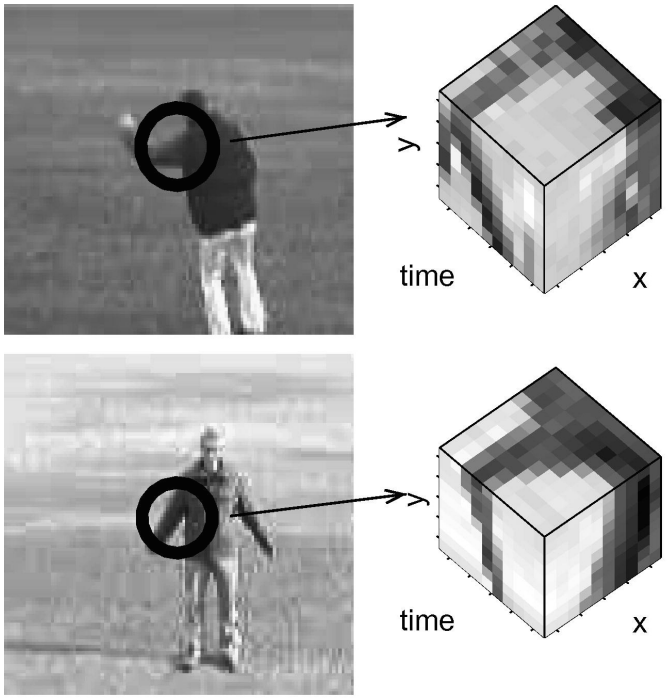
[Blank et al., 2005]



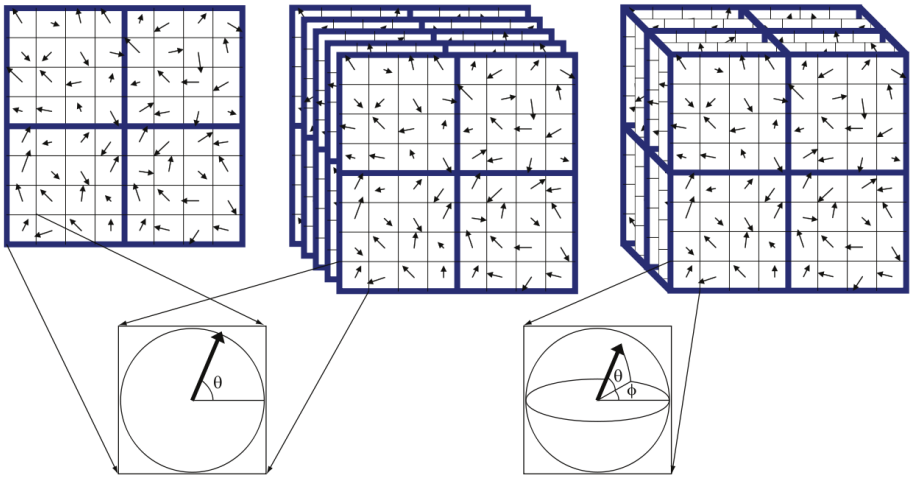
[Laptev et al., 2007]

Related Work [Poppe 2010]

- Local Representations:
  - Local features are detected and computed.
  - A video sequence is represented as ''bag of features''.



[Laptev et al., 2007]



[Scovanner et al., 2007]

Contributions

- **Challenges & Limitations of current methods**:
  - Current methods do not account for:
    - Actions consisting of re-occurring patterns.
    - Actions not following a clear sequential pattern (nevertheless, they can contain key motions).
  - Experiments often done on controlled video data.
- **Contributions**:
  - **A novel human centric way to represent actions temporally as a loose collection of movement patterns.**
  - **Human action recognition performed in controlled and realistic video data.**
  - **Evaluation and validation on public datasets involving 14 different action types.**

Outline

- Context and Goal.
- **Human Action Description.**
- Experimental Results:
  - Action Classification.
  - Action Localization.
- Conclusions.

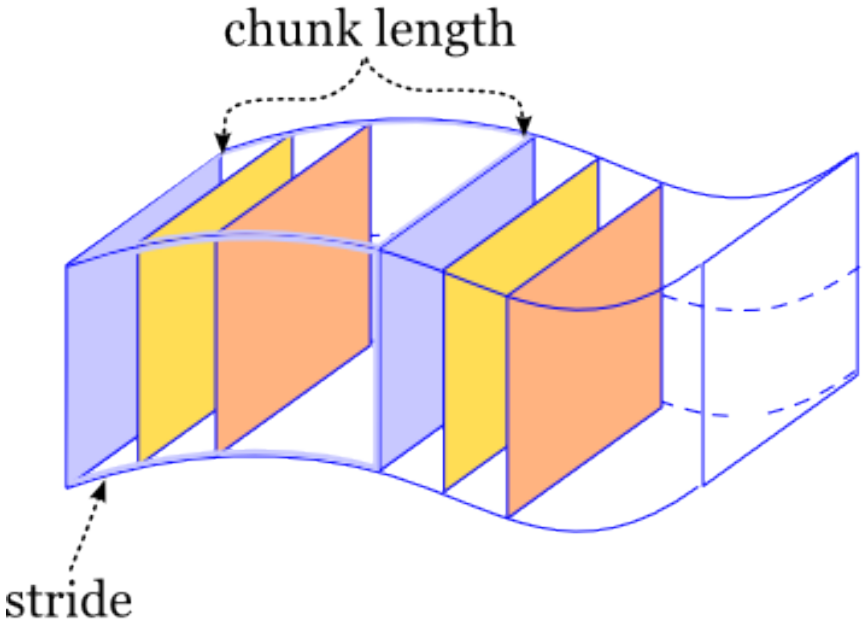
Human Action Description

1. Detect and track humans throughout time [Kläser et al. 2010]:
    1. Pedestrian or upper body detector [Dalal et al. 2005].
    2. Linking detections together + smoothing/interpolation.
    3. Track classifier to increase precision for high recall.
2. Represent *human tracks* as set of overlapping *chunks* (extracted at fixed time steps and with fixed lengths).
3. *Chunk* descriptor [Kläser et al. 2010]
    1. Combination of *appearance* and *motion* information using a descriptor based on histograms of 3D gradient orientations.
4. Classification/Localization.

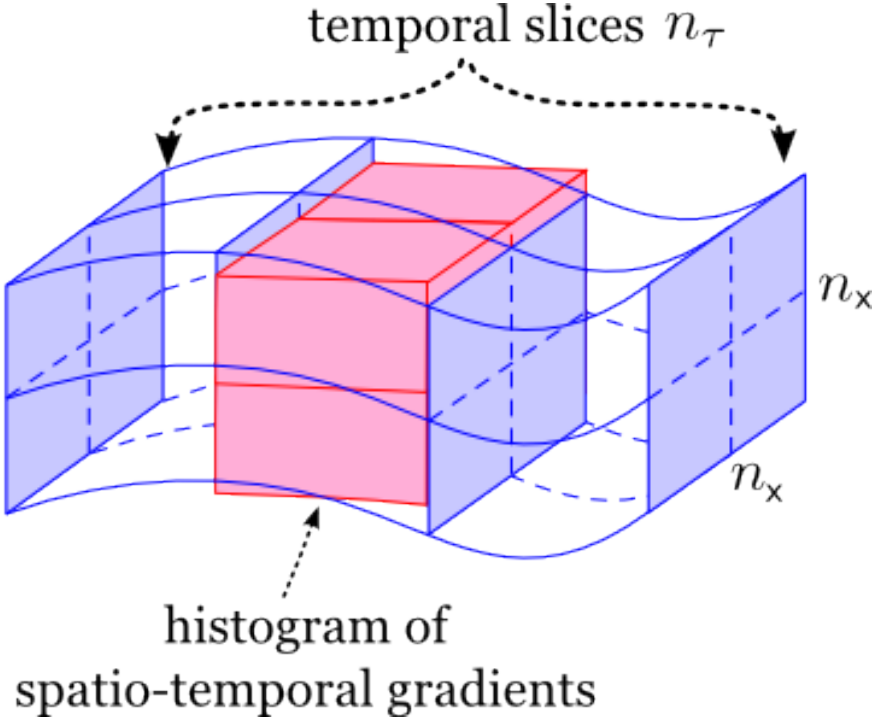
Human Action Representation

Action representation as a loose collection of movement patterns:

Track chunk



Chunk Descriptor



Human Action Classification

- **Training:**
  - SVM classifier.
  - *Positive* and *negative* sequences are given which contain a particular action or not.
  - Chunks belonging to a positive sequences are used as positive training samples; others are negatives.
- **Testing:**
  - *All* chunk descriptors of a track are classified => probability score p(Chunk|Action).
  - Chunk scores are *summed up* and *averaged*.
  - Classifier that *maximizes* the *average score* determines action class of video sequence.

Human Action Localization

- **Assumption**: chunks that overlap with the action obtain a higher score.
- Localization by *clustering* chunks based on their classification scores.
- Different clustering approaches:
  - Multi-level clustering.
  - Peak Threshold clustering.
  - Variable band-width Mean-shift clustering [Comaniciu et al. 2002].

Outline

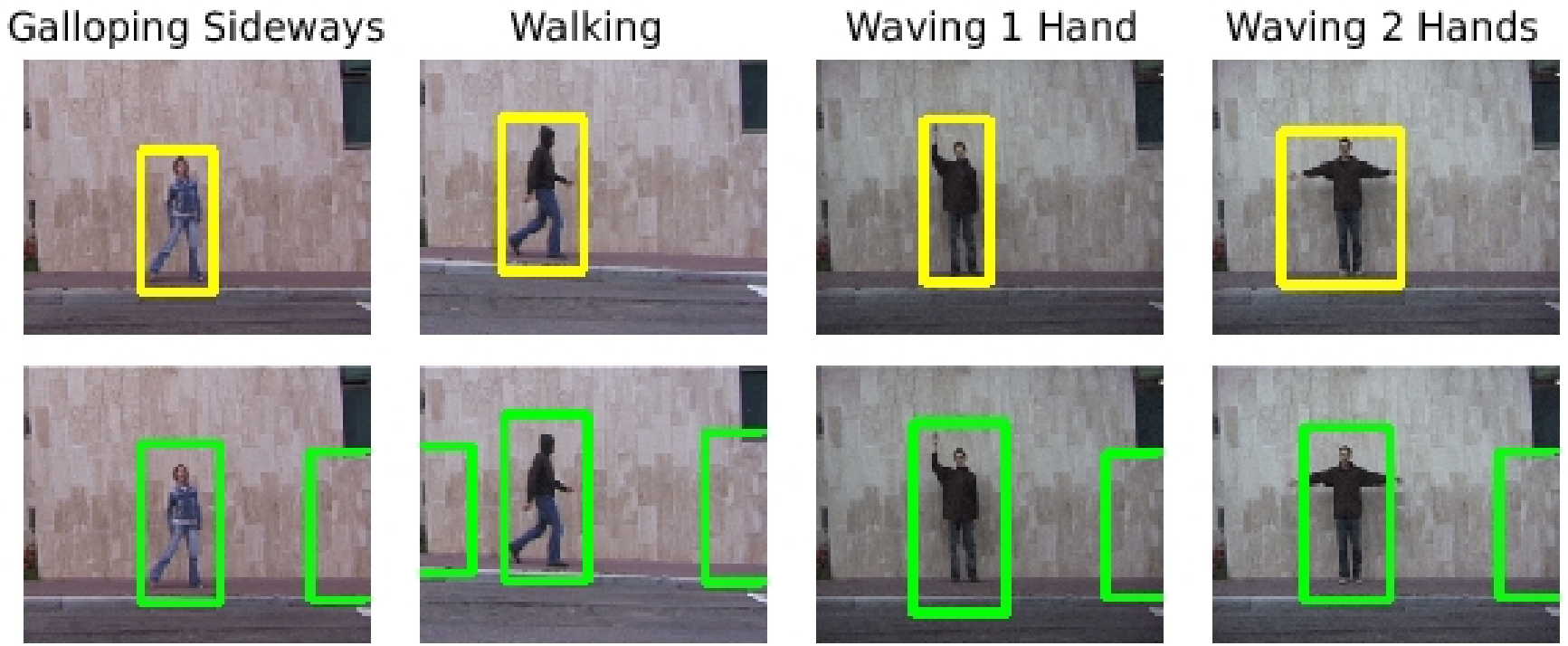
- Context and Goal.
- Human Action Description.
- **Experimental Results:**
  - **Action Classification.**
  - Action Localization.
- Conclusions.

Action Classification Results

- **Weizmann dataset** [Blank et al., 2005]**:**
  - 9 type of actions performed by 9 individuals.
  - Performance measure: *average classification accuracy.*

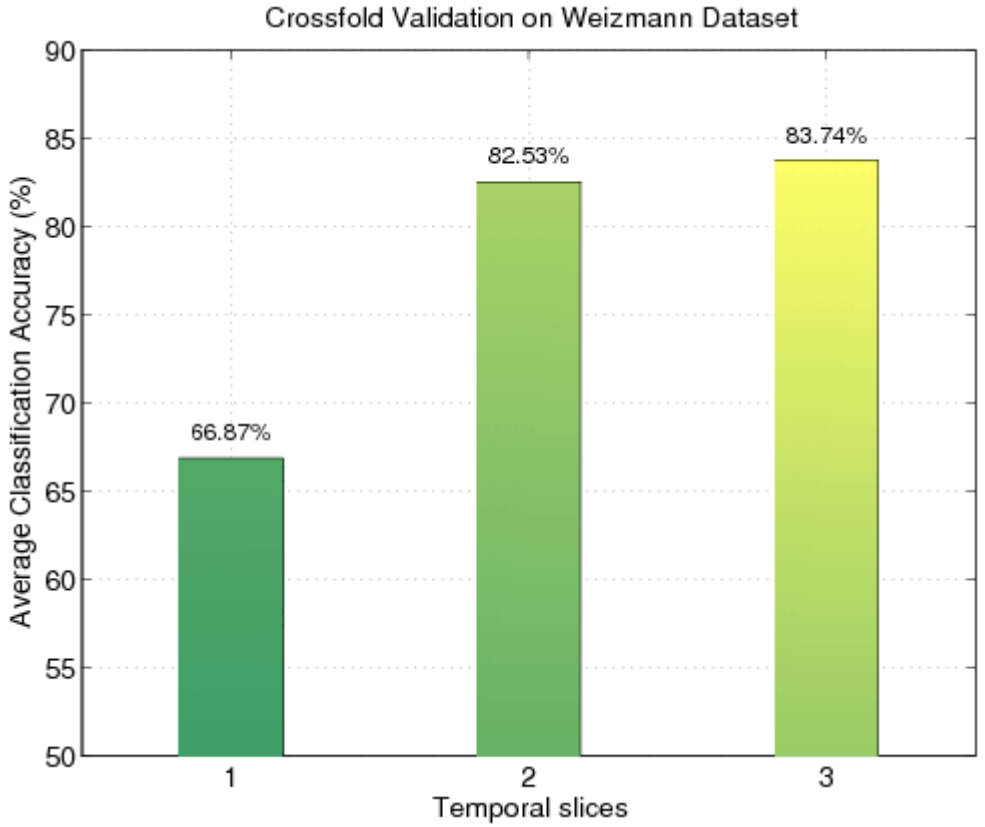
Detection Results

- Yellow: ground truth tracks (via foreground segmentation).
- Green: automatic tracks.



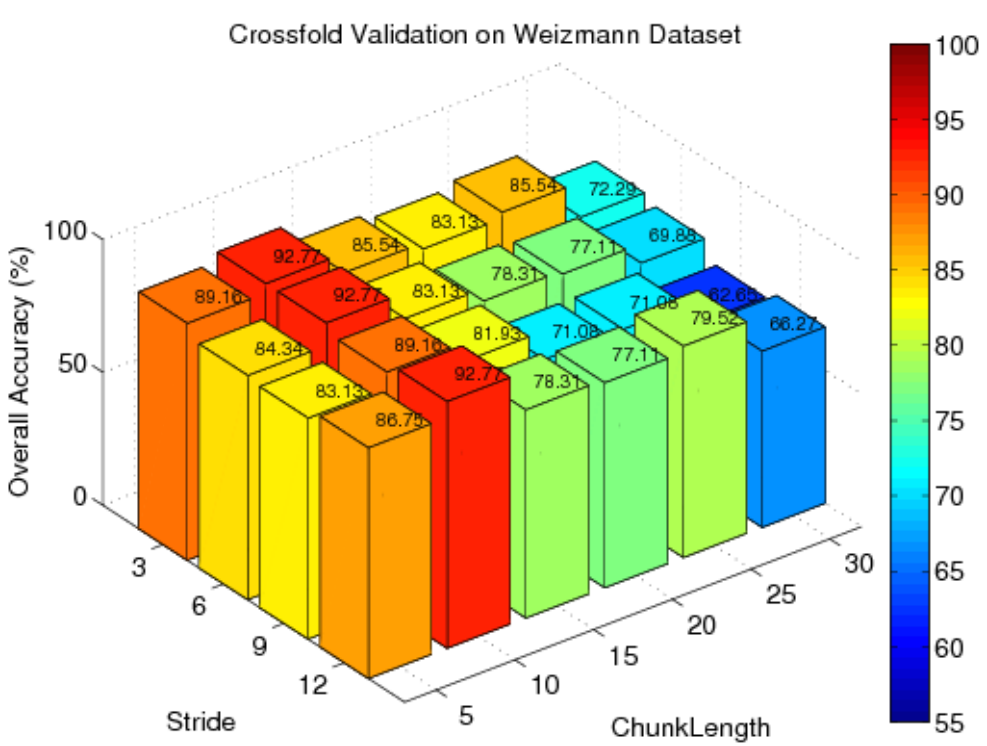
Baseline

- Global action representation [Kläser et al. 2010].
- Evaluation of different number of temporal slices (n_t).
- Number of spatial cells fixed to value suggested by [Kläser et al. 2010] ($n_x=5$) .

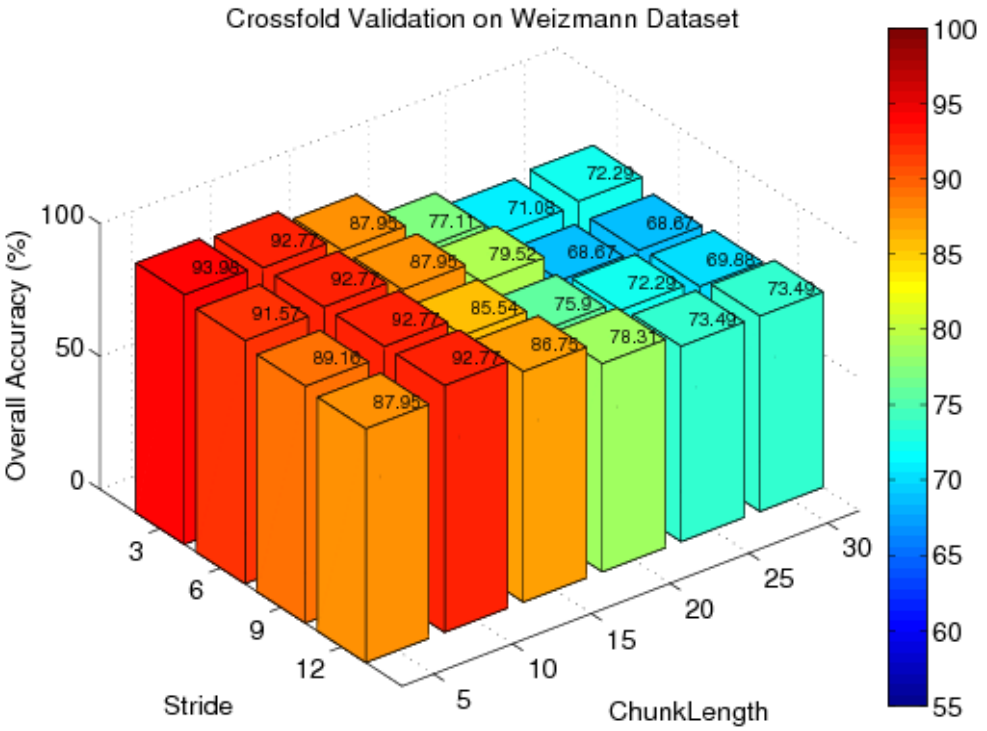


Parameter Evaluation

- Systematic evaluation of *stride* and *length* parameters.
- Analyze the gain of modeling actions as a loose collection of movement patterns.



Temporal divisions ($n_t = 1$)



Temporal divisions ($n_t = 2$)

Comparison to the state-of-the-art

Comparison of action classification results with state-of-the-art methods.

Method	Accuracy
Niebles [NL07]	72.8%
Dollar [DRCB05]	86.7%
Ali [ABS07]	92.6%
Our Approach	94.0%
Jhuang [JSWP07]	98.8%
Schindler [SG08]	100.0%

Outline

- Context and Goal.
- Human Action Description.
- **Experimental Results:**
  - Action Classification.
  - **Action Localization.**
- Conclusions.

Action Localization Results

- **Coffee and Cigarettes dataset** [Laptev et al., 2007]**:**
  - *Drinking* and *smoking* actions.
  - Total test time: *drinking* ~24 min, *smoking* ~21 min.
  - Ground truth annotations: 3D cuboids + keyframe.
  - Evaluation: overlap between *GT* cuboid and track segment (> 0.2).

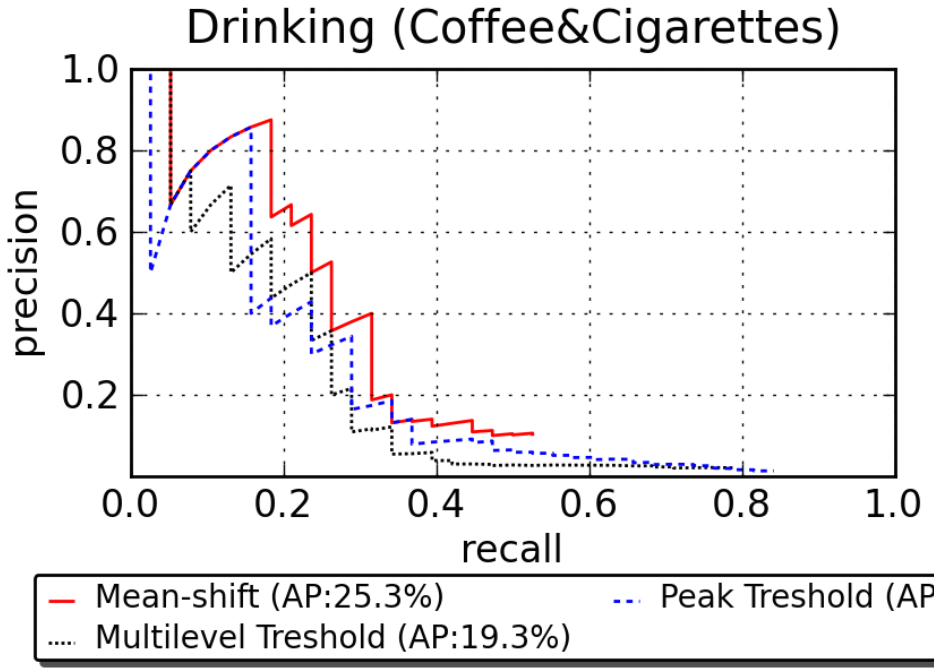


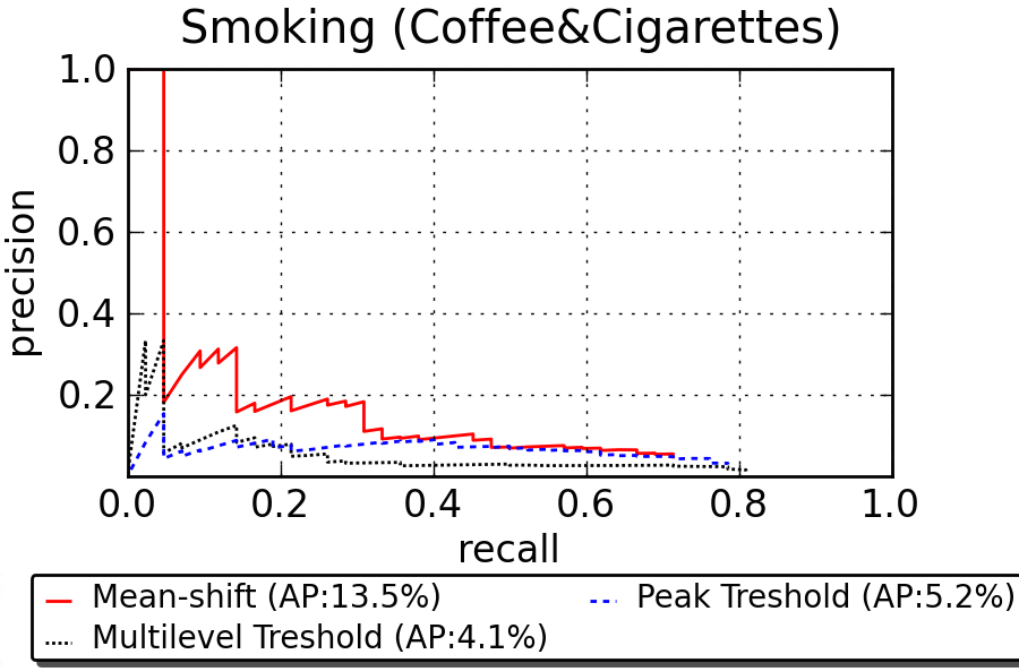

Detection Results

- Humans usually visible with their upper body.
- Detector of [Dalal et al. 2005] but trained for upper bodies as proposed by [Kläser et al. 2010].

Evaluation of localization strategies

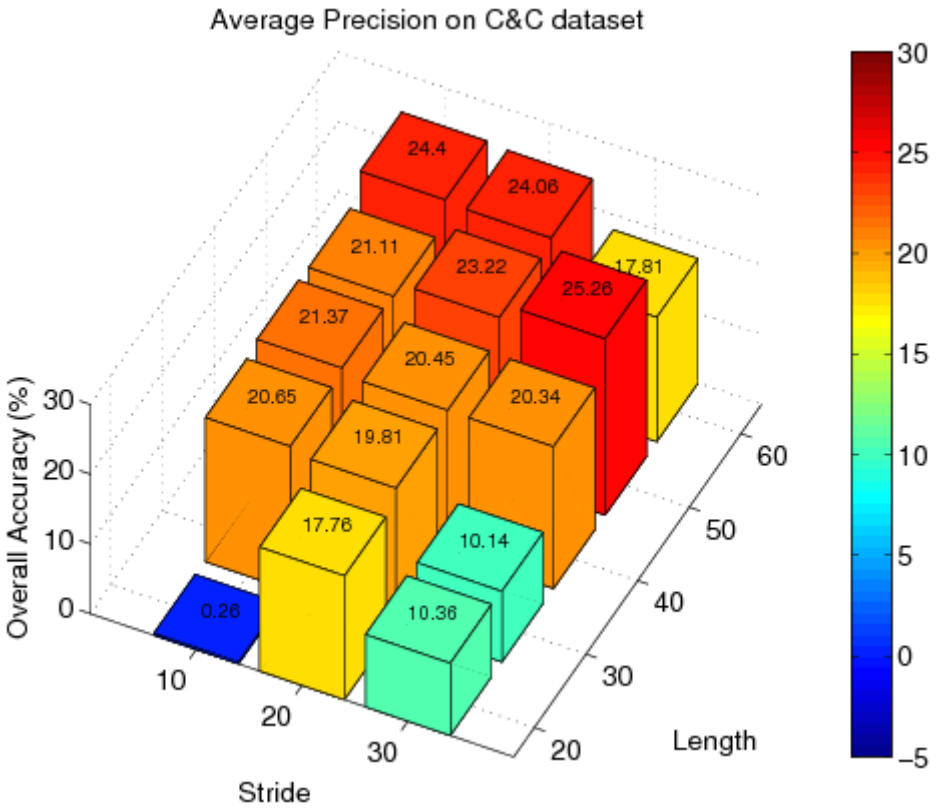
- Comparison of different clustering strategies for localization.
- For each strategy, best combination of *stride* and chunk *length* is plotted.
- Mean-shift is the most versatile.





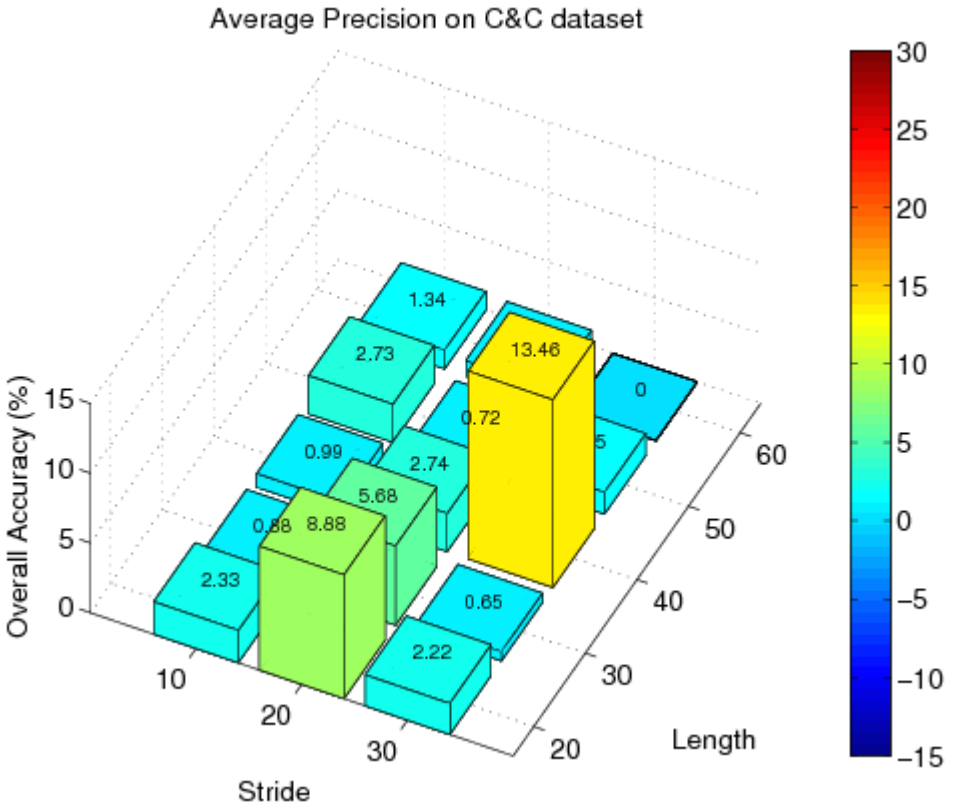
Parameter Evaluation

drinking



Temporal divisions (n_t=3)

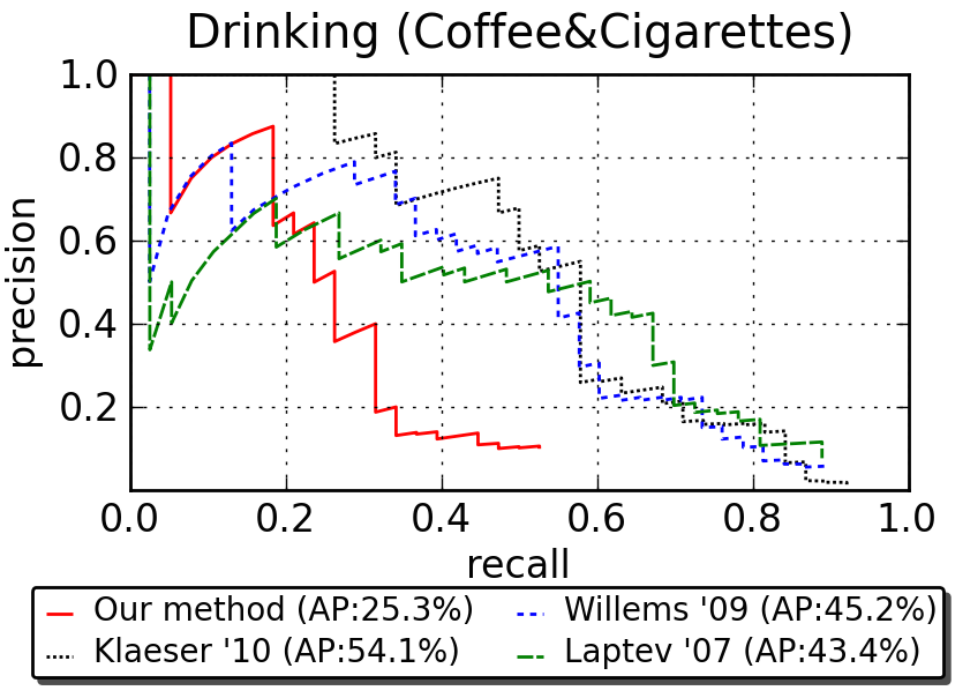
smoking

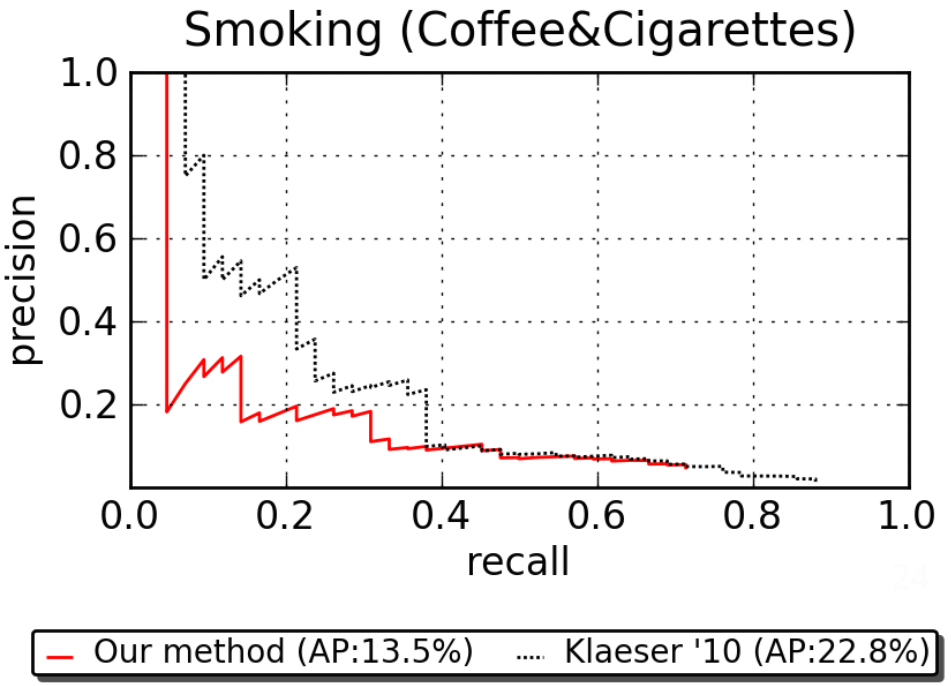


Temporal divisions (n_t=3)

Comparison to state-of-the-art

- Comparison with: [Kläser et al. 2010, Willems et al. 2009, Laptev et al. 2007].
- No gain in modeling actions as a loose collection of movement patterns.
  - Difficulty to capture different speeds with fixed chunk length.
  - Global representation more appropriate.





Action Localization Results

- **High ranked *drinking* detections:**

- **High ranked *smoking* detections:**

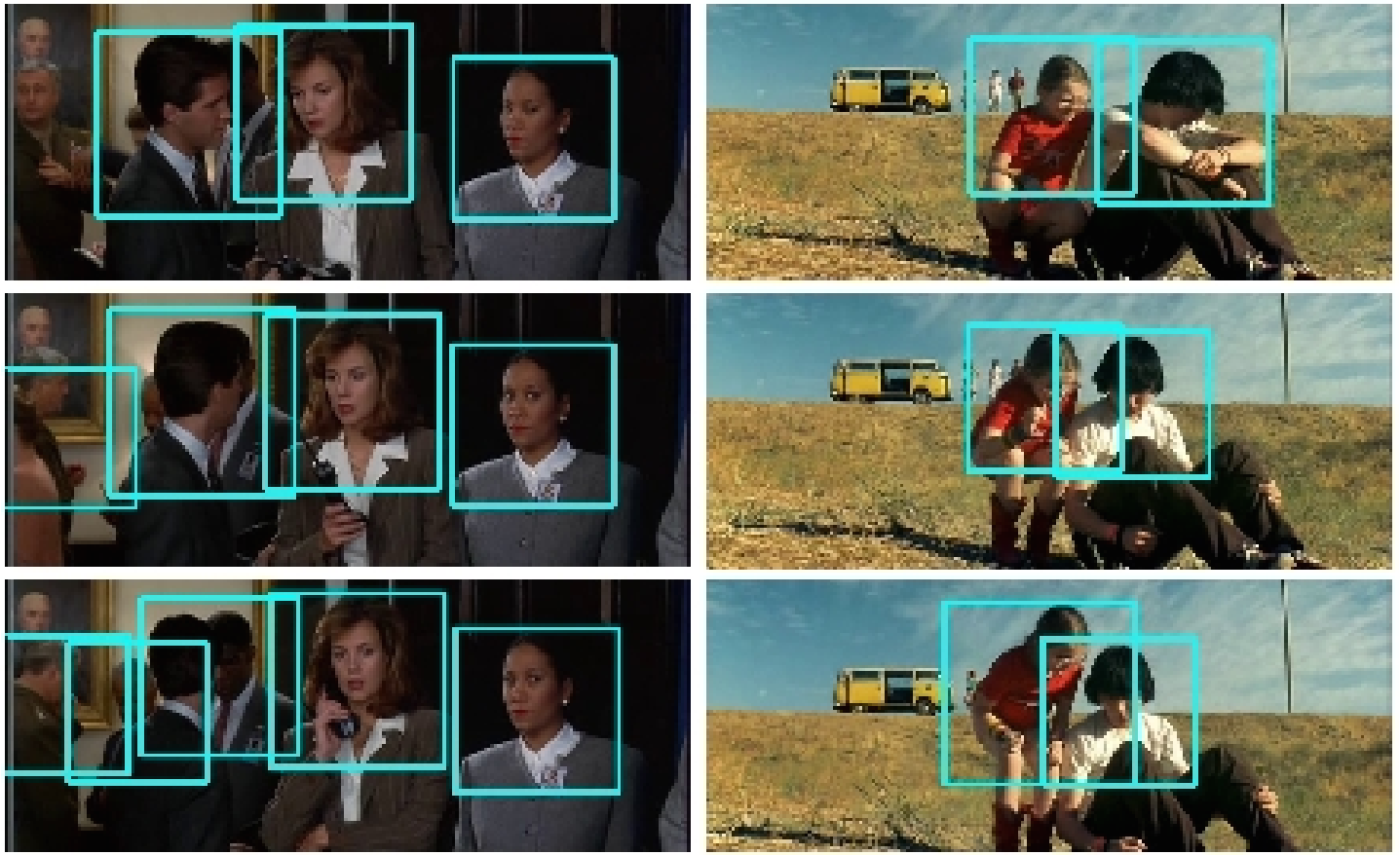
Action Localization Results

- **Hollywood Localization dataset** [Kläser et al., 2010]**:**
  - *Answer phone* and *stand up* actions.
  - Total test time: *answer phone* ~17.5 min, *stand up* ~39 min.
  - Ground truth: start, keyframe (2D box), end.
  - Evaluation: overlap between *GT* 2D box/time annotation and track segment (> 0.2).



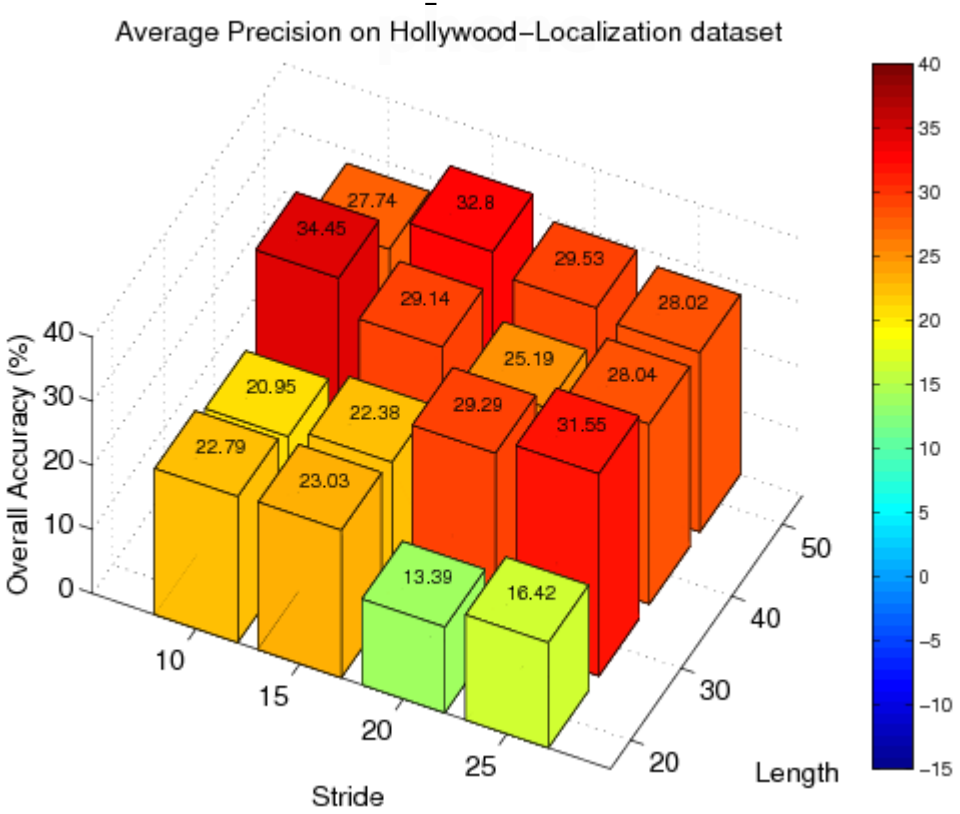

Detection Results

- Humans usually visible with upper body.
- Detector of [Dalal et al. 2005] but trained for upper bodies as proposed by [Kläser et al. 2010].



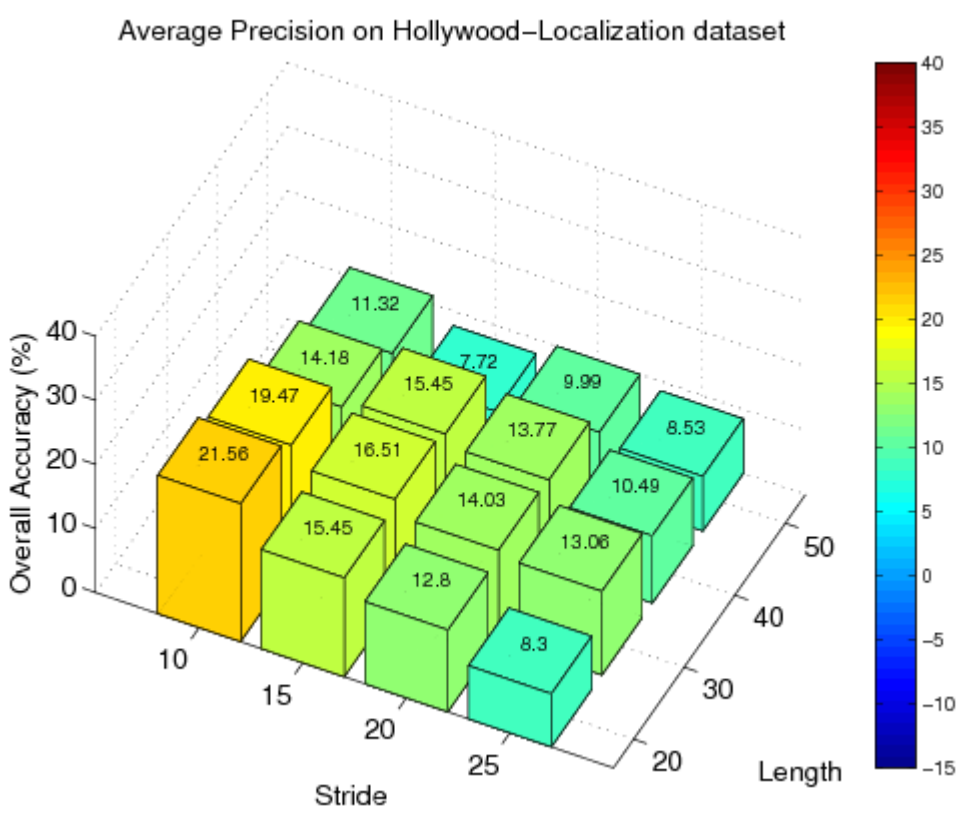
Parameter Evaluation

answer



Temporal divisions (n_t=3)

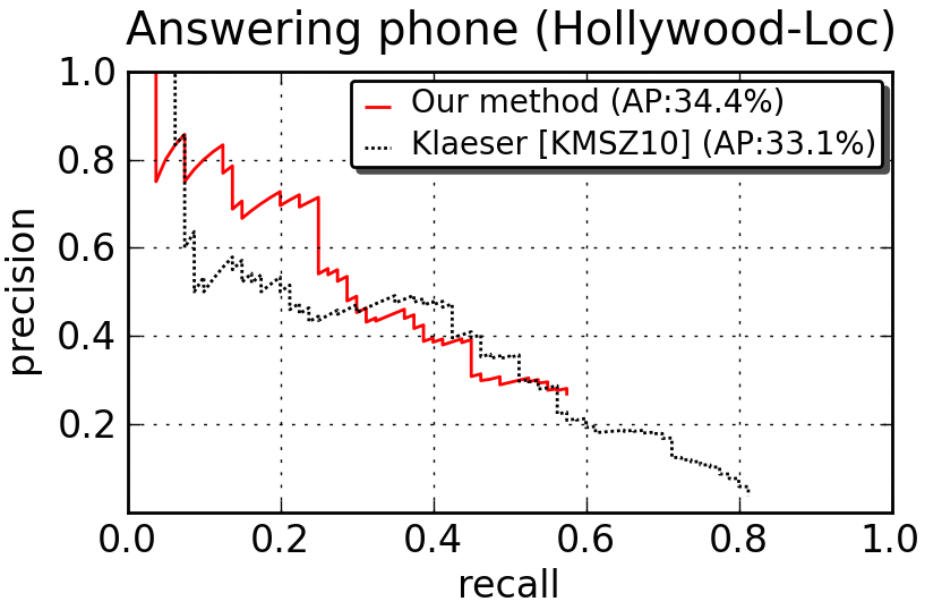
stand up

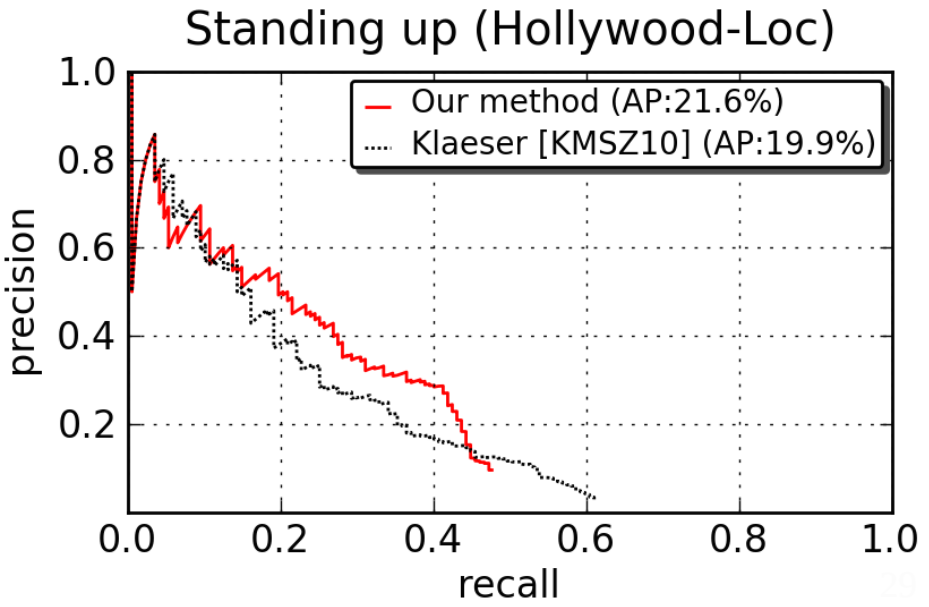


Temporal divisions (n_t=4)

Comparison to state-of-the-art:

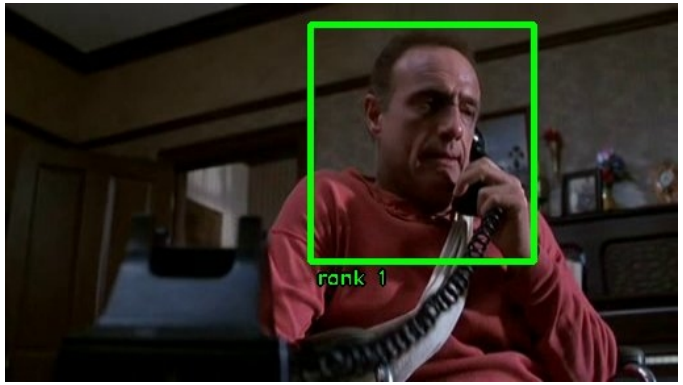
- Comparison to: [Kläser et al. 2010].
- Our approach compares favorably to the state-of-the-art.
- Actions in this dataset are rather short and well localized in time.
  - Suitable representation with fixed chunks



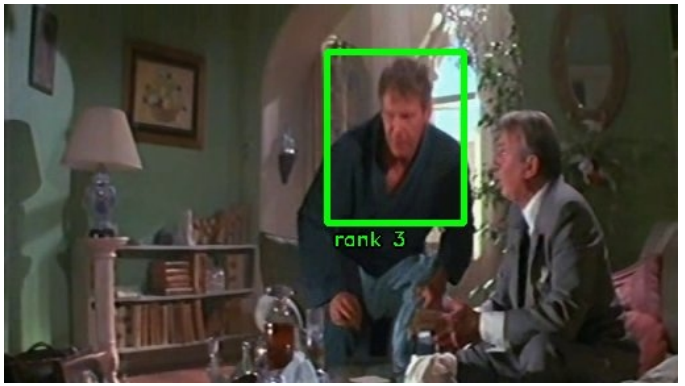


Action Localization Results

- **High ranked *answering phone* detections:**



- **High ranked *stand up* detections:**



Outline

- Context and Goal.
- Human Action Description.
- Experimental Results:
  - Action Classification.
  - Action Localization.
- **Conclusions.**

Conclusions

- We proposed and evaluated an approach to represent actions temporally as a loose collection of movement patterns.
- Division into four main parts:
  - Detection+tracking of humans.
  - Representation of tracks as a sequence of overlapping track segments (*chunks*).
  - Evaluation of chunks with an SVM classifier.
  - Classification and localization (start/end) of actions by voting/clustering approaches.

Conclusions

- Temporal divisions for descriptor improve performance

  (they encode more temporal information).
- Sufficiently long *chunks* + rather short *stride* (=large overlap) important.
- For *action classification* (Weizmann):
  - the optimal chunk length coincides with the cycle length of the repetitive action classes.
- For *action localization*:
  - Coffee and Cigarettes: actions of variable length/speed.
  - Hollywood: actions of fixed length, well localized in time.

Thank you very much for your attention!

I am glad to answer your questions.

Bibliography

- [BAV06] Moeslund Thomas B., Hilton Adrian, and Krüger Volker. *A survey of advances in vision-based human motion capture and analysis*. Computer Vision and Image Understanding, 104(2):90–126, 2006.
- [FW01] Bobick Aaron F. and Davis James W. *The recognition of human movement using temporal templates*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3):257-267,2001.
- [BGS+05] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. *Actions as spacetime shapes*. In Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 1395-1402.

Bibliography

- [LP07] Ivan Laptev and Patrick Perez. Retrieving actions in movies. In ICCV '07: Proceedings of the Eleventh IEEE International Conference on Computer Vision, pages 1-8, 2007.
- [SAS07] Paul Scovanner, Saad Ali, and Mubarak Shah. *A 3-dimensional sift descriptor and its application to action recognition*. In MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia, pp. 357-360, New York, NY, USA, 2007. ACM.
- [DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In CVPR'05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1, pages 886-893, 2005.

Bibliography

- [ESZ09] Mark Everingham, Josef Sivic, and Andrew Zisserman. *Taking the bite out of automatic naming of characters in TV video*. Image and Vision Computing, 27(5):545-559, 2009.
- [KMSZ10] Alexander Klaeser, Marcin Marszalek, Cordelia Schmid, and Andrew Zisserman. Human focused action localization in video. International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV, 2010.
- [CM02] Dorin Comaniciu and Peter Meer. *Mean shift: A robust approach toward feature space analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):603-619, 2002.