Bottom-up vision: initial findings with players action recognition

Teo de Campos, DPhil

CVSSP – Centre for Vision Speach and Signal Processing
Univerisity of Surrey

ACASVA kick-off meeting, 01 September 2009

Outline

1. Current Vision Sytem
2. Recognising Players Actions
   - Overview of the approaches
   - A Closer Look at [Ramanan et al., 2007]
3. Plan for the next 6 months
4. References

Outline

1. Current Vision Sytem
2. Recognising Players Actions
   - Overview of the approaches
   - A Closer Look at [Ramanan et al., 2007]
3. Plan for the next 6 months
4. References

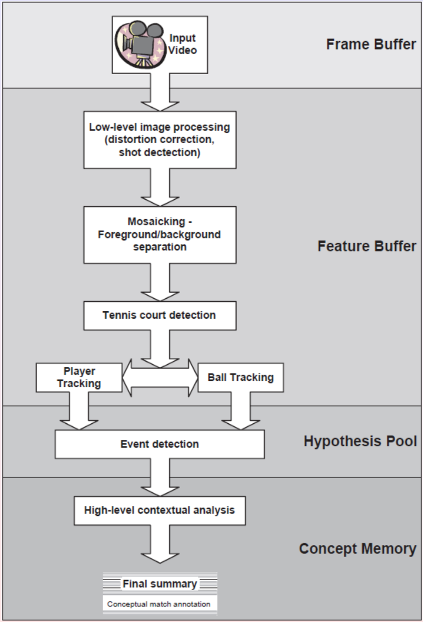
Outline

1. Current Vision Sytem
2. Recognising Players Actions
   - Overview of the approaches
   - A Closer Look at [Ramanan et al., 2007]
3. Plan for the next 6 months
4. References

Outline

1. Current Vision Sytem
2. Recognising Players Actions
   - Overview of the approaches
   - A Closer Look at [Ramanan et al., 2007]
3. Plan for the next 6 months
4. References

Current Vision Sytem from the VAMPIRE project


Input Video
Frame Buffer
Low-level image processing (distortion correction, shot dectection)
Mosaicking - Foreground/background separation
Feature Buffer
Tennis court detection
Player Tracking
Ball Tracking
Event detection
Hypothesis Pool
High-level contextual analysis
Concept Memory
Final summary
Conceptual match annotation

Issues with the VAMPIRE system

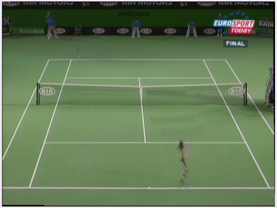
Heavily hard coded:

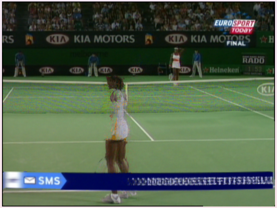
- Court detection
- Mosaic and homography computation
- Ball detection and tracking
- Players detection and action recognition
- Event/score recognition

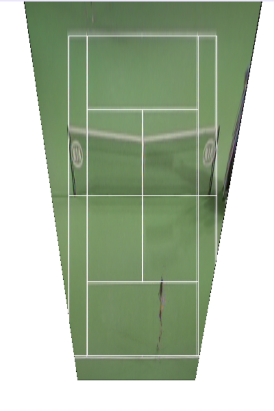
Issues with the VAMPIRE system

Heavily hard coded:

- Court detection
- Mosaic and homography computation
- Ball detection and tracking
- Players detection and action recognition
- Event/score recognition



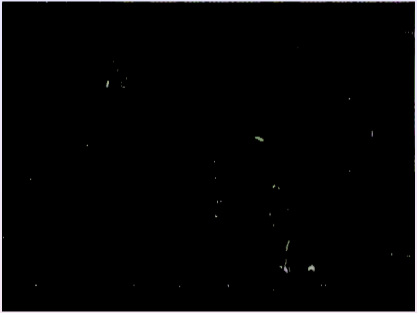


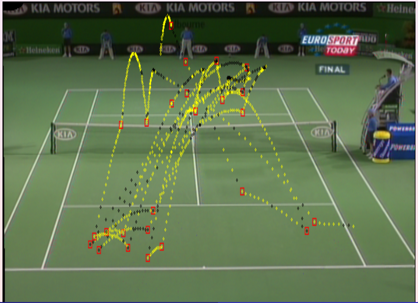


Issues with the VAMPIRE system

Heavily hard coded:

- Court detection
- Mosaic and homography computation
- Ball detection and tracking
- Players detection and action recognition
- Event/score recognition

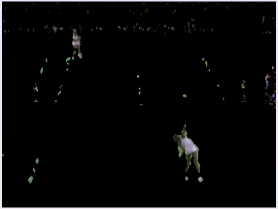


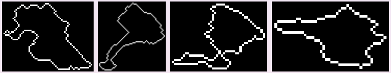


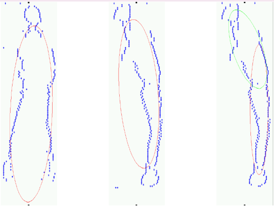
Issues with the VAMPIRE system

Heavily hard coded:

- Court detection
- Mosaic and homography computation
- Ball detection and tracking
- Players detection and action recognition
- Event/score recognition



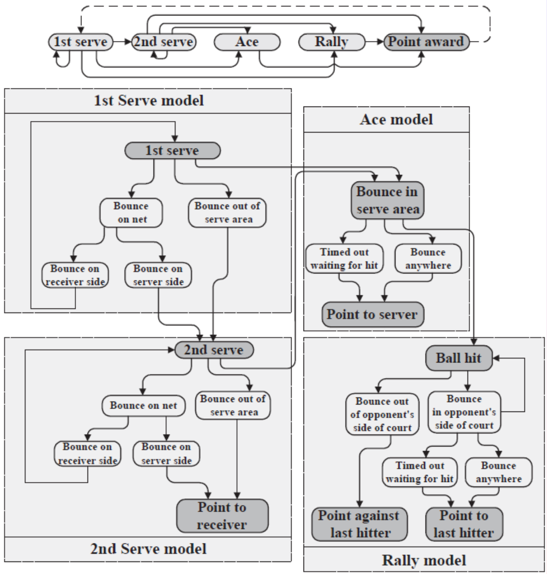




Issues with the VAMPIRE system

Heavily hard coded:

- Court detection
- Mosaic and homography computation
- Ball detection and tracking
- Players detection and action recognition
- Event/score recognition



Outline

1. Current Vision Sytem

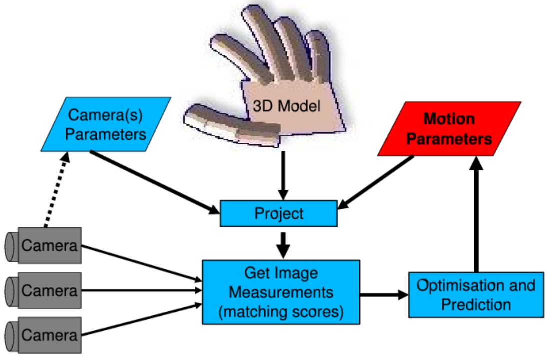
2. Recognising Players Actions
   - Overview of the approaches
   - A Closer Look at [Ramanan et al., 2007]

3. Plan for the next 6 months

4. References

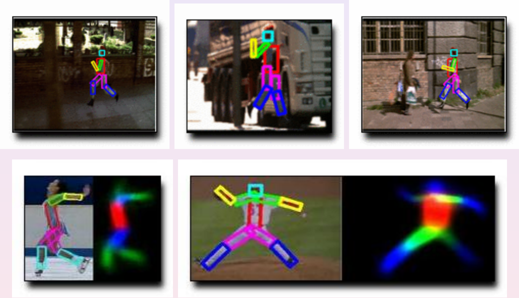
Recognising Players Actions

- Model-based (analysis as synthesis) approaches
- Discriminative approaches:
  - HOG [Dalal, 2006]
  - SIFT [Agarwal, 2006]
  - 3D SIFT in spatio-temporal blocks [Scovanner et al., 2007]
  - 3D corners [Gilbert et al., 2008]



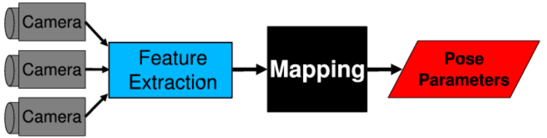
Recognising Players Actions

- Model-based (analysis as synthesis) approaches [Ikisler and Forsyth, 2007], [Ramanan et al., 2007]
- Discriminative approaches:
  - HOG [Dalal, 2006]
  - SIFT [Agarwal, 2006]
  - 3D SIFT in spatio-temporal blocks [Scovanner et al., 2007]
  - 3D corners [Gilbert et al., 2008]



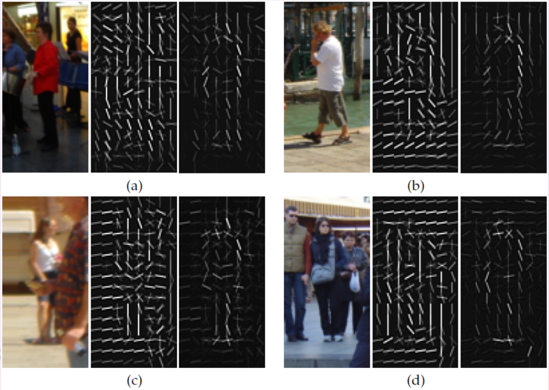
Recognising Players Actions

- Model-based (analysis as synthesis) approaches
- Discriminative approaches:
  - HOG [Dalal, 2006]
  - SIFT [Agarwal, 2006]
  - 3D SIFT in spatio-temporal blocks [Scovanner et al., 2007]
  - 3D corners [Gilbert et al., 2008]



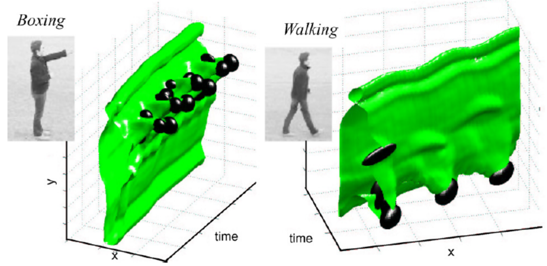
Recognising Players Actions

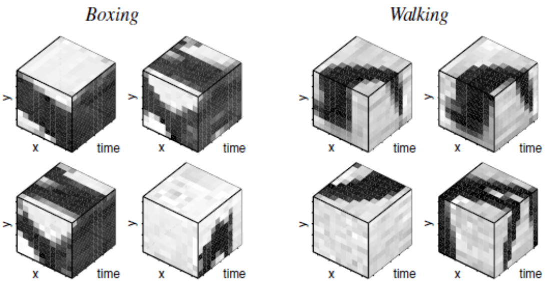
- Model-based (analysis as synthesis) approaches
- Discriminative approaches:
  - HOG [Dalal, 2006]
  - SIFT [Agarwal, 2006]
  - 3D SIFT in spatio-temporal blocks [Scovanner et al., 2007
  - 3D corners [Gilbert et al., 2008]



Recognising Players Actions

- Model-based (analysis as synthesis) approaches
- Discriminative approaches:
  - HOG [Dalal, 2006]
  - SIFT [Agarwal, 2006]
  - 3D SIFT in spatio-temporal blocks [Scovanner et al., 2007
  - 3D corners [Gilbert et al., 2008]





Outline

1. Current Vision Sytem

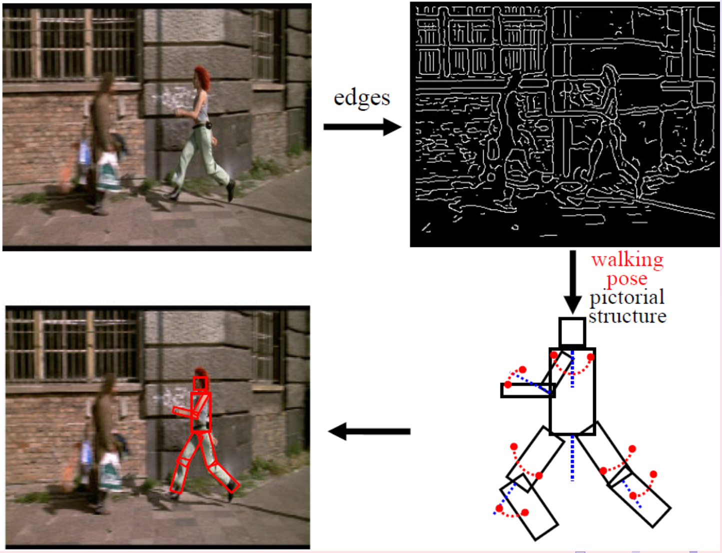
2. Recognising Players Actions
   - Overview of the approaches
   - A Closer Look at [Ramanan et al., 2007]

3. Plan for the next 6 months

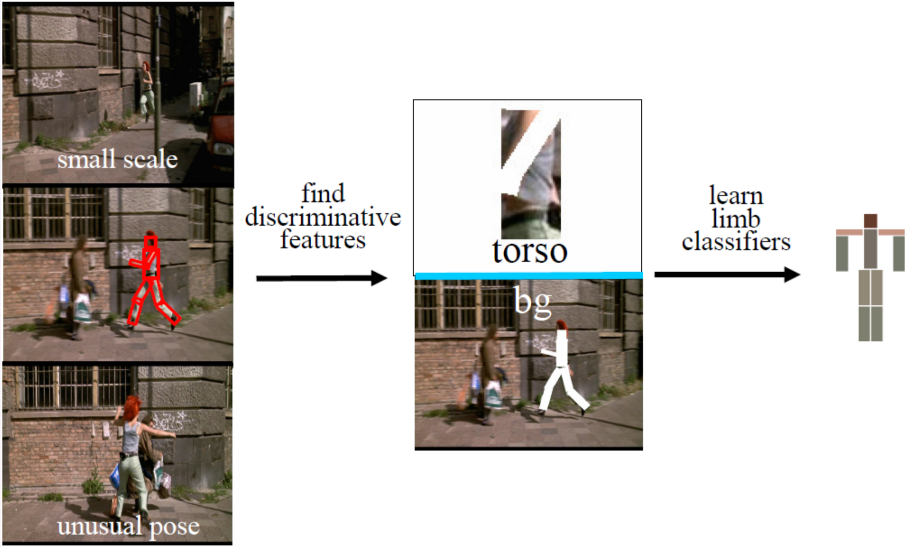
4. References

A Closer Look at [Ramanan et al., 2007]

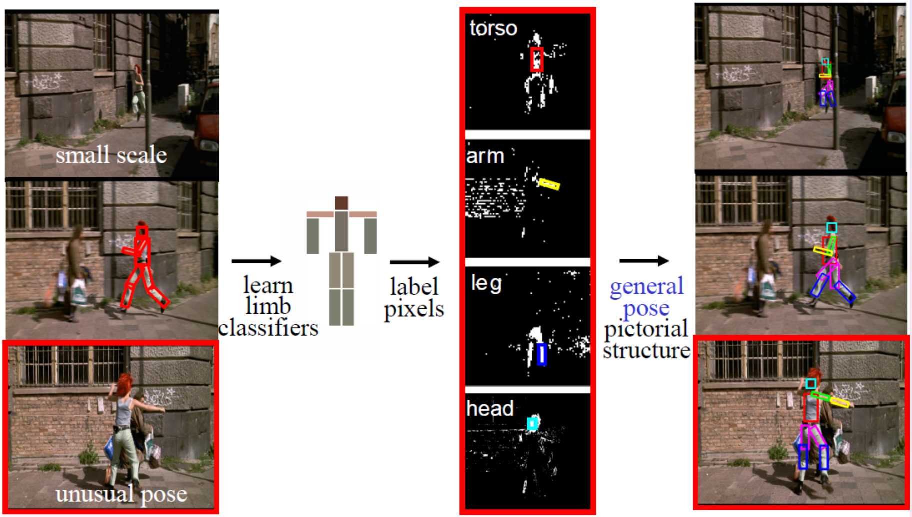
Stylised Pose Person Detector



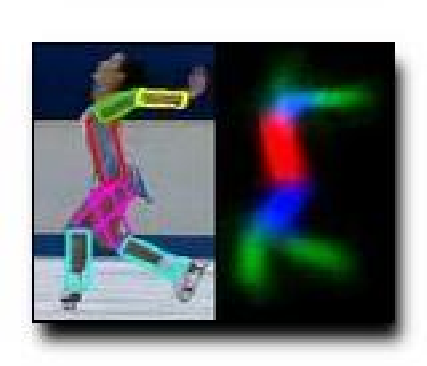
Building the Model [Ramanan et al., 2007]



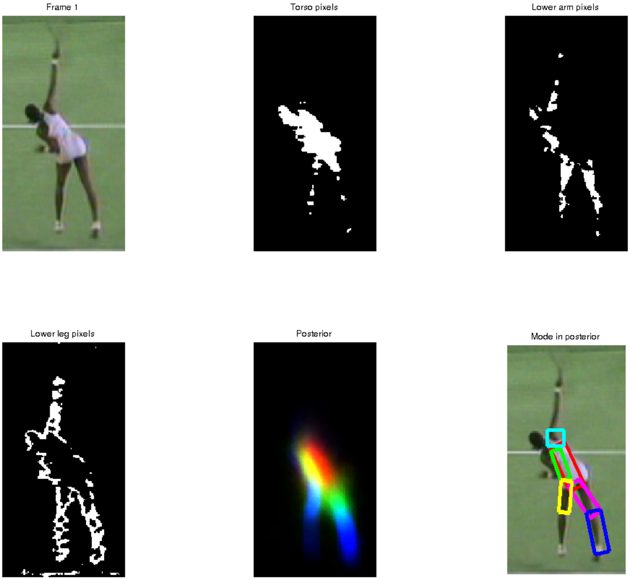
Detecting Known People [Ramanan et al., 2007]



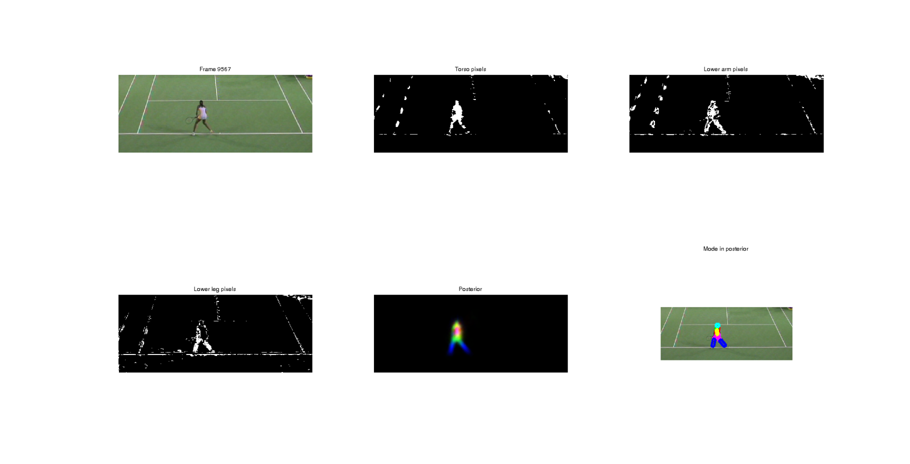
Result provided by the authors [Ramanan et al., 2007]



Results training with a serve frame



Results training with “walking person”



Initial Plan

1. First steps: human action recognition
    1. Try an action recognition method based on 3D SIFT (soon)
    2. Decide which method(s) to take forward
    3. Try to improve them and make them generalisable for different viewpoints, scales, etc (2-5 months)
2. Later: generalise other modules of the system:
    - Ball tracking
    - Court detection
3. In parallel: investigate uses of gaze data as part of the loop

References

Agarwal, A. (2006).
Machine Learning for Image Based Motion Capture.
PhD thesis, INRIA, GRAVIR, IMAG, Institut National Polytechnique de Grenoble, Grenoble, France.

Dalal, N. (2006).
Finding People in Images and Videos.
PhD thesis, INRIA, GRAVIR, IMAG, Institut National Polytechnique de Grenoble, Grenoble, France.

Forsyth, D. A., Arikan, O., Ikemoto, L., O'Brien, J., and Ramanan, D. (2006).
Computational studies of human motion: Part 1, tracking and motion synthesis.
Foundations and Trends in Computer Graphics and Vision, 1(2/3):77–254.

Gilbert, A., Illingworth, J., and Bowden, R. (2008).
Scale invariant action recognition using compound features mined from dense spatio-temporal corners.
In *European Conf. on Computer Vision, ECCV08*.

Gilbert, A., Illingworth, J., and Bowden, R. (2009).
Fast realistic multi-action recognition using mined dense spatio-temporal features.
In *Proc 12th Int Conf on Computer Vision, Kyoto, Japan, Sept 27 - Oct 4*.

Ikisler, N. and Forsyth, D. (2007).
Searching video for complex activities with finite state models.
In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*.

Laptev, I., Caputo, B., Schüldt, C., and Lindeberg, T. (2007).
Local velocity-adapted motion events for spatio-temporal recognition.
Computer Vision and Image Understanding, 108:207–229.

References (2)

Micilotta, A., Ong, E., and Bowden, R. (2006).
Real-time upper body detection and 3D pose estimation in monoscopic images.
In Leonardis, A., Bischof, H., and Pinz, A., editors, *Proc. European Conference on Computer Vision*, volume III of *LNCS*, pages 139 – 150. Springer Verlag 2006.

Moeslund, T., Hilton, A., and Kruger, V. (2006).
A survey of advances in vision-based human motion capture and analysis.
Computer Vision and Image Understanding, 104(2-3):90–127.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008).
Unsupervised learning of human action categories using spatial-temporal words.
Int Journal of Computer Vision.

Ramanan, D., Forsyth, D. A., and Zisserman, A. (2005).
Strike a pose: Tracking people by finding stylized poses.
In *Proc IEEE Conf on Computer Vision and Pattern Recognition, San Diego CA, June 20-25*.

Ramanan, D., Forsyth, D. A., and Zisserman, A. (2007).
Tracking people by learning their appearance.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(1):65–81.

Scovanner, P., Ali, S., and Shah, M. (2007).
A 3-dimensional sift descriptor and its application to action recognition.
In *Proc of the ACM Multimedia*, Augsburg, Germany.