Players action detection

Teo de Campos

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

3^{rd} ACASVA meeting, 24 June 2010

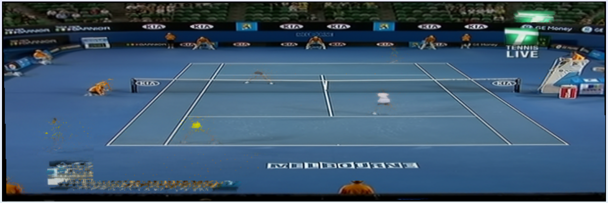
Outline

1. Player Detection
2. Action Recognition
3. Ongoing work
4. References

Outline

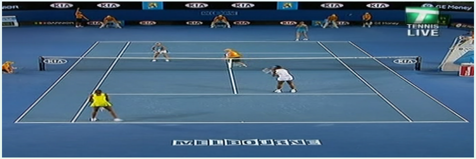
1. Player Detection
2. Action Recognition
3. Ongoing work
4. References

Detecting player foreground



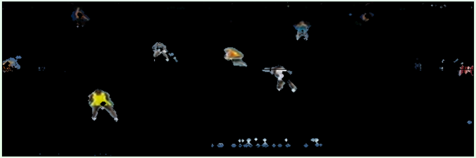
Mosaic, built per shot

-



Input image: de-interlaced field with radial distortion corrected, registered with the mosaic

=



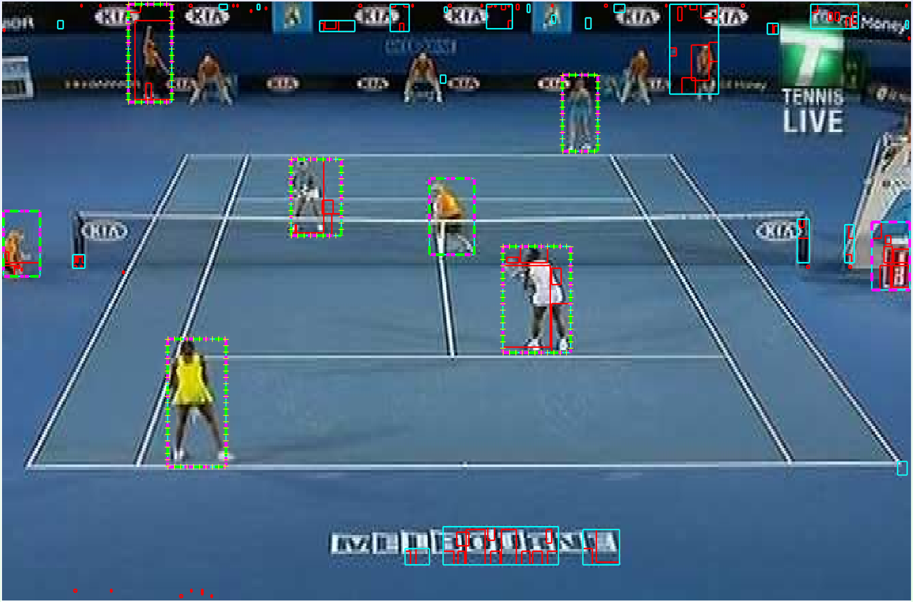
Moving blobs, filtered with a morphological opening operation (erosion → dilation)

Processing foreground blobs for player detection

Algorithm

1. Background subtraction
2. Morphological opening
3. Fit bounding boxes to all continuous blobs
4. Merge nearby boxes
5. Apply geometric constraints: area, aspect ratio, ratio area/BB_area
6. Apply temporal constraint
7. Apply foreground mask

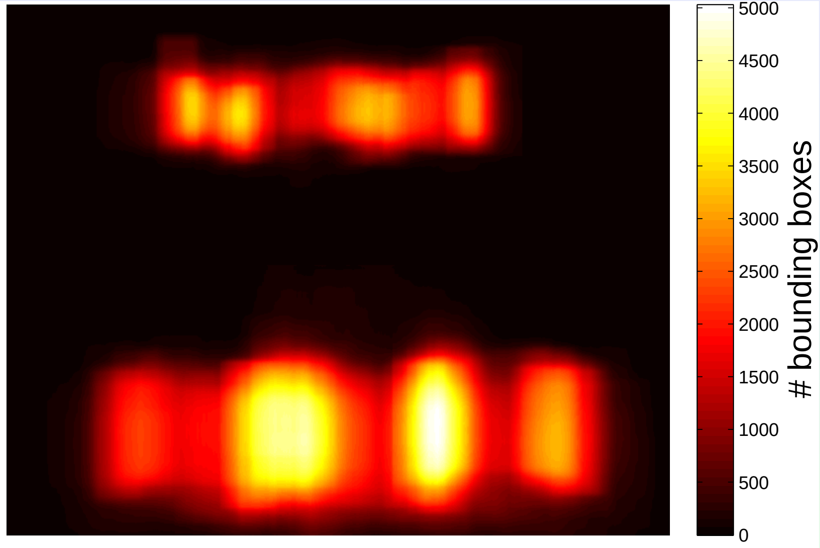
Processing foreground blobs for player detection



Resulting boxes for the previous image

- Initial background subtraction and pre-processing: 119 red boxes
- Merging: 32 cyan boxes
- Geometric constraints: 8 dashed magenta boxes
- Spatio-temporal consistence: 7 dashed green boxes
- Mask filter: 5 dotted yellow boxes

Player location pdf



Computed from a 35 minutes footage of singles.

Foreground mask

Statistics of the bounding boxes

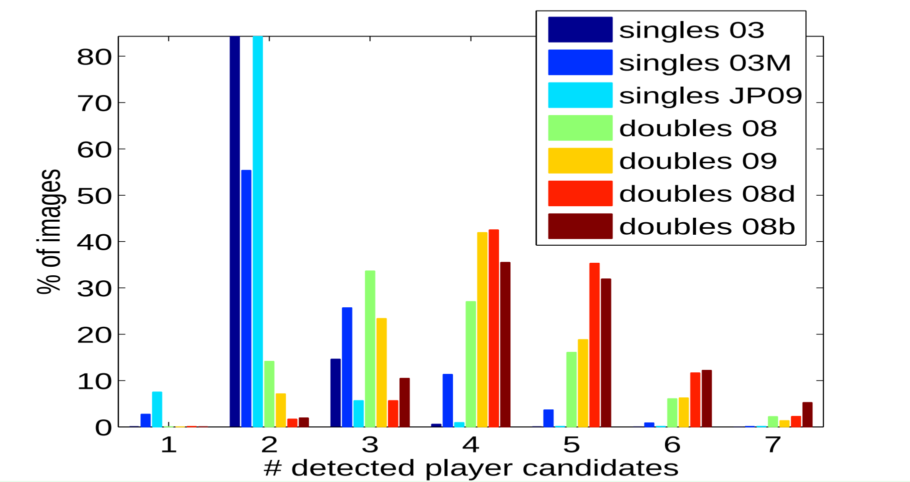
Number of player candidates per frame after some stages of the processing pipeline:

Footage	BS blobs	motion	mask
singles 03	177.4 ± 23.8	2.8 ± 1.5	2.2 ± 1.3
doubles 08	64.9 ± 21.9	4.7 ± 1.3	3.8 ± 1.2
doubles 09	50.4 ± 44.7	3.5 ± 2.2	3.0 ± 1.7

- *BS blobs*: initial blob detection from background subtraction;
- *motion*: application of a motion smoothness constraint;
- *mask*: application of the likely player location mask.

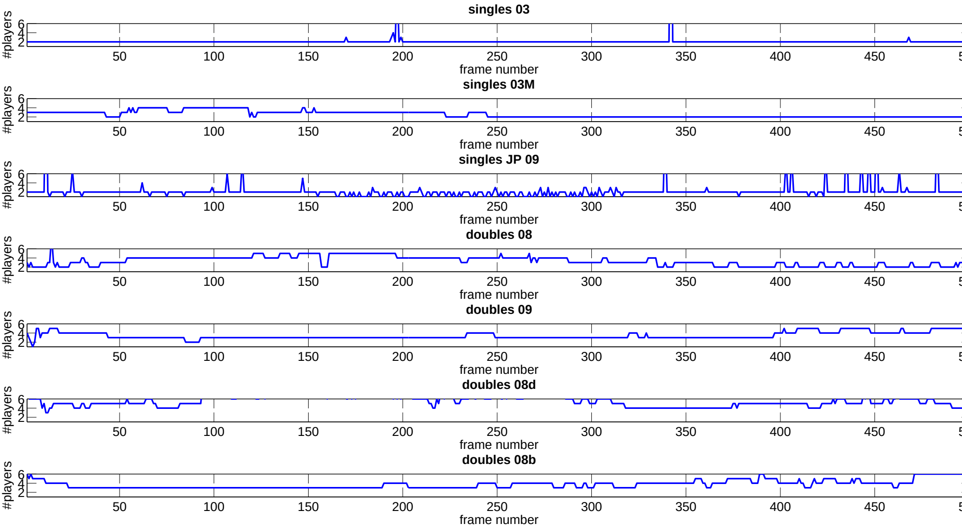
Statistics of the bounding boxes

Detected player candidates in each frame of play shots:



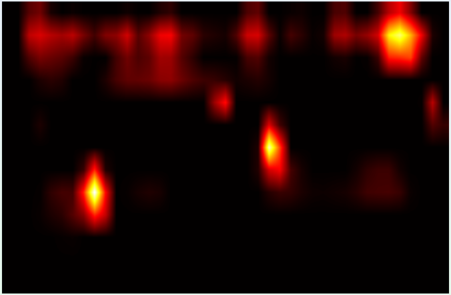
# play frames	sngl03	dobl08	dobl09
	25984	17737	33223

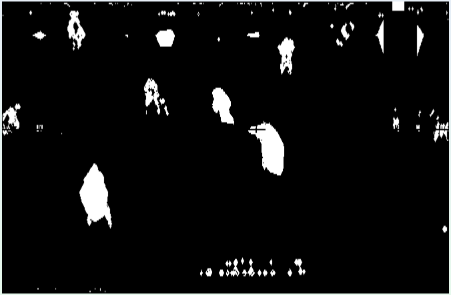
Player detections over time



Combining with visual saliency

Combining background subtraction with visual saliency maps from (Walther and Koch, 2006) by thresholding both and using an OR operation.



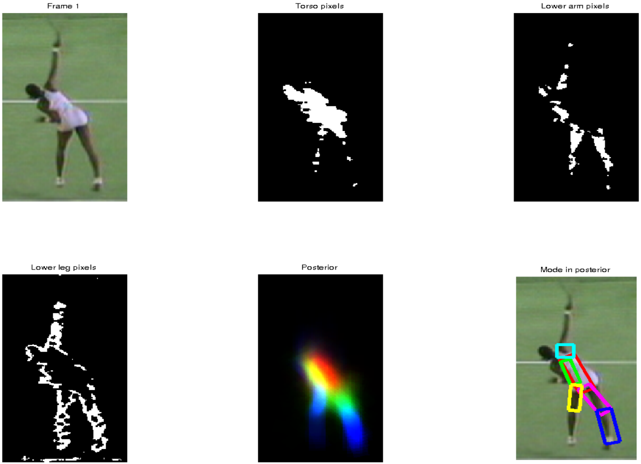


Bad idea!
Too many false positives.

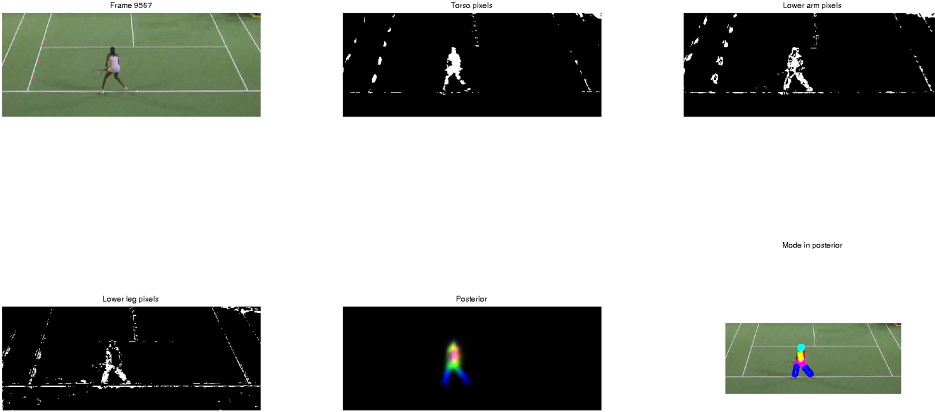
Ongoing work on player detection

Using parts-based person detector to locate players (Ramanan et al., 2007)

Results training with a *serve* frame:



(Ramanan et al., 2007)'s results training with *walking*



Frame 9567

Torso pixels

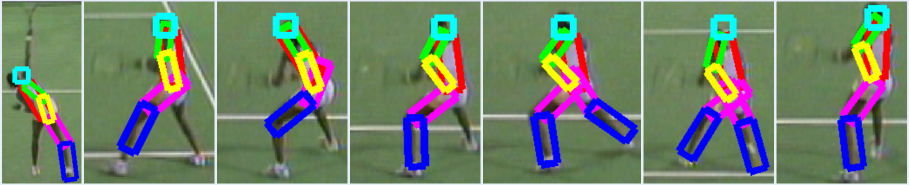
Lower arm pixels

Lower leg pixels

Posterior

Mode in posterior

More results with (Ramanan et al., 2007)

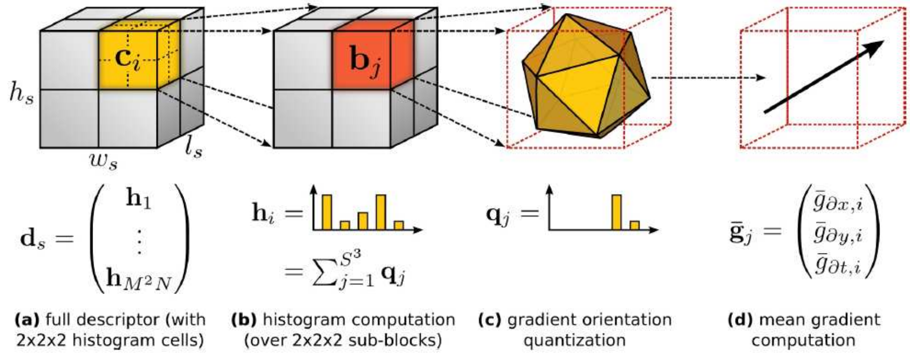


- ☺ Good to detect *near* players
- ☹ Bad to locate arms
- ☹ Joint localisation is not accurate
- 😐 For training, a search through the scale is required
- 😐 Training has to be done for each game and each player individually

Outline

1. Player Detection
2. Action Recognition
3. Ongoing work
4. References

3DHOG spatio-temporal descriptor (Kläser et al., 2008)



$\mathbf{c}_i$
$\mathbf{b}_j$
h_s
w_s
l_s
$\mathbf{d}_s = \begin{pmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{M^2 N} \end{pmatrix}$
$\mathbf{h}_i =$
$= \sum_{j=1}^{S^3} \mathbf{q}_j$
$\mathbf{q}_j =$
$\bar{\mathbf{g}}_j = \begin{pmatrix} \bar{g}_{\partial x,i} \\ \bar{g}_{\partial y,i} \\ \bar{g}_{\partial t,i} \end{pmatrix}$

(a) full descriptor (with 2x2x2 histogram cells)

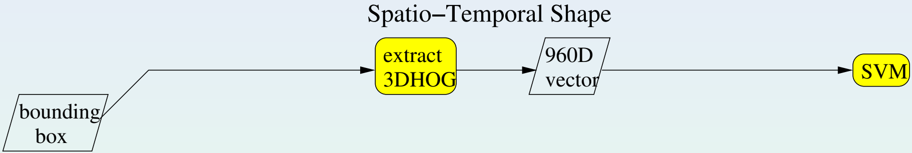
(b) histogram computation (over 2x2x2 sub-blocks)

(c) gradient orientation quantization

(d) mean gradient computation

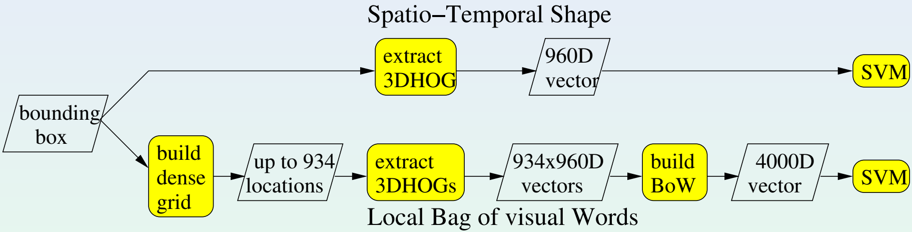
- Gives a $20f \times 4x \times 4y \times 3t = 960$D vector
- Proven to be among the state-of-the-art descriptors in (Wang et al., 2009)

Action classification methods – STS

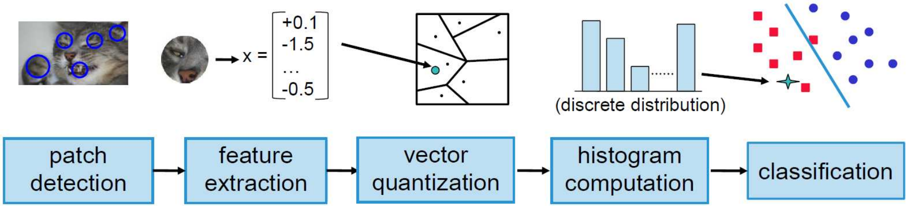


σ of the RBF kernel was set to the average distance between every pair of samples in the training set.

Action classification methods – STS and LBoW



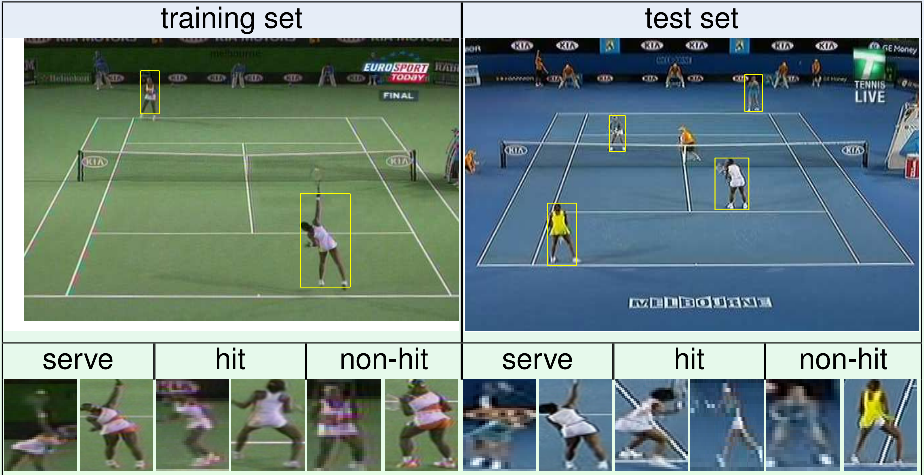
Bags of visual Words (Csurka et al., 2004)



- Up to 934 feature vectors per player, per event (we use a dense grid of $5s \times 9y \times 9x \times 9t$ locations but sampling is denser near the centre of the bounding box)
- 4000 visual words
- We also evaluated spatio-temporal pyramid kernels (Choi et al., 2008)

Primitive Actions Dataset

Footage	length	play shots	serve	hit	non-hit
singles 03	35min	80	76	219	943
doubles 09	30min	34	46	167	1351



training set

test set

serve

hit

non-hit

serve

hit

non-hit

Results

- LBoW – mean AUC (%) with different spatio-temporal pyramid kernels:

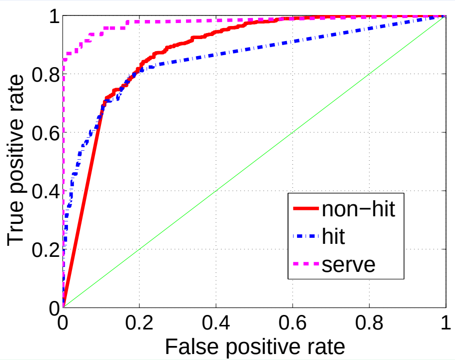
	spatial split				
temporal split	1x1	1x3	2x2	3x1	MK
x1	78.5	78.2	79.6	79.5	80.6
x3	84.4	82.3	82.8	84.4	**84.5**

- The STS single feature method resulted in mean AUC of **90.3**%.
- STS confusion matrix for thresholds selected so that the true positive rate is 77.62% and the false positive rate is 22.38%:

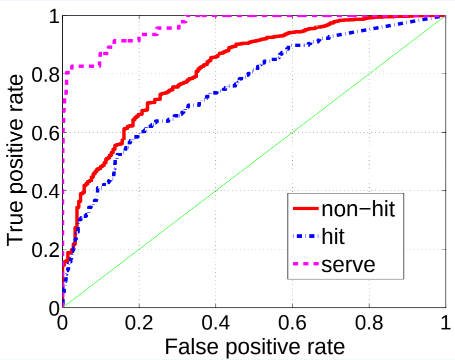
	non-hit	hit	serve
non-hit	1068	182	117
hit	36	119	14
serve	2	3	41

ROC curves

single feature STS



LBoW MKx3



Outline

1. Player Detection

2. Action Recognition

3. Ongoing work

4. References

Ongoing work

1. Improve player detection and tracking methods
2. Compare STS and BoW-based methods using well known datasets (de Campos et al., 2010)
3. Apply *n-gram-like* heuristics to filter action classification results
4. Separate near player from far player
5. Do experiments in larger datasets
6. Evaluate the *bags of locally weighted features* (de Campos et al., 2010)

Outline

1. Player Detection
2. Action Recognition
3. Ongoing work
4. References

References

Choi, J., Jeon, W. J., and Lee, S.-C. 2008.
Spatio-temporal pyramid matching for sports videos.
In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. 2004.
Visual categorization with bags of keypoints.
In *ECCV International Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic.

de Campos, T., Csurka, G., and Perronnin, F. 2010.
Images as sets of locally weighted features.
Computer Vision and Image Understanding.
under review.

de Campos et al., T. 2010.
Bags-of-words and spatio-temporal shapes for action recognition.
In *Proc Asian Conf on Computer Vision*, Queenstown, New Zealand. Springer.
submitted.

Kläser, A., Marszałek, M., and Schmid, C. 2008.
A spatio-temporal descriptor based on 3d-gradients.
In *Proc 19th British Machine Vision Conf, Leeds*, pages 995–1004. BMVA.

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.

Walther, D. and Koch, C. 2006.
Modeling attention to salient proto-objects.
Neural Networks, 19:1395–1407.

Wang, H., Ullah, M. M., Käser, A., Laptev, I., and Schmid, C. 2009.
Evaluation of local spatio-temporal features for action recognition.
In *Proc 20th British Machine Vision Conf, London, Sept 7-10*. BMVA.