# Real Time Human Pose Recognition in Parts from Single Depth Images [Shotton et al, CVPR 2011]

This paper describes human pose recognition algorithm with:

* auto-initialized tracking
* failure recovery
* handles variations in human poses, shapes and size
* limited compute budget ( real time games on Xbox 360)

Major steps in the pipeline include:

* **Capture depth image and remove background**
  + background subtraction is simple due to depth information from the infrared sensor
* **Infer body parts per pixel**
  + Learn discriminative classifier from training data
  + Synthetic training data is created from 500K mation capture frames containing 100K poses.
  + These are retargeted to 15 models and rendered using graphics pipeline
  + Invariance to shape, size, pose is built
  + ‘Fast’ depth image features are computed
  + Random forest classifier
* **Cluster pixels to hypothesize body joint positions**
  + Joint locations are hypothesized using density function and mean shift clustering is used for mode detection to obtain joint locations
* **Fit model and track skeleton**
  + Proposals for skeletons are made more robust by 3D join hypotheses, kinematic constraints and temporal coherence constraints

## Highlights of this method : speed and robustness

# Hollywood 3D: Recognizing Actions in 3D Natural Scenes [Hadfield, Bowden]

This paper extends action recognition in video to 3D video. A new dataset Hollywood3D is made available for 3D video action recognition.

Extensions considered include:

1. Interest points:
   1. Harris corners (Ha)
   2. Hessian points (He)
   3. Separable filters (S)
2. Feature descriptors
   1. Bag of visual words : HOG, HOF (HoDG)
   2. Relative motion Descriptors (RMD)

Important point to note is that combination of appearance and depth streams (I and D respectively) constitutes 3.5D rather than volumetric data – the measurements are not dense along the new dimension. Gradient calculations can not be performed directly on the z axis. The relation between the gradients is captured by the chain rule:

Iz = Ix / Dx + Iy / Dy + It / Dt

Hence the choice is between 4D representation or 3.5 representation using a pair of complimentary 3D spatio-temporal volumes for appearance and depth respectively.

## Results:

Average precision and correct classification rate are reported for the combination:

{RMD, RMD-4D, HoG/HoF, HoG/HoF/HoDG} x {3D-S, 3.5D-S, 4D-S,3D-Ha, 3.5D-Ha, 4D-Ha, 3D-He, 3.5D-He, 4D-He}

AP values are in the order of 10-15% percent.

## Comments

Recent approaches in 2D video action recognition like trajectories are not exploited on this dataset and there seems to be a scope of refinement.