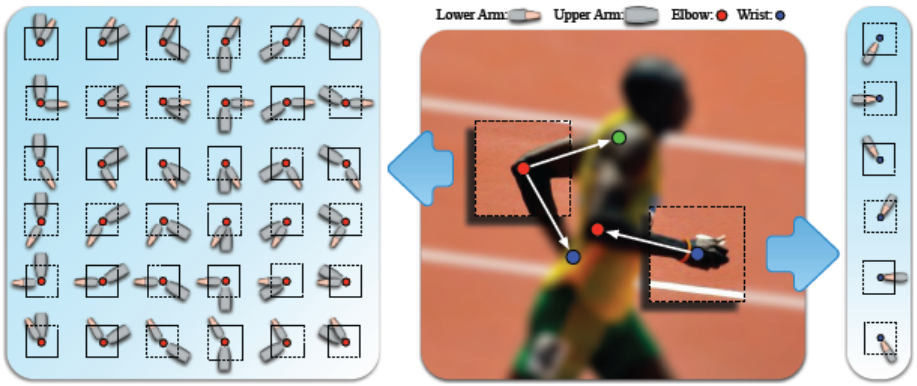
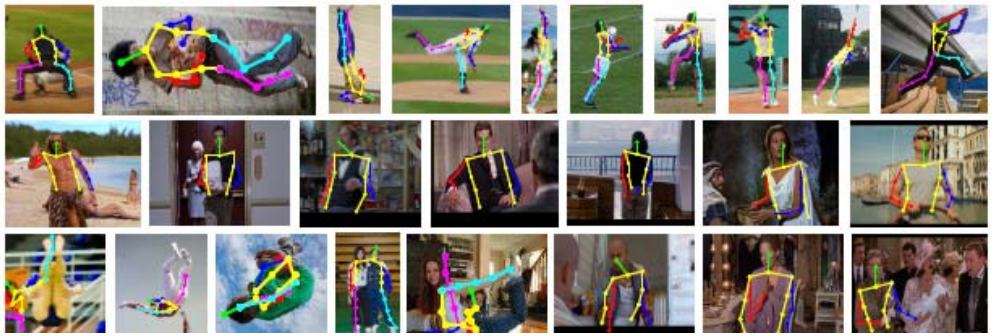
Human Parsing: 3D and Sparsity

Alan Yuille
Bloomberg Distinguished Professor
Cognitive Science and Computer Science
Johns Hopkins University

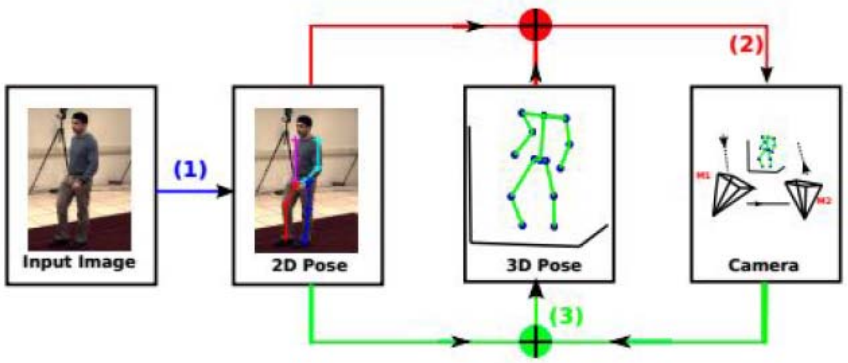
DeepNets for Joint-Types.
Pictorial Structures for Spatial Relations.

- Train DeepNets to detect/hypothesize positions and types of joints.
- Use graphical models to reason about spatial relations of joint types.
- Note: Joint-type predicts direction, distance, and type of neighboring joints. (Xianjie Chen and A. Yuille 2014).
- Psychophysics: Kersten et al. (2016).





From 2D to 3D: Joints



- Estimate 3D from 2D.
- Detect joints in 2D.
- Requires a prior on 3D shape of joints.

- Learn prior from dataset of 3D poses.
- PCA? – too many bases – not a linear space.
- Sparsity – not tight – impossible poses can have low encoding costs.
- Sparsity + limb length ratios –better (Chunyu Wang et al. 2014).
- But not generative and lacks geometric intuition.

Encoding 3D poses: Action recognition

- Learn a representation of 3D pose. (Chunyu Wang et al. 2016)
- The representation is generative. We can sample from it and obtain realistic poses.
- It is a variant of sparsity which involves modeling the manifold of poses by a set of simplices.

Geometric Interpretation of Sparse Coding.

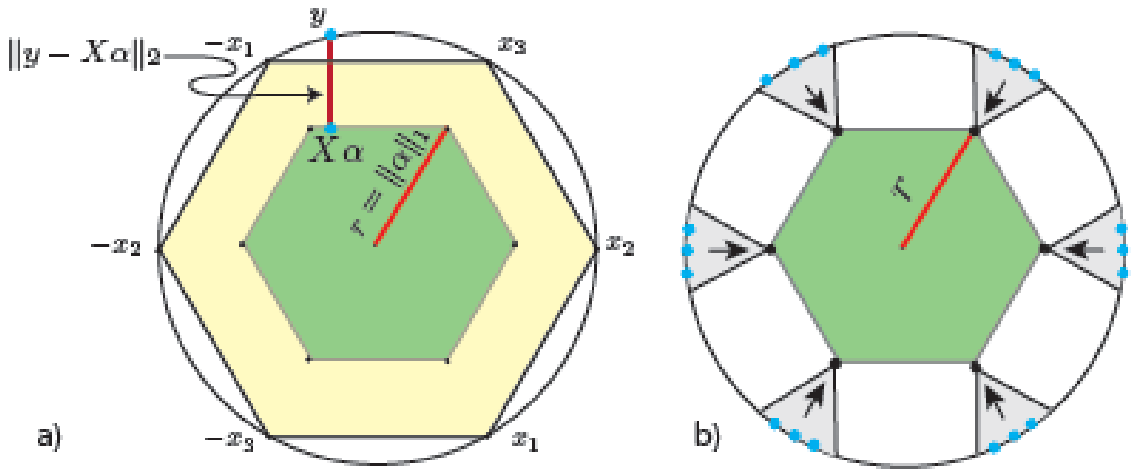
- Problem: Surely the 3D poses of joints lie on a manifold (e.g., L. Agapito et al). What do manifolds have to do with sparsity?

- Sparse coding: $$||y - \sum_{i=1}^{n} \alpha_i x_i||^2 + \lambda \sum_{i=1}^{n} |\alpha_i|.$$
- Data y, Bases x, Coefficients alpha, penalty lambda.
- Quadratic Penalty plus L1 regularizer.
- Learning a sparse dictionary from data {y}:

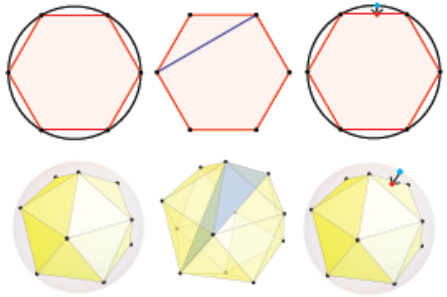
$$\min_{X,\alpha} \sum_{\mu=1}^{N} ||y_\mu - X\alpha^\mu||^2 + \lambda||\alpha^\mu||_1,$$

Geometric Interpretation

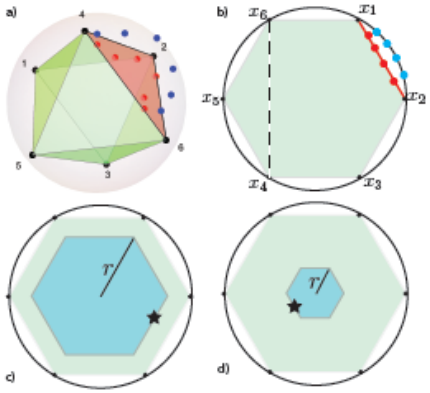
- Fix the sum of alphas. Then set of points X alpha is a convex hull of the bases X.



Learning Sparse Dictionary



- Normalize the data to lie on the unit sphere.
- Fix the radius of convex hull.
- Then learning a sparse dictionary corresponds to projecting data *onto the boundary of the convex hull*.
  Hence onto a subset of the boundary simplices.
- Exploit this fact. Standard sparsity encodes all the data *within* the convex hull.
- Instead only encode data by boundary simplices. Constraints on the co-activation of bases.



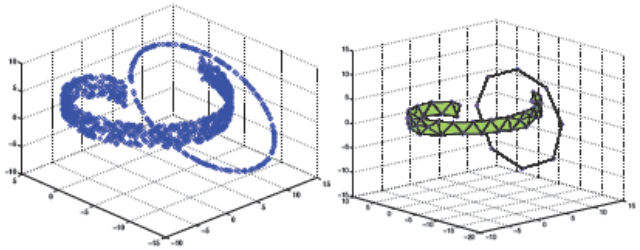
Activated Simplices: Algorithms

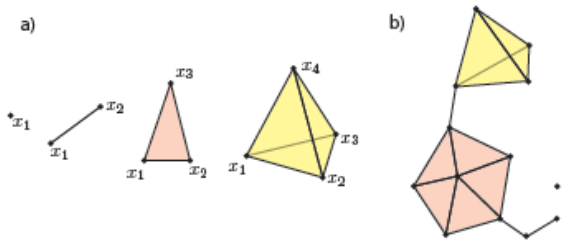
- Identify a subset of boundary simplices which encode the data. Bias towards boundary simplices which have low-dimensions and which contain many datapoints,

- Algorithm I: Learn the convex hull (bases) that minimize the projection error. Variant of spare coding algorithms.
- Algorithm II: Identify boundary simplices which have low-dimension and which encode large numbers of datapoints. Clustering algorithm.
- Output: a set of Activated Simplices.
- Note: both algorithms address non-convex problems so only local minima can be guaranteed.

Represent Manifold by activated simplices.

- The manifold of poses (projected onto sphere) is represented by a set of simplices.
- Note that activated simplices allows the manifold to be disconnected and to have variable dimension.

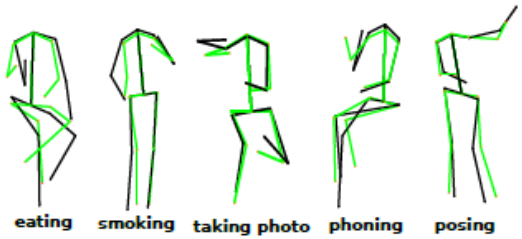




- Note: computer graphics represents 3D geometry by simplices.
- Note: prior work on representing manifolds by simplices.

Generative Model: Tight Representation.

- To sample from Activated Simplices.
- Learn a Dirichlet distribution for the datapoints on each activated simplex.
- (I) Select an activated simplex at random.
- (II) Sample a point within the activated simplex from Dirichlet distribution. Project onto sphere.



- The samples are realistic poses.
- The representation is tight.
- Can validate by reconstruction from noisy/contaminated data.

Synthesize 3D poses

- Synthesize humans.
- Dirichlet distributions

for positions on simplex.

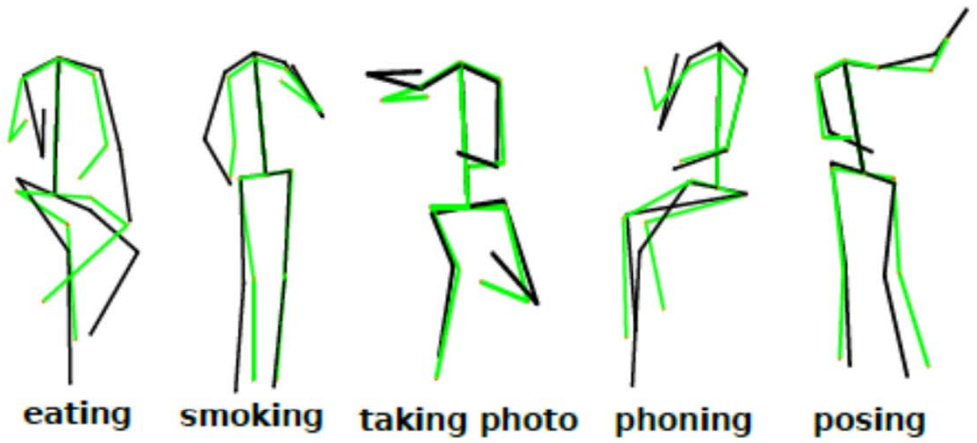
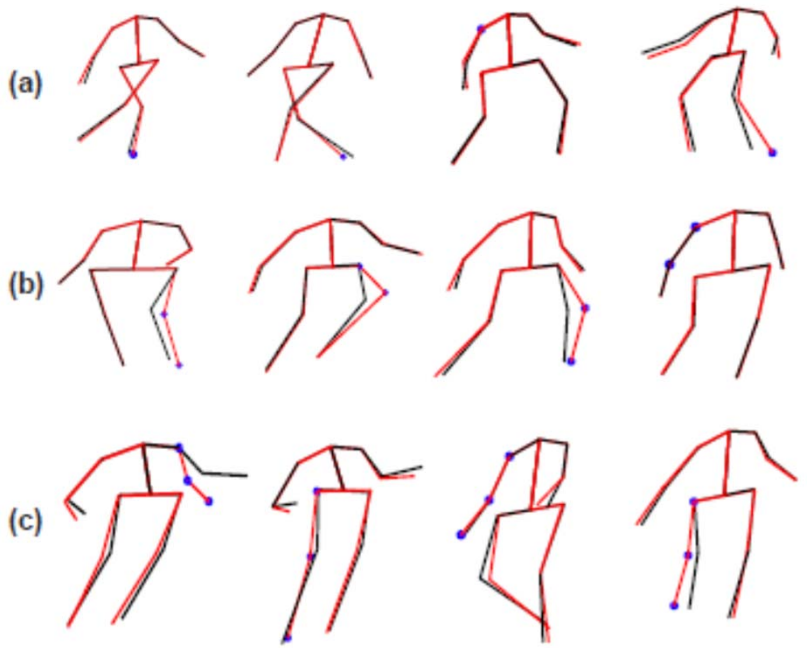


Figure 13: Synthesized human poses (in green) are overlaid with their nearest neighbours (in black) in the training dataset. The actions of the nearest neighbours are shown in the texts below the poses. We can see that the synthesized poses differ from the training data but they are still realistic which shows the simplices' generalization properties.

Reconstruct occluded poses

- Blue dots are missing points.
- Black are the reconstruction



Action Recognition

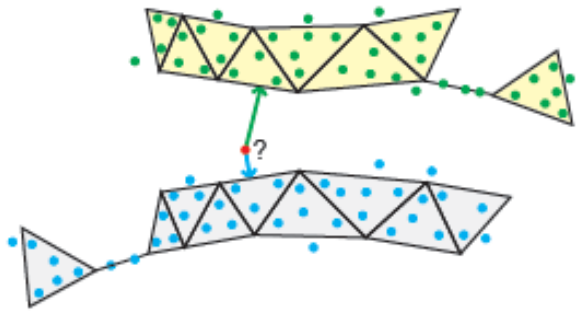
- Use activated simplices for classification of actions.

Table 2: Action recognition accuracy on MSR-Action3D.

Methods	Accuracy (%)
Sparse Coding	80.22
Action Graph on Bag-of-3D joints [13]	74.70
Actionlet Ensemble [24]	88.20
Spatial-Temporal-Part model [23]	90.22
Our Approach	**91.30**

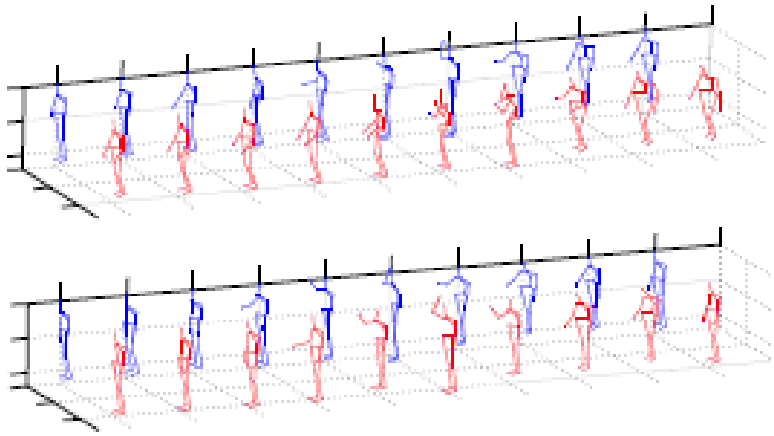
Advantages of Activated Simplices.

- Tight representation – anything that can be represented cheaply by activated simplices is a legitimate human pose. (Learning based).
- The representation can be use for multiple tasks. Only needs a simple classifier for each task.
- For example, action classification – learn a Manifold of Activated Simplices for each action. Classify an input sequence by finding the manifold closest to the data.

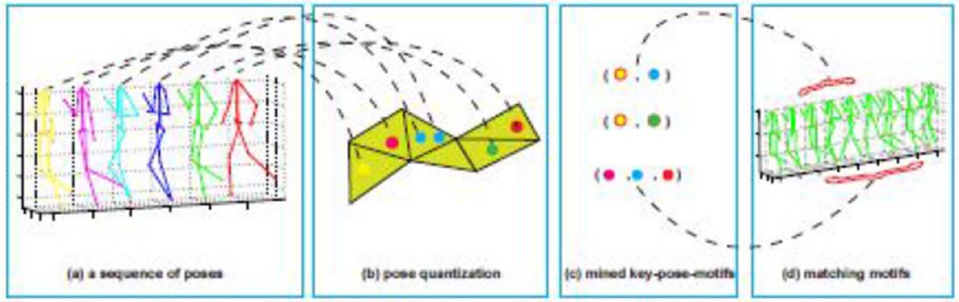


Pose-Snippets

- Pose-snippets are sequences of poses. E.g., ten consecutive poses.
- Can apply activated simplices to pose sequence.
- Single poses can occur in many different actions. But pose-snippets are more discriminative between sequences. (Also sample from them).



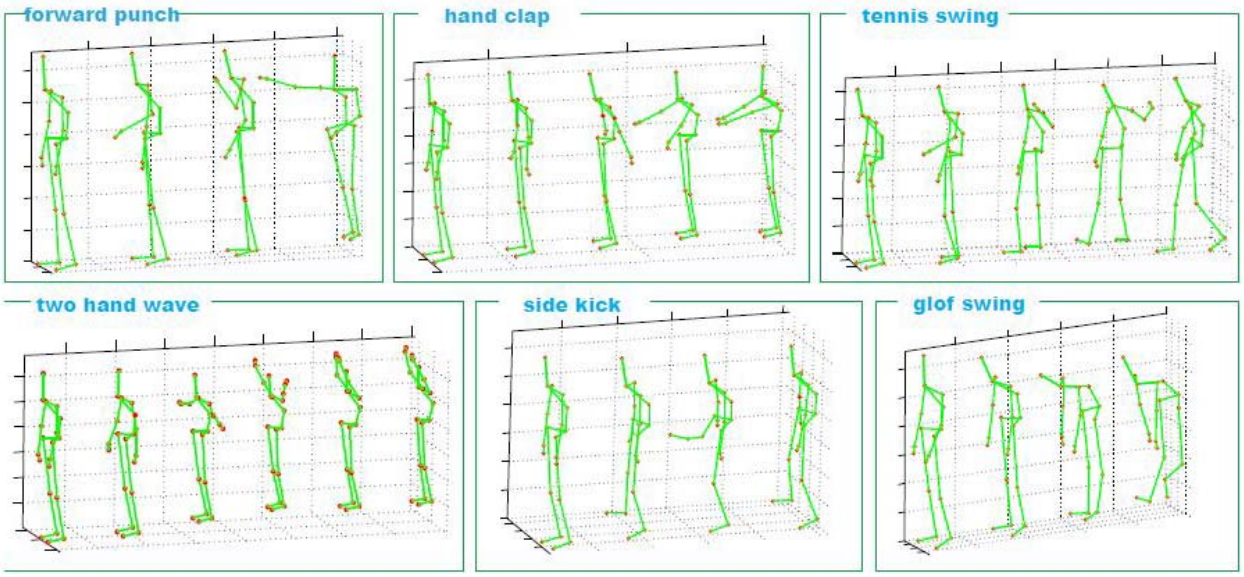
Key Point Motifs



- Actors perform actions in different styles, speeds, and with outlier poses.
- Limited amounts of training data.
- Key-point motifs. A set of ordered poses which are close, but not necessarily adjacent, in an action sequence.
- A key-point motif is adaptive to the input sequence and is robust to variations in speed and style. Also insensitive to outliers.
- Chunyu Wang et al. 2016B

Mining Key-Point motifs

- Compose elementary sequences to form larger sequences.
- Algorithm: use dynamic programming, exploit the linear structure.

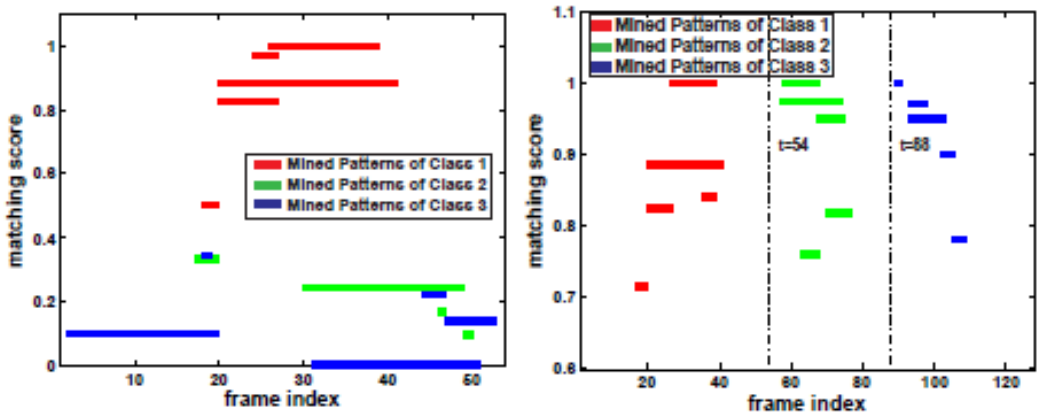


Key Point Motifs

- Strategy: find small key-point-motifs, compose them recursively to find bigger ones.

Algorithm 1 Key-pose-motif Mining Algorithm

T^1={1-motifs}
for $(k = 2; T^{k-1} \neq \emptyset; k{+}{+})$ **do**
  $T^k = \text{expand}(T^{k-1})$
  for $(i = 1; i \leq |T^k|; i{+}{+})$ **do**
    support=0
    for $(j = 1; j \leq |D|; j{+}{+})$ **do**
      support=support+$\eta(t_i^k, D_j)$
      If $\frac{\text{support}}{|D|} \leq \epsilon$
        $T^k \leftarrow T^k - \{t_i^k\}$
      endif
    end for
  end for
end for



Key Point Motifs

- Intuitively learn “fuzzy templates” for sequences of human poses.
- Fuzzy because we allow variability in time distance between adjacent frames in the motif. Robust.
- Adapts to new data. Needs little training data.
- Good results on benchmarked datasets.