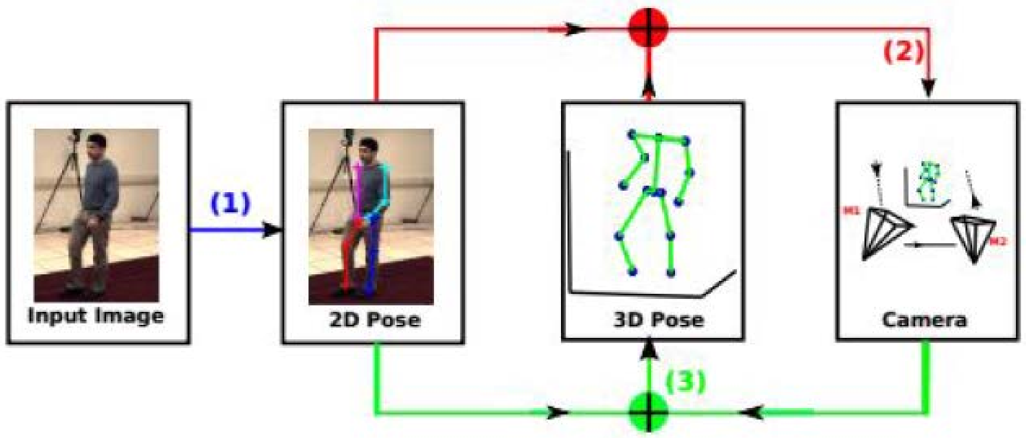
Human Parsing from Static Images and Sequences

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

Modeling 3D joint poses.

- Estimate 3D from 2D.
- Activated Simplex Models of Human Actions (3D data).
- Mining Key-Point Motifs.

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.

Overview of the Method

The main idea is to minimize the discrepancy between the projection of the inferred 3D pose and the 2D joint detection

- Alternately estimate Camera parameters & Human poses
- Reduce the ambiguity by **basis representation** & **anthropomorphic constraints**
- Reduce the influence of inaccurate 2D joint detections by minimizing L1-norm projection error

Representation

- 2D and 3D poses $x \in R^{2n}$ and $y \in R^{3n}$ are represented by n joint locations
- 2D and 3D poses are related by weak perspective camera parameters M

$$x = M \bullet y + t$$

- We assume 2D and 3D poses are mean-centered (t=0)

Representation

- 3D Human poses are known to lie on low-dimensional manifold
- Represent a pose y by a linear combination of bases (linear assumption)

b_i

$$y = \sum_{i=1}^{k} \partial_i \bullet b_i + \mu$$

- The representation reduces the ambiguity in 2D-3D pose mapping by restricting the set of 3D poses

Anthropomorphic Constraints

- Human poses are highly structured, e.g., limb ratios are almost the same for different people, which can explored to remove implausible configurations

Anthropomorphic Constraints

- Define joint selection matrix $E_j = [0, \cdots, I, \cdots 0]$, the j_{th} block is an identity matrix
- The product between E_j and y returns the j_{th} joint
- Let $C_i = E_{i1} - E_{i2}$, the product between C_i and y returns the limb
- We enforce limb length constraints on the inferred 3D pose

$$\| C_i \bullet (B\partial + \mu) \|_2^2 = L_i, i = 1 \cdots t$$

Objective Function

- We propose to minimize the discrepancy between the projection of the 3D pose and the 2D joint detections

$$\| x - M(B\alpha + \mu) \|$$

- This yields accurate 3D poses (were state-of-the-art when the work was published).

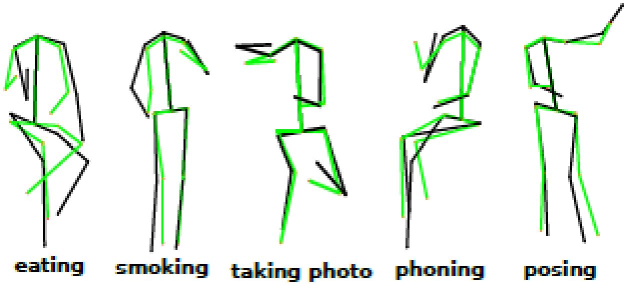
Encoding 3D poses: Action recognition

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

Generative Model: Tight Representation.

- 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.

Actions as Compositions of Key-Point-Motifs

- Recognizing actions from sequences of images is challenging.
- Challenges:
- Viewpoint – best to represent actions in 3D.
- Variability – different actors can perform the same task in different ways.
- Lack of Data – there may not be enough data for every action sequence.

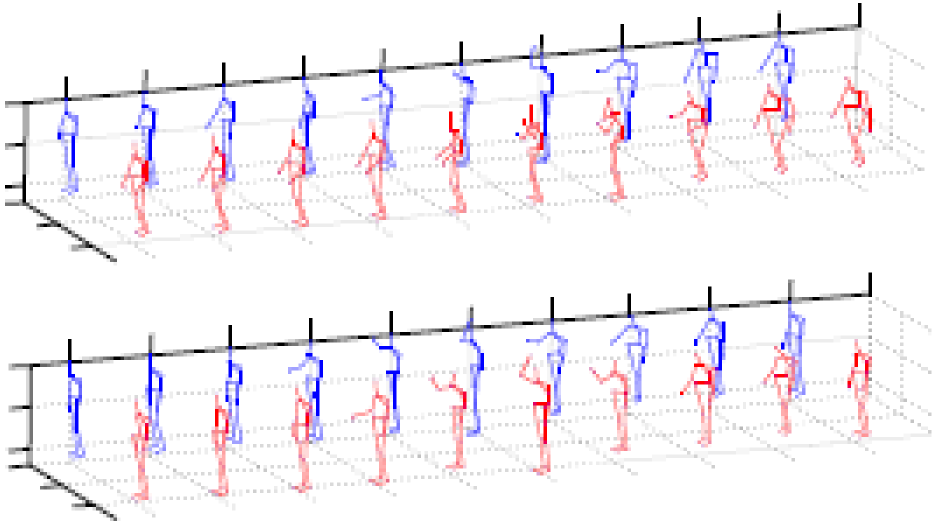
- Humans can generalize from few examples. Humans can sometimes recognize actions from single images. Suggests that humans can recognize actions for a few key poses instead of the whole pose sequence.

Pose Dictionaries and Key-Pose-Motif mining.

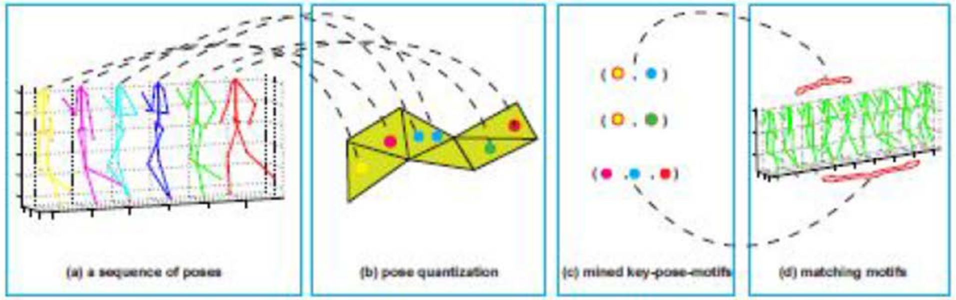
- We address this task by mining a set of key-pose-motifs for each action class.
- Represent each pose by a dictionary of possible poses (obtained by clustering). A pose sequence is represented as a sequence of these dictionary elements. Technically, we use soft-encoding so that poses can be represented by a weighted combination of a small number of dictionary elements.
- A key-pose-motif consists of a set of ordered poses, which are required to be close but not necessarily adjacent in the action sequences. A motif is called a key-pose-motif of a particular class if it appears in a sufficient number of sequences of that class.
- We mine several key-pose-motifs for each action class and classify an input sequence by finding the action class whose key-pose-motifs best match the sequence.
- Observe that this approach also has the ability to detect the start and finish of an action sequence by inspecting the matching results,

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).



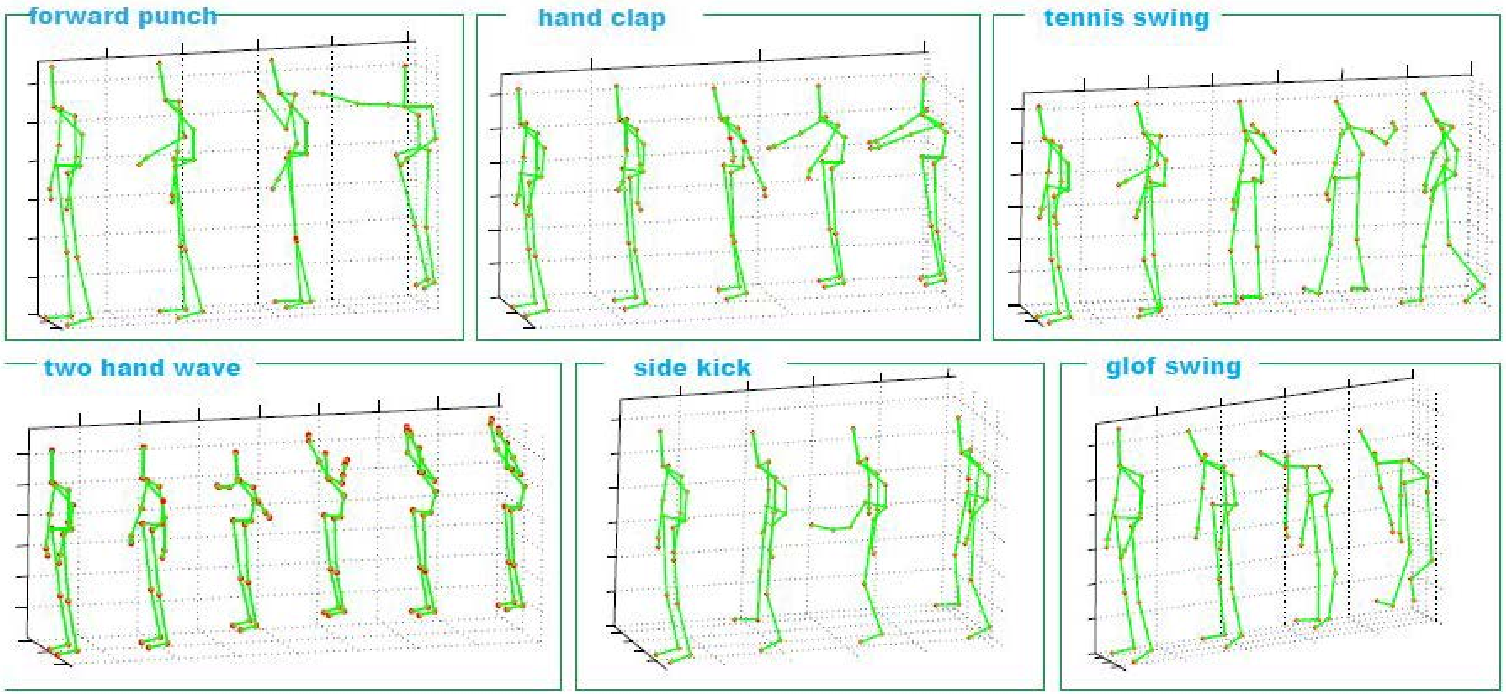
Key Point Motifs



- 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.

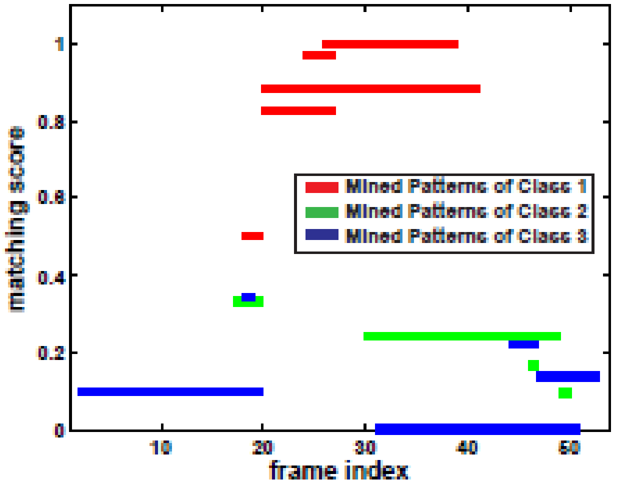


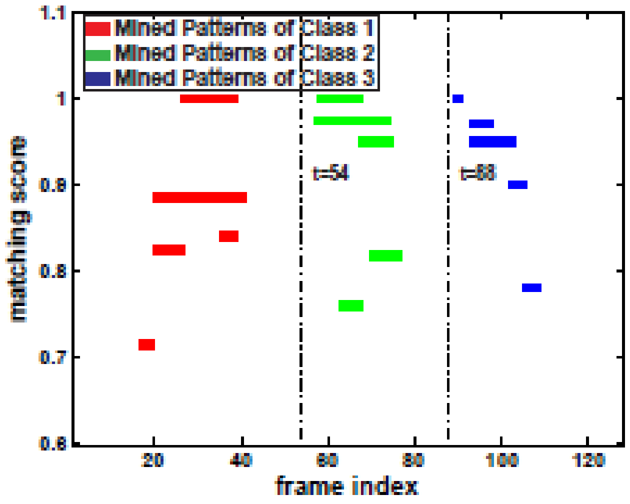
Key Point Motifs

- 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.
- Can also directly extend to detecting the start-points and end-points of activity sequences.

Action Sequences as Compositions

- Can represent action actions as compositions of key-point-motifs.
- Can represent more complex action sequences as compositions of elementary actions, each represented as compositions of key-point-motifs.