Hierarchical Models for 3D Visual Inference

Atul Kanaujia, *Rutgers University*
Cristian Sminchisescu, *TTI- Chicago*
Dimitris Metaxas, *Rutgers University*
3D Human Pose Inference `in the Wild’

Towards automatic monocular methods

• Background clutter
• Geometric transformations
  – Scale and viewpoint change
• Illumination changes
• Fast motions
• Occlusion / self-occlusion
• Variability in the human body proportions
Basic Approach

• Train structured model to predict 3D human poses given image descriptor inputs
  - Multi-valued predictor necessary

• Training using realistic CG impostor animated using human motion capture, and placed on real backgrounds
Problems

- Lack of typical training data (Vicon, quasi-real)
- Image descriptors are unstable \( w.r.t. \) to geometric deformations and background clutter
This Talk

• Learning image descriptors
  – Multi-level / coarse-to-fine encodings stable w.r.t. deformation and misalignment in the training set
  – Learning metrics for noise suppression / clutter removal

• Semi-supervised multi-valued prediction
What is wrong with existing features?

- Global histograms (bag of features) are robust to local deformation but sensitive to background clutter.

- Fine-scale, regular grid descriptors can be robust to clutter but sensitive to training set misalignments / local deformations.
Hierarchical Descriptor

- Multilevel coarse-to-fine encoding based on either multi-level histograms or successive object part matching and max pooling operations (e.g. HMAX)

SIFT Descriptor

Independent Vector Quantization in each spatial region

Concatenate to Descriptor
Hierarchical Features

HMAX Features

\( \Omega = \text{MAX} \)

S1

16 Gabor Filter response

C1

\( \Omega \)

S2

C2

Ω

Select Patches / Object Parts to match against results from previous layer

Encoding
Dealing with Background Clutter

- Multi-level encodings still affected by background
- Need to suppress noise
  - Eliminate features corresponding to irrelevant, fluctuating image regions
- Feature selection based on e.g. sparse linear regression tends to be ineffective for descriptors globally perturbed by clutter (e.g. global histograms)
Distance Metric Learning

- Learn a Mahalanobis distance that maximizes similarity within chunklets = sets of images of people in similar poses, but differently proportioned and placed on different backgrounds
  - Relevant Component Analysis (Hillel et. al. 2003)
  - Alternatively, maximize correlation between pairs
Background Clutter Removal

Clean | Quasi-real (QReal) | Real

- The distance between the learned multilevel descriptors computed on different backgrounds is diminished.
Canonical Correlation Analysis

- Learn bases that maximize correlation between pairs of images of people on clean and cluttered backgrounds.

![Graphs showing projection on canonical correlations.](image-url)
This Talk ... Flexible Training

- Multi-level / coarse-to-fine Image Descriptors stable \( w.r.t. \) image deformation and misalignment in training set.
- Learning metrics in the space of descriptors for noise suppression/ clutter removal
- Semi-Supervised multi-valued prediction
Semi-supervised Multi-valued Prediction

- Manifold Assumption
- Expert Ranking Assumption (mixture of experts)
  - If two inputs are closed in the intrinsic image geometry (e.g. graph Laplacian), their outputs should be smooth only if predicted by the same expert
Experiments

• Multi-level encodings
  – 5 Hierarchical Features, 5 Scale levels, 3 Coarse-to-Fine levels
  – ~1500D image descriptor

• Dataset of human poses
  – 56D human joint angle state vector
  – 5 Motions obtained with motion capture
    • Side walk, Pantomime, Bending Pickup, Dancing and Running
  – Multi-view walking with unlabelled images from INRIA Person Dataset
    – 3247 x 3 images, 1000 unlabeled images

• Multi-valued predictor with 5 experts
Prediction Accuracy
Multilevel vs. Global Descriptors

- Multilevel / hierarchical descriptors performs significantly better than global histograms or fine grids of local descriptors
Prediction Accuracy
Before / after Metric Learning

- Metric learning improves global histogram descriptors
- CCA improves HMAX
Run Lola Run Movie
Automatic 3D Pose Reconstruction

Notice: camera motion, occlusion (trees), self-occlusion, fast motions
Run Lola Run Movie
3D Pose Reconstruction

Notice scale changes, camera motions
Conclusions and Perspectives

• Learning multilevel image encodings for 3D inference
  – Stable w.r.t. to local deformation and misalignment in the training set
  – Metric learning for background clutter removal
• Flexible training based on semi-supervised multi-valued manifold regularization methods

• Ongoing work
  – Jointly learning both the image features and the predictor – can fit well within a hierarchical mixture of experts models
  – Scaling to large datasets
Semi-Supervised Learning

- Test on walking poses with unlabelled images from INRIA person database.
- Number of unlabelled data improves the prediction accuracy initially.
Thank You