Semi-supervised Hierarchical Models for 3D Human Pose Reconstruction

Atul Kanaujia
Rutgers University
kanaujia@cs.rutgers.edu

Cristian Sminchisescu
University of Bonn
Cristian.Sminchisescu@ins.uni-bonn.de

Dimitris Metaxas
Rutgers University
dnm@cs.rutgers.edu

Abstract

Recent research in visual inference from monocular images has shown that discriminatively trained image-based predictors can provide fast, automatic qualitative 3D reconstructions of human body pose or scene structure in real-world environments. However, the stability of existing image representations tends to be perturbed by deformations and misalignments in the training set, which, in turn, degrade the quality of learning and generalization. In this paper we advocate the semi-supervised learning of hierarchical image descriptions in order to better tolerate variability at multiple levels of detail. We combine multilevel encodings with improved stability to geometric transformations, with metric learning and semi-supervised manifold regularization methods in order to further profile them for task-invariance – resistance to background clutter and within the same human pose class *differences. We quantitatively analyze the effectiveness of both descriptors and learning methods and show that each one can contribute, sometimes substantially, to more reliable 3D human pose estimates in cluttered images.*

1. Introduction

One line of recent research in the area of monocular 3D human pose reconstruction has studied discriminative, feedforward models that can be trained to automatically predict pose distributions directly from image descriptors. This is in contrast with generative algorithms that search the pose space for configurations with good image alignment. Each class of methods has complementary trade-offs. Generative models are flexible at representing large classes of poses and useful for training and hypothesis verification, but inference is expensive and good observation models are difficult to construct without simplifying assumptions. Discriminative (feedforward) predictors offer the promise of speed, full automation and complete flexibility in selecting the image descriptor[1], but have to model multivalued image-to-3D relations and their reliance on a training set makes generalization to very different poses, body proportions, or scenes where people are filmed against background clutter, problematic. (N.B. Clearly, these remain hard problems for any method, be it generative or discriminative.)

The design of multi-valued feedforward pose predictors and the temporal density propagation in conditional chains is, at present, well understood, but the tradeoffs inherent in the acquisition of a sufficiently representative training set or the design of image descriptors with good resistance to clutter and intra-class variations is less explored. The construction of realistic *pose labeled* human databases (images of humans and their 3D poses) is inherently difficult because no existing system can provide accurate 3D ground truth for humans in real-world, non-instrumented scenes. Current solutions rely either on motion acquisition systems like Vicon, but these operate in engineered environments, where subjects wear special costumes and markers and the background is simplified, or on quasi-synthetic databases, generated by CG characters, animated using motion capture, and placed on real image backgrounds [3, 20]. In either case, there is a risk that models learned with these training sets may not generalize well when confronted with the diversity of real world scenes. A more flexible and scalable solution for model acquisition is necessary.

A second difficulty for reliable pose prediction is the design of image descriptors that are distinctive enough to differentiate among different poses, yet invariant to *within the same pose class differences* – people in similar stances, but differently proportioned, or photographed on different backgrounds. Exiting methods have successfully demonstrated that bag of features or regular-grid based representations of local descriptors (*e.g.* bag of shape context features, block of SIFT features [3, 20]) can be effective at predicting 3D human poses, but the representations tend to be too inflexible for reconstruction in general scenes. It is more appropriate to view them as two useful extremes of a multilevel, hierarchical representation of images – a family of

[1] Overcomplete bases or overlapping features of the observation can be designed without simplifying independence assumptions.

descriptors that progressively relaxes block-wise, rigid local spatial image encodings to increasingly weaker spatial models of position / geometry accumulated over increasingly larger image regions. Selecting the most competitive representation for an application – a typical set of people, motions, backgrounds or scales – reduces to either directly or implicitly learning a metric in the space of image descriptors, so that both good invariance and distinctiveness is achieved, *e.g.*, for 3D reconstruction, suppress noise by maximizing correlation within the desired pose invariance class, yet keep different classes separated, and turn off components that are close to being statistical random for the task of prediction, disregarding the class.

Our research brings together several innovations at the incidence of object recognition, metric learning and semi-supervised learning, as follows:

- We learn hierarchical, coarse to fine image descriptors that combine multilevel image encodings (inspired from object recognition, but here in the *different* context of 3D reconstruction) and metric learning algorithms. HMAX [16, 2], spatial pyramids [12], and vocabulary trees [14] are complemented with noise suppression and metric learning algorithms based on Canonical Correlation Analysis and Relevant Component Analysis. These refine and further align the image descriptors within individual pose invariance classes in order to better tolerate deformation, misalignment and clutter in the training and test sets.

- We construct models based on both labeled and unlabeled data, in order to make training with diverse, real-world datasets possible. We generalize semi-supervised regression models [7] to the more general problem of learning multi-valued predictors. We follow a manifold regularization approach in order to construct smoothness priors that automatically propagate outputs (poses) from labeled image descriptors to those unlabeled ones close in their intrinsic geometry, as represented, *e.g.*, by the graph Laplacian of a training set.

The two components are strongly dependent in practice. To make unlabeled data useful for generalization, perceptually similar descriptors have to be close in the selected input metric.[2] Learning an appropriate one becomes a necessary preliminary step.

1.1. Related Work

Our research relates to work in areas like discriminative human pose reconstruction, feature design, correlation analysis and semi-supervised learning. Several methods exists for discriminative pose prediction [15, 19, 20, 21, 3, 18] but they primarily concentrate on its multi-valuedness [15, 19, 18], or on single levels of feature encodings [19, 20, 21], based on global histograms or regular grids of SIFT blocks. Here we also study the general problem of 3D prediction for models trained using multilevel image encodings and many motions and tested on images with background clutter. Recent studies in object recognition [16, 12, 14, 2] have shown that multilevel image encodings can lead to improvements in image classification and retrieval performance. However, to our knowledge their potential, possibly in conjunction with metric learning and feature selection techniques, has not been investigated for 3D reconstruction. Learning a metric for clustering and image classification has been studied by [4, 24, 17] with methods differing in their treatment of equivalence constraints and the optimization performed. Some methods constrain the problem using only similar (image) instances, referred as chunklets, others build contrastive cost functions based on both similar and dissimilar class constraints, or learn projections that maximize mutual within-class correlation. Metric learning and correlation analysis can be useful for suppressing noise and discovering intrinsic, latent shared structure in data. They are adequate for our problem where image descriptors are affected by background differences and *within the same pose class* variations.

There is substantial work in semi-supervised learning [7], a methodology that uses both labeled and unlabeled data in order to construct more accurate models. Recent work in human tracking [13] showed promising results when learning mixtures of joint human poses and silhouettes, based on Expectation Maximization applied to partially labeled data. We follow a different approach inspired by manifold regularization [5], here *generalized to multi-valued prediction*. This is necessary because noise and perspective projection ambiguities manifest as distant 3D solutions for similar input descriptors. The smoothness prior assumptions typically used in semi-supervised regression only hold if appropriately qualified by the additional data partitioning constraints of multi-valued predictors.

2. Hierarchical Image Encodings

In this section we review the modified multilevel, hierarchical image descriptors we use as a basis for subsequent metric learning, described in §3.

HMAX [16] is a hierarchical, multilayer model inspired by the anatomy of the visual cortex. It alternates layers of template matching (simple cell) and max pooling (complex cell) operations in order to build representations that are increasingly invariant to scale and translation. Simple layers use convolution with local filters (template matching against a set of prototypes), in order to compute higher-

[2] This holds broadly, for both supervised and unsupervised methods.

order (hyper)features, whereas complex layers pool their afferent units over limited ranges, using a MAX operation, in order to increase invariance. Rather than learning the bottom layer, the model uses a bank of Gabor filter simple cells, computed at multiple positions, orientations and scales. Higher layers use simple cell prototypes, obtained by randomly sampling descriptors in the equivalent layer of a training set (k-means clustering can also be used), hence the construction of the hierarchical model has to be done stage-wise, bottom-up, as layers become available.

Hyperfeatures [2] is a hierarchical, multilevel, multi-scale encoding similar in organization with HMAX, but more homogeneous in the way it repeatedly accumulates / averages template matches to prototypes (local histograms) across layers, instead of winner-takes-all MAX operations followed by template matching to prototypes.

Spatial Pyramid [12] is a hierarchical model based on encodings of spatially localized histograms, over increasingly large image regions. The bottom layer contains the finest grid, with higher layers containing coarser grids with bag of feature (SIFT) encodings computed within each one. Originally, the descriptor was used to build a pyramid kernel as a linear combination of layered, histogram intersections kernels, but it can also be used stand-alone, in conjunction with linear predictors. It aligns well with the design of our 3D predictors, that can be either linear or kernel-based.

Vocabulary Tree [14] builds a coarse-to-fine, multilevel encoding using hierarchical k-means clustering. The model is learned divisively – the training set is clustered at top level, then recursively split, with a constant branching factor, and retrained within each subgroup. Nistér & Stévenius collect measurements on a sparse grid (given by MSER interest points) and encode any path to a leaf by a single integer. This is compact and gives good results for object retrieval, but is not sufficiently smooth for our continuous pose prediction problem, where it collapses qualitatively different poses to identical encodings. We learn the same vocabulary tree, but construct stage-wise encodings by concatenating all levels. At each level we store the continuous distances to prototypes and recursively descend in the closest sub-tree. Entries in unvisited sub-trees are set to zero. For each image, we accumulate tree-based encodings of patches on a regular grid and normalize.

Multilevel Spatial Blocks (MSB) is an encoding we have derived and consists of a set of layers, each a regular grid of overlapping image blocks, with increasingly large (SIFT) descriptor cell size. We concatenate descriptors within each layer and across layers, orderly, in order to obtain encodings of an entire image or sub-window.

3. Metric Learning and Correlation Analysis

Multilevel hierarchical encodings are necessary in order to obtain image descriptors with good resistance to deformations, clutter or misalignments in the training / test set. But they do not entirely eliminate the need for problem dependent similarity measures for descriptor comparison. Albeit multilevel encodings are in general more stable at preserving invariance to geometric transformations, their components may still be perturbed by clutter or may not be relevant for the task.

Linear predictors implicitly assume a Euclidean metric in input space, whereas kernel methods use an explicit metric induced by the selected kernel. In each case, there is no guarantee that an ad-hoc selected metric – a Euclidean distance or an RBF kernel with arbitrary covariance, would provide the best invariance w.r.t. the task *e.g.*, for 3D prediction, the invariance *within the same pose class*. In this section we review learning techniques to build a metric – or alternatively, to compute representations with implicit Euclidean metric, for a desired level of invariance. The training set consists of image descriptor examples of the same invariance class, here different people in roughly the same pose, but with different body proportions or viewed on different backgrounds.[3] In practice, each pose can define an invariance class but we will need to train with only a few qualitatively different poses in order to learn a useful metric.

Relevant Component Analysis (RCA): [4] is a metric learning method that minimizes the spread within each chunklet – a subset of examples obtained by applying a transitive closure on given equivalence relations, *e.g.* pairwise constraints on examples, here image descriptors. We use RCA to optimize a Mahalanobis distance for the image descriptors, with a constraint on the covariance matrix, in order to avoid trivial solutions that shrink the entire space. The cost function is:

$$\min_{\mathbf{D}} \frac{1}{U} \sum_{j=1}^{k} \sum_{i=1}^{U_j} ||\mathbf{r}_{ji} - \mathbf{m}_j||_{\mathbf{D}}, \text{ s.th. } |\mathbf{D}| \geq 1 \qquad (1)$$

where U is the total of examples and U_j is the number of examples $\mathbf{r}_{ji}, i \in \{1 \dots U_j\}$ in chunklet j, and $\mathbf{m}_j$ its mean. The solution to (1) can be obtained in closed form in 3 steps, for details see [4]: (1) subtract the mean of each chunklet from all its points, (2) compute the covariance matrix of each chunklet, and (3) sum the covariance matrices of all chunklets, scaled by the inverse number of examples $1/U$) and use it as a Mahalanobis distance (the RCA matrix that needs to be inverted has dimension $\dim(\mathbf{r})$).

Canonical Correlation Analysis (CCA): In its standard form, CCA [17] is a method to identify shared struc-

[3] No explicit 3D pose information is necessary.

ture among two classes of variables: the algorithm estimates two basis vectors so that, after linear projection[4], the correlation between the two classes is mutually maximized. Given two sets of vectors $\mathbf{r}$ and $\mathbf{u}$, as samples: $\mathbf{S} = ((\mathbf{r}_1, \mathbf{u}_1), (\mathbf{r}_2, \mathbf{u}_2), \ldots, (\mathbf{r}_n, \mathbf{u}_n))$, and their projection on two arbitrary directions, $\mathbf{w}_r$ and $\mathbf{w}_u$, with $\mathbf{S}_r = (\langle \mathbf{w}_r, \mathbf{r}_1 \rangle, \ldots, \langle \mathbf{w}_r, \mathbf{r}_n \rangle)$, and $\mathbf{S}_u = (\langle \mathbf{w}_u, \mathbf{u}_1 \rangle, \ldots, \langle \mathbf{w}_u, \mathbf{u}_n \rangle)$, CCA maximizes the cost:

$$f = \max_{\mathbf{w}_r, \mathbf{w}_u} \frac{\langle \mathbf{S}_r \mathbf{w}_r, \mathbf{S}_u \mathbf{w}_u \rangle}{||\mathbf{S}_r \mathbf{w}_r|| ||\mathbf{S}_u \mathbf{w}_u||} = \tag{2}$$

$$\max_{\mathbf{w}_r, \mathbf{w}_u} \frac{\mathbf{w}_r \top \mathbf{C}_{ru} \mathbf{w}_u}{\sqrt{\mathbf{w}_r^\top \mathbf{C}_{rr} \mathbf{w}_r \mathbf{w}_u^\top \mathbf{C}_{uu} \mathbf{w}_u}} \tag{3}$$

with $\mathbf{C}_{rr}$ and $\mathbf{C}_{uu}$ *within-sets* covariance matrices and $\mathbf{C}_{ru} = \mathbf{C}_{ur}^\top$ *between-sets* covariances. A closed form solution to (2) can be computed by solving a generalized eigenvalue problem of size $\dim(\mathbf{r}) + \dim(\mathbf{u})$. Large problems can be solved efficiently using predictive low-rank decomposition with partial Gram-Schmidt orthogonalization [17].

4. Manifold Regularization for Semi-supervised Multivalued Prediction

In this section we introduce semi-supervised extensions to multivalued predictive models. Existing models are primarily designed for supervised problems and represent the solution space (*e.g.* for 3D human pose estimation, the joint angles) using a mixture of image-based predictors. Each expert is paired with an observation dependent gate function that scores its competence in predicting states (3D) when presented with different inputs / images. As the input changes, different experts are active and their rankings (relative probabilities) change. The model is essentially a mixture of predictors with input-sensitive mixing proportions:

$$p(\mathbf{x}|\mathbf{r}) = \sum_{i=1}^{M} g_i(\mathbf{r}) p_i(\mathbf{x}|\mathbf{r}) \tag{4}$$

$$g_i(\mathbf{r}) = \frac{\exp(\boldsymbol{\lambda}_i^\top \mathbf{r})}{\sum_k \exp(\boldsymbol{\lambda}_k^\top \mathbf{r})} \tag{5}$$

$$p_i(\mathbf{x}|\mathbf{r}) = \mathcal{G}(\mathbf{x}|\mathbf{W}_i \mathbf{r}, \boldsymbol{\Omega}_i^{-1}) \tag{6}$$

with $\mathbf{r}$ image descriptors, $\mathbf{x}$ state outputs, g_i input dependent gates, computed using linear regressors *c.f.* (5), with weights $\boldsymbol{\lambda}_i$. g are normalized to sum to 1 for any given input $\mathbf{r}$ and p_i are Gaussian functions (6) with covariance $\boldsymbol{\Omega}_i^{-1}$, centered at the expert predictions, here chosen to be linear regressors with weights $\mathbf{W}_i$.

A semi-supervised extension of this model would combine *both* labeled *and* unlabeled data in order to constrain the parameter estimates of each expert p_i. A standard semi-supervised learning assumption can be stated as follows: *if two inputs* $\mathbf{r}$ *in a high density region are close, so should be their corresponding outputs* $\mathbf{x}$. For our problem and predictor, this assumption is adapted as follows:

- Manifold assumption: If two points are close in the intrinsic geometry of $p(\mathbf{r})$ (given by a manifold or graph regularizer, see below), their conditional distributions $p(\mathbf{x}|\mathbf{r})$ should be similar.
- Expert responsibility assumption: If two inputs $\mathbf{r}$ are close (in the intrinsic geometry) *and can be predicted by expert i with confidence* g_i, their corresponding conditional distributions $p(\mathbf{x}|\mathbf{r})$ should be smooth to the same extent (*i.e.* modulated by g_i).

For the linear case, the semi-supervised manifold assumption manifests as a prior on the weights of each expert i, or equivalently, a (negative log-likelihood) regularization term (constants and scaling factors dropped for simplicity):

$$\mathcal{R}_i = \sum_{u,j=1}^{U} (\mathbf{W}_i \mathbf{r}_u - \mathbf{W}_i \mathbf{r}_j) N_{uj} (\mathbf{W}_i \mathbf{r}_u - \mathbf{W}_i \mathbf{r}_j)^\top \tag{7}$$

$$= \mathbf{W}_i \mathbf{R}^\top \mathbf{L} \mathbf{R} \mathbf{W}_i^\top \tag{8}$$

where U is the size of the entire training set including the unlabeled points, $\mathbf{R}$ is a $\dim(\mathbf{r}) \times U$ matrix that stores all the input vectors $\mathbf{r}$ in the training set (supervised and unsupervised), $\mathbf{L}$ is a graph Laplacian regularizer constructed over all the training set[5]: $\mathbf{L} = \mathbf{D} - \mathbf{N}$ with $\mathbf{N}$ a matrix of graph weights N_{ij} and $\mathbf{D}$ a diagonal matrix with elements $D_{ii} = \sum_{j=1}^{U} N_{ij}$. The input geometry is not distorted because the manifold regularization framework does not compute a low-dimensional embedding explicitly. The framework only implicitly assumes that *some* intrinsic geometry, embedded in $\mathbb{R}^{\dim(\mathbf{r})}$, exists.

Eq. (7) can be interpreted as a ridge-style regression prior on the expert weights, with a special covariance matrix given by the graph Laplacian. The prior is computationally tractable – it contributes as yet another matrix to the existing ones corresponding to the labeled data or the expert weight priors[6] – which are inverted in order to compute each expert. Learning is performed iteratively, using an EM algorithm that computes soft assignments of each datapoint to

[4] Non-linear extensions can be obtained in the usual way, using 'kernelization' [17].

[5] This is (typically) a sparse graph construction, obtained by connecting each training point to its k-nearest neighbors and computing local Gaussian distances to them. A global regularizer based on geodesic distances can also be used.

[6] We also use a sparsity weight prior, but this does not appear explicitly in the sum that accumulates the matrix to be inverted – for linear methods sparsity contributes by decreasing the effective input dimension, hence the size of the matrix.

experts and learns both the experts and their gates using a double loop (expert-expert) estimation scheme.

5. Experiments

In this section we report experiments obtained using 5 different multilevel image encodings further profiled using 2 different metric learning and noise suppression methods (RCA and CCA), and the semi-supervised manifold regularization framework based on both labeled and unlabeled data.

Multilevel Encodings: We use 5 different hierarchical encodings, calibrated to roughly similar dimensionality: HMAX (1600, 4 levels with patch size 4-16 codebooks, obtained by sampling from 1600 real images - the same set was used to generate codebooks for all methods which need them), Spatial Pyramid (1400, 3 levels, SIFT descriptors, 6x6 pixel cells, 4x4 cells per block, 4 angular bins of gradient orientations, $0 - 180^o$ unsigned,10 pixel grid overlap), Hyperfeatures (1400, SIFT descriptors, 4x4 blocks, 4x4 pixels per cell, 3 levels having 200, 400, and 800 centers with scales 2, 4, 6), Vocabulary Tree (1365, 5 levels, branching factor 4, SIFT descriptors computed at 5 different scales of a Gaussian pyramid, SIFT descriptors, 4x4 blocks, 4x4 pixels per cell) and Multilevel Spatial Blocks (1344, 3 levels with 16, 4, 1 SIFT block, 4x4 cells per block, 12x12 cell size).

Database and Multivalued Predictor: For qualitative experiments we use images from a movie (Run Lola Run) and the INRIA pedestrian database [8]. For quantitative experiments we use our own database consisting of $3 \times 3247 = 9741$ quasi-real images, generated using a computer graphics human model that was rendered on real image backgrounds. We have 3247 different 3D poses from the CMU motion capture database [1], rendered to produce different viewing patterns of walks, either frontal or parallel to the image plane, dancing, conversation, bending and picking, running and pantomime (one of the 3 sets of 3247 poses is placed on a clean background). We collect three test sets of 150 poses for each of the five motion classes. The motions executed by different subject are not in the training set. We also render one test set on a clean background as baseline (Clean). The other two test sets are progressively more complicated: one has the model *randomly* placed at different locations, but on the same images as in the training set (Clutter1), the other has the model placed on unseen backgrounds (Clutter2). In all cases, a 320x240 bounding box of the model and the background is obtained, possibly using rescaling [20]. There is significant variability and lack of centering in this dataset because certain poses are vertically (and horizontally) more symmetric than others (*e.g.* compare a person who picks an object with one who is standing, or pointing the arm in one direction). We train a multivalued predictor (a conditional Bayesian mixture of 5 experts) on the entire dataset, as opposed to training models for each motion / activity separately. The model uses linear experts with sparsity priors, complementing the closeform pre-processing from RCA / CCA. Empirically, we observe sparsity patterns in the range $15\% - 45\%$, with lower values usually associated with models that generalize better. For the quantitative experiments, the 56d human joint angles were reduced to 8d using PCA. This fast and mapping to joint angles is exact, as opposed to approximate in kernelPCA or other latent variable models. For our experiments, we wished to factor out the variability in the dimensionality reduction mapping, but non-linear methods can be used, alternatively.

Metric Learning and Correlation Analysis: This stage does not require explicit 3D pose information – it works entirely on sets of image descriptors. We train using corresponding doublets in 750 images, each pair $(\mathbf{r}, \mathbf{u})$ shows our model rendered in the same pose, on both on clean and a (varying) cluttered background. Each pair is given as a separate chunklet to RCA. For CCA we give the corresponding pairs of vectors. RCA requires a matrix inversion and CCA solves an eigenvalue problem. In each case, regularization with a scale of identity matrix usually helps performance (*e.g.* for CCA, the dimensionality of the image descriptors is larger than the size of the training set). The behavior of the two methods is illustrated in fig. 1 and fig. 2. After metric learning, the dimensionality of the image encoding changed to (this was the input descriptor used to train multivalued predictors): HMAX – 1174, Hyperfeatures – 1073, Spatial Pyramid – 1076, Vocabulary Tree – 1059, Multilevel Spatial Blocks – 1048.

Cumulative results of our tests on the quasi-real databases, on the Clutter2 set, previously unseen are given in fig. 3 (details for each motion on both Clutter1 and Clutter2 will be available in an upcoming TR – in general, performance on Clutter1 is better, but the problem is arguably simpler). The plots in the first two rows give prediction error per joint angle for different multilevel encodings and the two metric learning methods. The bottom row in fig. 3 shows 'marginal projections', computed for one class of activities (bending and picking). In our experiments HMAX works best, followed closely by the MSB and Hyperfeatures. For such features little improvement or even performance drops are observed following metric learning on Clutter2. One reason may be that features are well encoded already with localized entries that are contaminated by clutter. Our linear experts being sparse, they are capable of feature selection and noise suppression at the (more informed, albeit greedy) level of 3D prediction. While the Spatial Pyramid and the Vocabulary tree performed less well in their original encoding, RCA improved them substantially. We assume this happens because both the Vocabulary Tree

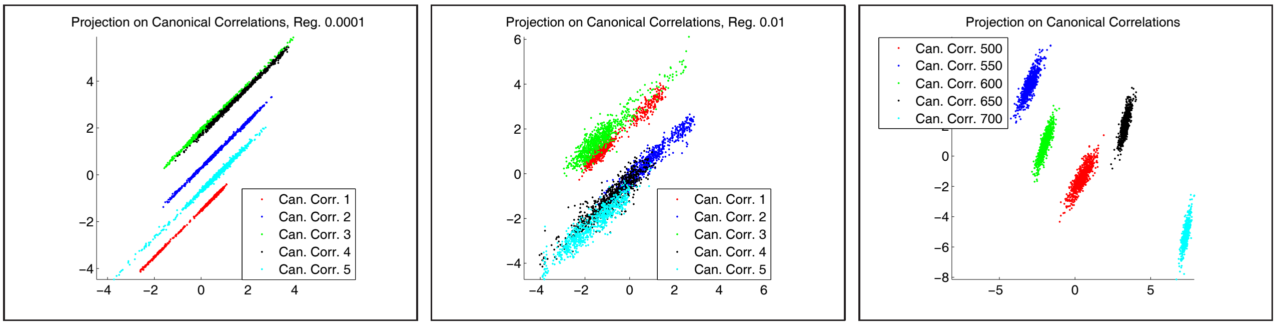

Figure 2. Projection of the training set on different canonical correlations. *(a) Left* and *(b) Middle* show the top most correlated components, differently colored (these correspond to the largest eigenvalues), for two levels of regularization - we plot pairs of vector components, hence good correlations (*i.e*. similar values) are achieved when their slope is close to 45^o. *(c) Right* shows un-correlated directions corresponding to low eigenvalues – notice the deviation from 45^o.

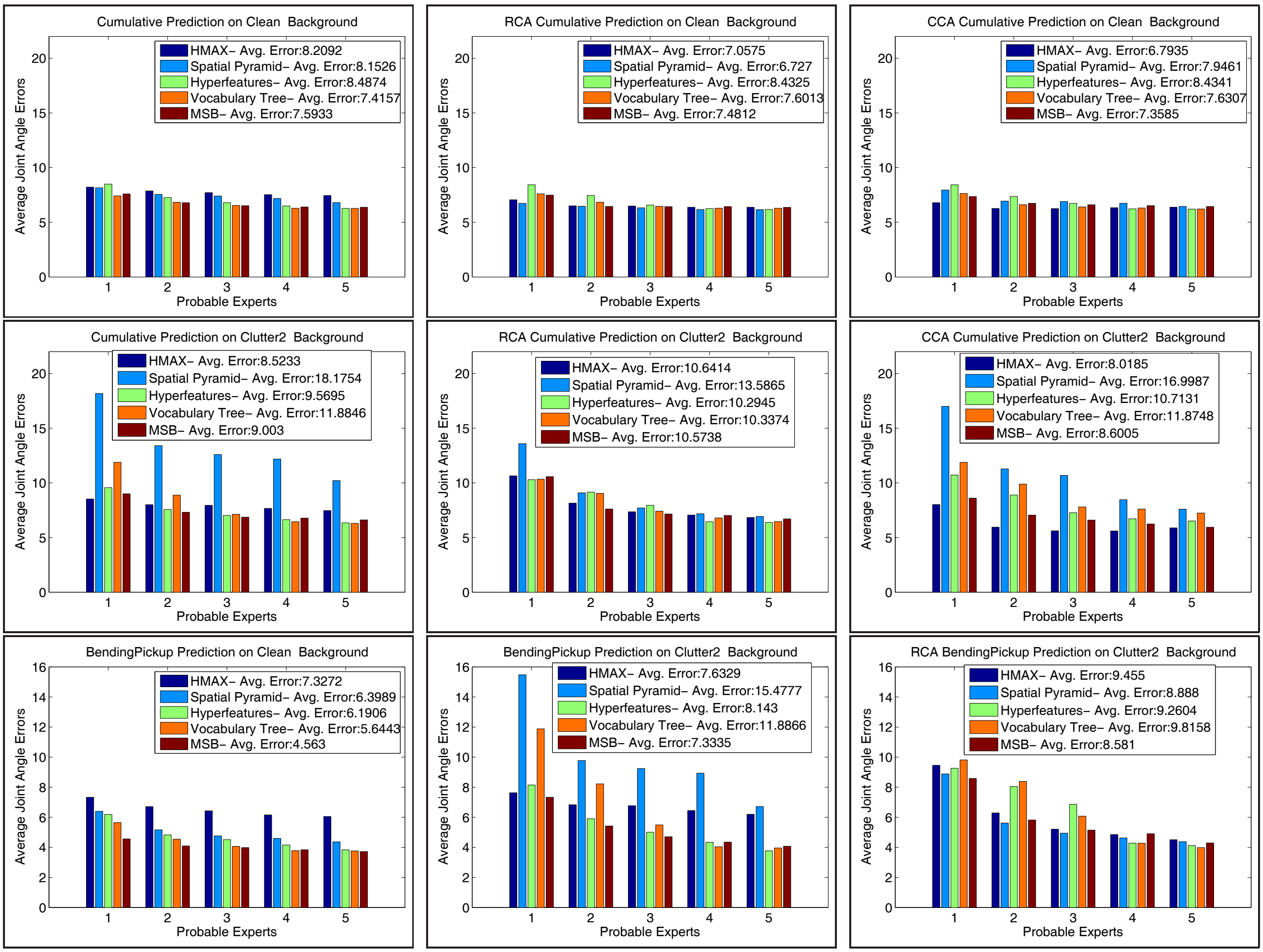

Figure 3. *(a) Top row* and *(b) Middle Row:* Quantitative results cumulated for 5 different motions, 5 image encodings: HMAX, Hyperfeatures, Spatial Pyramid, Vocabulary Tree, Multilevel Spatial Blocks (MSB), and 2 metric learning and correlation analysis methods (RCA, CCA). *(c) Bottom row:* Details from *(a)* and *(b)* for the subset of bending and picking motions in the training set. A single (global) model (a conditional Bayesian mixture of experts) was trained on the entire dataset. Each plot shows the error in the best-k experts, for $k = 1 \dots 5$, the total number of experts used. The k-th bar was computed by selecting the value closest to ground truth among the ones predicted by the most probable k experts.

and the Spatial Pyramid are based on more globally computed histogram blocks. This tends to perturb a large number of descriptor components and makes noise suppression by sparsification at the level of individual entries less effec-

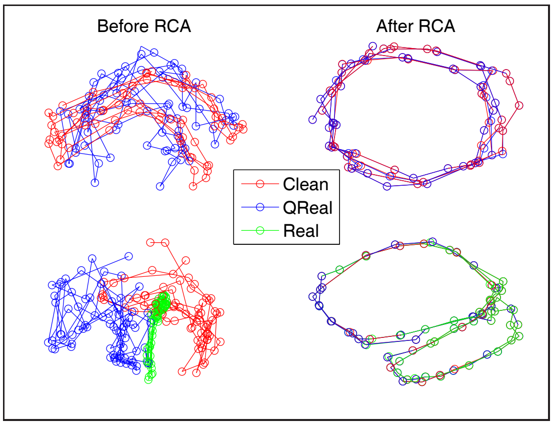

Figure 1. 3d embedding of images encoded using Multilevel Spatial Blocks (MSB), before and after metric learning with RCA (images of walking parallel to the image plane). We use use different training sets with different degrees of realism. The inclusion of pairs of clean and real images of similar 3D poses as chunklets in RCA significantly improves the descriptor invariance to clutter. This *does not* introduces walking half cycle ambiguities, the bottom-rights shows the 2d projection of a somewhat twisted (not self-intersecting) 3d loop.

tive. For low training error, the predictor is forced to either keep a large number of entries – but then it usually overfits, or to sparsify them aggressively. But for globally noisy descriptors, any remaining subset is noisy / unstable and this increases training error. In this case, preprocessing using more global noise suppression methods like RCA seems appropriate. An alternative may be to use problem-dependent kernels, *e.g*. histogram intersections [12], with good resistance to noise and image mismatches. Our kernel-based multivalued predictors can, in principle, use histogram kernels (this is currently investigated).

We have also run experiments using the manifold regularization framework (fig. 4), where we trained several models on a small dataset of only 150 cluttered poses (we subsampled our database by a factor of 20) and progressively added unlabeled data in the range of $150 - 450$ datapoints. Here we show examples for the Multilevel Spatial Block (MSB) encoding, with different distance metric learning algorithms (NML refers to a model with no metric learning). The addition of unlabeled data improves performance, especially for models trained using RCA. One reason NML or CCA show performance drops may the use of an incorrect descriptor metric – manifold regularization relies on good input similarity in order to smooth the output. If this doesn't hold, semi-supervised learning may be less effective.

The finals set of results we show in fig. 5 is based on real images from the INRIA pedestrian dataset [8], and the movie 'Run Lola Run', where our model is seduced by Lola and runs after her. These are all automatic 3D reconstructions of fast moving humans in non-instrumented environments, filmed under significant viewpoint and scale changes. We use a model trained with 2000 walking and running labeled poses (quasireal data of our graphics model placed on real backgrounds, rendered from 8 different viewpoints) with an additional 1000 unlabeled (real) images of humans running and walking in cluttered scenes. The 3D reconstructions have good perceptual accuracy, although the solutions are not entirely accurate in a classical alignment sense. This is mostly caused by the lack of typical data in the CMU dataset – *e.g*. we only trained on pedestrians walking, yet in many images pedestrians are standing with one hand in their pocket, hold a purse, *etc*. More diverse training sets are likely to significantly improve accuracy.

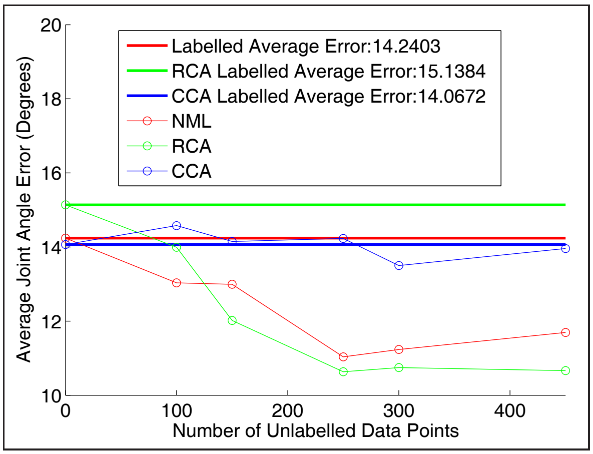

Figure 4. Semi-supervised learning. We compare manifold regularization with up to 450 unlabeled data points with baseline models trained using 150 samples in a supervised training set, for the MSB encoding and different combinations of metric learning methods (NML uses MSB, without any metric learning). While NML and CCA based models show improvement followed by performance degradation as more data is added, models based on descriptors with learned metrics perform best.

6. Conclusions

We have argued that the robustness of discriminative 3D predictors is affected by three main factors: (1) the image encoding, (2) the metric chosen in the space of encodings, and (3) the capacity to flexibly extend the training set with unlabeled real world data (here images), which, for 3D problems, is significantly easier to collect than realistically looking labeled one. To make this possible, we have advocated the learning of hierarchical descriptors by profiling multilevel, coarse to fine, image encodings using metric learning and correlation analysis. Finally, we showed how unlabeled data can be incorporated for 3D reconstruction by generalizing semi-supervised manifold regularization to multivalued prediction – propagating information from labeled to unlabeled inputs using their intrinsic geometry and the different expert (predictors) responsibility constraints.

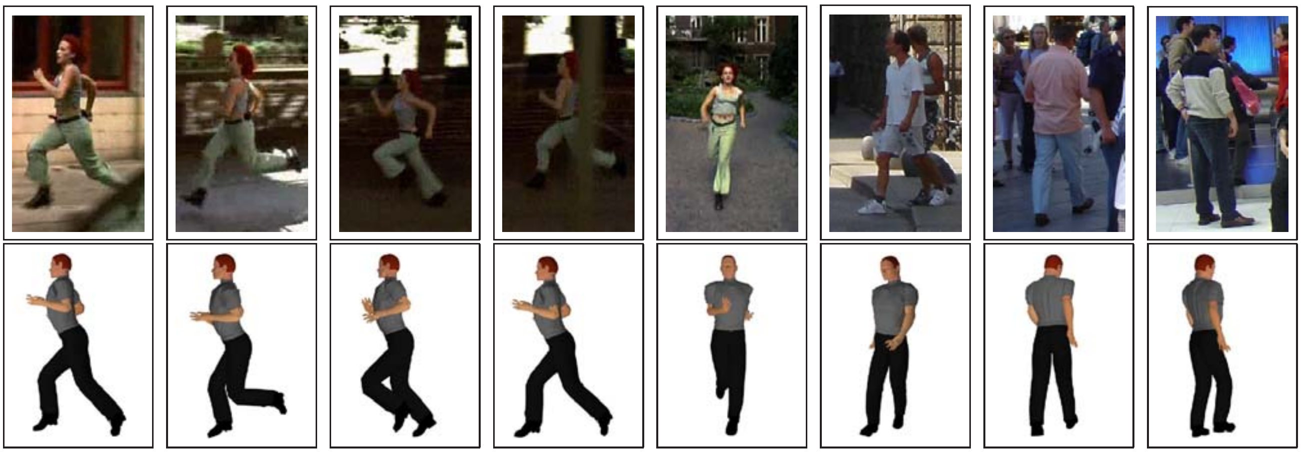

Figure 5. Qualitative 3D reconstruction results obtained on images from the movie 'Run Lola Run' (block of leftmost 5 images) and the INRIA pedestrian dataset (rightmost 3 images) [8]. *(a) Top row* shows the original images, *(b) Bottom row* shows automatic 3D reconstructions.

In our tests, each of the three components provided performance gains. Empirically, we also observe that a combined system improves the quality of 3D human pose prediction in images and video.

Ongoing work: We currently investigate methods to scale the existing algorithms to large datasets, possibly with a large unlabeled component and the use of nonlinear correlation methods. We also plan to study the construction of models that gracefully degrade with occlusion.

Acknowledgements: This work has been partly supported by the NSF, under award IIS-0535140 and EC MCEXT-025481.

References

[1] CMU Human Motion Capture DataBase. Available online at http://mocap.cs.cmu.edu/search.html, 2003.

[2] A. Agarwal and B. Triggs. Hyperfeatures – Multilevel Local Coding for Visual Recognition. In *ECCV*, 2006.

[3] A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV*, 2006.

[4] A. Bar-hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *ICML*, 2003.

[5] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *AISTATS*, 2005.

[6] M. Bray, P. Kohli, and P. Torr. Posecut: Simultaneous segmentation and and 3d pose estimation of humans using dynamic graph cuts. In *ECCV*, 2006.

[7] O. Chapelle, B. Scholkopf, and A. Smola. *Semi-supervised Learning*. MIT Press, 2006.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR*, 2004.

[10] T. Jaeggli, E. Koller-Meier, and L. V. Gool. Monocular tracking with a mixture of view-dependent learned models. In *AMDO*, pages 494–503, 2006.

[11] X. Lan and D. Huttenlocher. Beyond trees: common factor models for 2d human pose recovery. In *ICCV*, 2005.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[13] R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. Semi-supervised learning of joint density models for human pose estimation. In *BMVC*, 2006.

[14] D. Nistér and H. Stévenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[15] R. Rosales and S. Sclaroff. Learning Body Pose Via Specialized Maps. In *NIPS*, 2002.

[16] T. Serre, L. Wolf, and T. Poggion. Object recognition with features inspired by visual cortex. In *CVPR*, pages 994–1000, Washington, DC, USA, 2005.

[17] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[18] L. Sigal and M. Black. Predicting 3d people from 2d pictures. In *AMDO*, 2006.

[19] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, volume 1, pages 390–397, 2005.

[20] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning Joint Top-down and Bottom-up Processes for 3D Visual Inference. In *CVPR*, 2006.

[21] C. Sminchisescu, A. Kanaujia, and D. Metaxas. BM^3E: Discriminative Density Propagation for Visual Tracking. *PAMI*, 2007.

[22] C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *ICCV*, 2003.

[23] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking in small training sets. In *ICCV*, 2005.

[24] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2002.