Learning Markerless Human Pose Estimation from Multiple Viewpoint Video

Matthew Trumble, Andrew Gilbert, Adrian Hilton and John Collomosse

Centre for Vision Speech and Signal Processing (CVSSP)
Univeristy of Surrey, Guildford, UK.

Abstract. We present a novel human performance capture technique capable of robustly estimating the pose (articulated joint positions) of a performer observed passively via multiple view-point video (MVV). An affine invariant pose descriptor is learned using a convolutional neural network (CNN) trained over volumetric data extracted from a MVV dataset of diverse human pose and appearance. A manifold embedding is learned via Gaussian Processes for the CNN descriptor and articulated pose spaces enabling regression and so estimation of human pose from MVV input. The learned descriptor and manifold are shown to generalise over a wide range of human poses, providing an efficient performance capture solution that requires no fiducials or other markers to be worn. The system is evaluated against ground truth joint configuration data from a commercial marker-based pose estimation system.

Keywords: Deep Learning, Pose Estimation, Multiple Viewpoint Video

1 Introduction

Performance capture is used extensively within the creative industries for character animation and visual effects. Current performance capture requires the actor to wear a special suit either augmented with retro-reflective markers (e. g. Vicon, Optitrack), or illustrated with high contrast multi-scale fiducials (e. g. ILM Fractal suit) from which an estimate of human pose is derived, usually as a sequence of skeletal joint angles. The use of unsightly markers prohibits co-capture of the pose with principal footage (i. e. roll visible in the final production), requiring multiple takes and so time and expense. Furthermore, many commercial performance capture systems require a large number of specialist cameras (typically infra-red) to be setup which takes time and restricts shooting to artificially lit locations. The contribution of this paper is a technique to estimate human pose sequences from the principal footage, using a set of synchronized video sequences shot from multiple static views. The acquisition of multiple viewpoint video (MVV) on-set is commonplace, and so this represents a practical cost saving to production. The proposed algorithm is the first to leverage deep convolutional neural networks (CNNs) to obtain a robust 3D human pose estimate from volumetric data recovered from MVV footage.

2 Related Work

Human pose estimation (HPE) is the task of estimating either a skeletal pose or a probability map indicating likely positions of skeleton limbs. HPE commonly

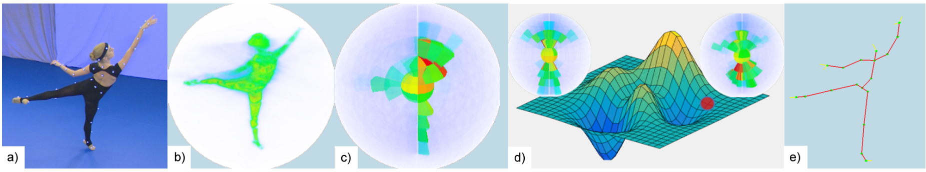

Fig. 1: Overview of proposed technique. MVV is captured (a) and a geometric proxy / PVH built (b). The PVH is sampled into log-polar form at multiple scales and passed through a CNN to learn a rotationally invariant descriptor (c). A non-linear manifold embedding of the combined CNN and joint angle space (d) is learned under supervision to regress a pose estimate (e).

begins with the localization of people in images. The localization problem can be solved by background subtraction [1, 2] or in cluttered scenarios, sliding window classifiers can robustly identify the face [3] or torso [4] to bootstrap limb labelling and subsequent pose estimation. Following localization, pose estimation can be approached either by a) top-down fitting of an articulated model, via optimizing joint parameters and evaluating the correlation of the fitted model with image data; or b) detection-led strategies in which body parts are labeled independently and their poses integrated to estimate full body pose in a bottom-up manner.

Following the results of Krizhevsky et al. [5], the benefits of deeply learned convolutional neural networks (CNN) have been explored for both 2D HPE [6] and more general 3D object pose estimation [7, 8] within photographs. Deeply learned descriptors have recently shown promise in estimating 2D limb positions within very low-resolution images of human pose [9]. Yet although the problem of aligning pairs of 3D human body poses has been explored using deep learning [10], the estimation of 3D pose from MVV remains largely unexplored. Arguably the most closely related work is that within free-viewpoint video reconstruction where skeletal pose may be recovered by manually attaching limbs to vertex clustered in a tracked 4D mesh [11]. Other methods reliant on frame to frame tracking use a CNN for body part detections in 2D which are fused into 3D pose [12]. However both tracking and detection are reliant upon strong surface texture cues and absence of surface deformation.

3 Markerless Pose Estimation

Our approach accepts a multiple viewpoint video (MVV) sequence as input, shot using synchronised calibrated cameras surrounding the performance. A geometric proxy of the performer is built for each frame of the sequence via an adapted form of Grauman et al.'s probabilistic visual hull (PVH) [13] over a grid of voxels of resolution 5cm^3, computed from soft foreground mattes extracted from each camera image using a chroma key.

A dynamic threshold is applied to the voxel distribution to normalise against appearance variation between performers and across datasets. This is computed by analysing the voxel occupancy distribution to identify the background noise level. The proxy is then resampled into a log-polar representation $\mathcal{S}(\phi, \theta)$, quantizing longtitude and latitude into N regular intervals, aggregating voxels from the subvolume of the PVH within a particular distance interval from the centroid, and fed into a convolutional neural network (CNN) configured for a supervised

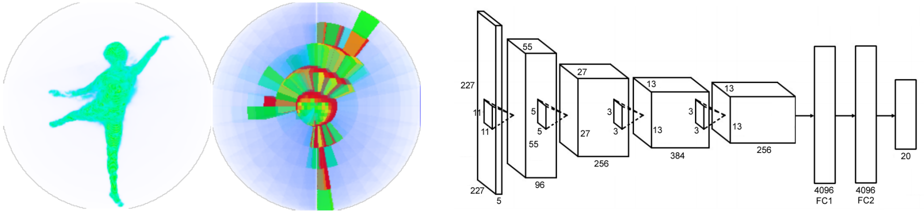

Fig. 2: Architecture of the proposed CNN (right) operating over the multi-scale log-polar representation parsed from MVV (left) and normalised against appearance variation via the dynamic thresholding operation (middle).

classification task. The use of a log-polar representation follows successes in prior work on human 3D mesh alignment [14] and general 3D object retrieval [15] that employ spherical histogram representations to match on coarse shape. We investigate removal of phase information by computing a frequency domain (DFT) representation of each row of the spherical histogram, and considering only the complex i. e. frequency magnitude. Similar to classical Fourier Descriptors this results in a shorter descriptor invariant to rotation of signal.

The CNN is trained using labeled examples of several distinct poses exercising the full range of typical human motion. Descriptors are extracted from the second fully connected layer of the network and a non-linear manifold embedding learned over a combined space of the CNN descriptors and joint angle estimates (subsec. 3.2). The manifold enables pose regression from descriptors derived from each MVV frame. Fig. 1 illustrates the full pipeline.

3.1 CNN Training

Our CNN adapts the architecture of [5] and is illustrated in Fig. 2. Similar to modern image classification work, which now extensively employs CNNs, we sample a high-dimensional descriptor from the second fully connected layer following training convergence. We evalute (Sec. 4) fully connected layers of 1024 (1K) and 4096 (4K) leading to descriptors of similar dimension. The CNN was trained to perform a supervised pose classification task using a purpose-built dataset of labeled MVV footage comprising $\sim 25k$ multiple-view frames from 8 cameras. 25 individuals in a variety of clothing were filmed executing repetitions of 20 distinct poses following the Vicon "Range of Motion" (ROM) sequence used to calibrate commercial motion capture equipment to exercise all major modes of human pose variation. Soft-max loss was used to train the CNN using 80% of this data to recognize the 20 poses, subject to two data augmentation strategies: **DA1:** Longitude Jitter. $\mathcal{S}(\phi, \theta)$ subject to random rotation of $\theta = [0, 2\pi]$. **DA2:** As DA1 with the addition of Gaussian noise and blur at random scale. Training proceeded over 100 epochs in our experiments, using a mini-batch size of 200. At test time, the CNN is truncated at the second fully connected layer yielding a vector of convolutional feature responses $\mathcal{C}$ that serves as our pose descriptor.

3.2 Joint Manifold Embedding

We perform human pose estimation via supervised learning in which a correlation is learned between exemplar pairs of descriptors (in CNN space $\mathcal{C}$) and a vector of 21 skeletal joint angles expressed in quaternion form (we denote this space

Table 1: Classification accuracy of DFT and CNN based descriptors

Descriptor type	Classifier	MAP(%)
DFT	SVM	75.62
4K CNN+DA1	CNN	77.55
4K CNN+DA2	CNN	80.97
4K CNN+DA2	SVM	**87.99**

$\mathcal{Q} \in \Re^{21\times 4}$). We investigate three approaches to the generalisation of these sparse training correspondences to a dense mapping $\mathcal{C} \mapsto \mathcal{Q}$ suitable for inferring performer pose $P \in \mathcal{Q}$ from a query point $c \in \mathcal{C}$ derived from MVV at test time.

Nearest Neighbour (Baseline) The naïve approach to creating a dense mapping is to snap a query pose descriptor to the closest $c_i \in \mathcal{C}$ i. e. perform a nearest neighbour lookup to obtain pose estimate P_{nn}. This can be implemented in real-time (i. e. 25 frames/second) using a kd-tree pre-built over c_i. Under this approach no constraints are imposed to guard against invalid poses, since no generalisation beyond training is performed.

Piecewise linear embedding (Baseline) A linear subspace model is learned local to each c_i based on the local K most proximate training samples c'_j where $j = \{1..K\}$. We construct this model as an undirected graph connecting c_i to c'_j, forming a piecewise linear manifold over $\mathcal{C}$ covering likely poses and (linear) interpolations between similar poses. In our experiments $K = 5$ provides a balanced trade-off between speed and accuracy. We estimate the pose P_{ple} under this model as $P_{ple} = \sum_{j\in J} d(c, c_j)q_j$, where $d(a, b)$ is a value proportional to geodesic distance between two points on the graph manifold, and J is the set of K nearest neighbours to c in $\mathcal{C}$.

Non-linear embedding Gaussian processes (GP) [16] are a popular approach for creating smooth non-linear mappings between continuous spaces of differing dimension. We adopt the Gaussian Process Latent Variable Model (GP-LVM) [17] as a supervised means for learning a non-linear manifold embedding within joined space $\mathcal{C} \times \mathcal{Q}$ i. e. to model the manifold upon which vectors $[c_i \ q_i]$ lie, from which we can generate a pose estimate P_{nle}.

4 Experiments and Discussion

Pose Classification Experiments We evaluated the CNN architecture under the two proposed data augmentation strategies on direct CNN classification and a non-linear SVM classifier using the FC2 descriptor. The mean average precision (MAP) score over the test data is shown in Table 1. Comparing augmentation strategy **DA2** against the **DFT** encoding of the log-polar representation, performance increases around 5%. However when the FC2 layer is used as descriptor in conjunction with an SVM, there is a 12% increase of classification performance. Much of the remaining limited confusion occurs between left and right variants of the pose classes. The high performance of FC2 derived descriptors implies the CNN has not only learned strong pose discrimination but that we are able to use it to produce a descriptor for pose estimation.

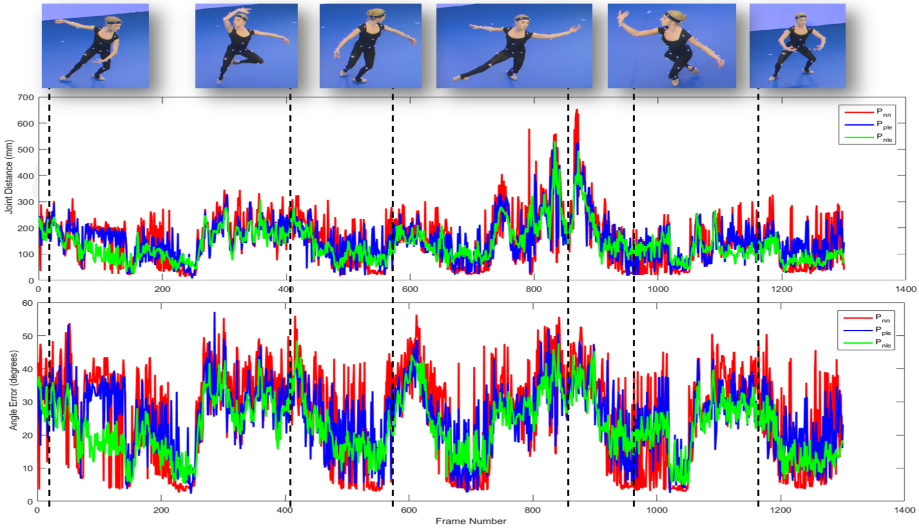

Fig. 3: Comparing total joint position and angular error of the three manifold embedding techniques over the *Ballet* dataset.

Pose Estimation Experiments To evaluate the pose estimation accuracy, we used a hybrid dataset *Ballet* comprising five MVV sequences, totalling 9434 frames, each of which were accompanied by ground-truth measurement of 21 skeletal joint angles, produced by a professional motion capture engineer using a Vicon motion capture system. Fig. 3 (top row) illustrates sample frames. We applied the optimal descriptor (CNN+DA2) learned on the *ROM* dataset to this dataset to extract pose descriptors.

The three manifold embeddings of Sec. 3.2 were learned using 4 of the 5 MVV sequences in *Ballet* with the remaining sequence used for testing. Table 2 shows the results of the three approaches, for different descriptor dimensionality and with or without the dynamic threshold (**DynThrs**). Two metrics are used to evaluate performance, the average angular error between the estimated quaternion angle of each joint and its groundtruth, and the average positional error, via simple euclidean distance. Dynamic thresholding the representation uniformly reduces error both in terms of joint angle and location. Without appropriate data scaling via this method, the log-polar representation poorly encodes extremities of the performer containing expressive arm and leg movements.

Although a general trend rewarding higher dimensionality is observed in P_{nn} and P_{ple}, this is not true for the non-linear embedding via GP-LVM where the best performing configuration is a dimensionality of 1k (with dynamic thresholding). Fig. 3 quantifies per frame error for each of the three manifold techniques over this best-performing descriptor. Not only does the non-linear embedding result in a lower average error under both metrics, but the graph also reflects a more temporally coherent estimate. Fig. 4 provides qualitative comparisons via representative examples of pose estimates from each of the three approaches.

Table 2: Estimation error: Avg. angular error (deg); Avg. location error (mm)

Descriptor	P_{nn}		P_{ple}		P_{nle}	
	Angle Err	Joint Err	Angle Err	Joint Err	Angle Err	Joint Err
1K	26.1	154.5	25.3	152.6	22.8	139.2
4K	23.2	136.1	22.9	140.1	21.4	132.2
1K+DynThrs	22.5	128.1	21.2	124.5	**20.3**	**122.9**
4K+DynThrs	21.7	124.9	21.5	127.5	20.8	129.8

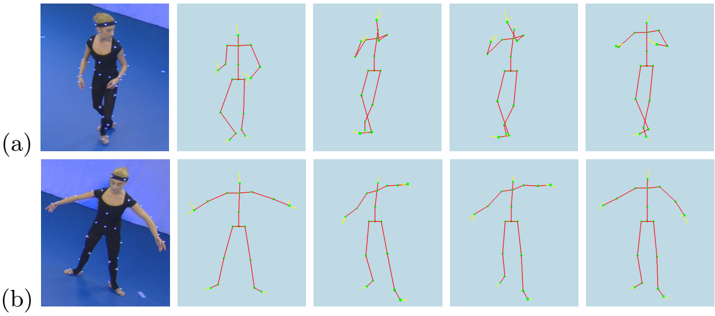

Fig. 4: Representative examples of source data and corresponding pose estimations. From left to right: Source MVV frame; Ground truth (Vicon); Nearest-neighbour (P_{nn}); Piecewise linear embedding (P_{ple}); Non-linear GP-LVM embedding (P_{nle}). Frames sampled at a) 582 and b) 660 from *Ballet*.

5 Conclusion

We presented a technique for markerless performance capture from MVV. A CNN was trained to discriminate between a broad range of motions using volumetric data derived from the MVV sequence. We reported experiments indicating that the pose descriptors learned by the CNN performed strongly in both pose classification and at pose estimation (regression). Furthermore, we demonstrated that the robustness of pose estimation was greatly improved by modeling the manifold of likely poses in the CNN descriptor space via a GP-LVM.

Future work will consider the fusion of additional forms of sensor data, such as wearable inertial sensors, to further enhance accuracy. On-axis rotation of limbs (e. g. wrist) are poorly captured by a silhouette-based representation (visual hull) whereas such movements may be captured with ease using inertial sensors. Although a prior toward valid poses is implicit within the learned manifold, explicit kinematic constraints might also be built in to further refine accuracy. The use of synthetic 3D animation data to boost the training set could also prove valuable. Nevertheless, we believe such improvements are unnecessary to demonstrate the potential for deep learning in pose estimation from MVV.

Acknowledgements

The work was supported by the REFRAME project, InnovateUK grant agreement 101854. The Ballet dataset is courtesy of the EU FP7 RE@CT project.

References

1. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: Proc. Computer Vision and Pattern Recognition. Volume 2. (2003) 459–466
2. Aggarwal, A., Biswas, S., Singh, S., Sural, S., Majumdar, A.: Object tracking using background subtraction and motion estimation in MPEG videos. In: Proc. Asian Conf. on Computer Vision (ACCV). Volume 3852., Springer LNCS (2006)
3. Viola, P., Jones, M.: Robust real-time object detection. Intl. Journal of Computer Vision **2**(57) (2004) 137–154
4. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: Proc. British Machine Vision Conf. (BMVC). (2009)
5. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proc. NIPS. (2012)
6. Toshev, A., Szegedy, C.: Deep pose: Human pose estimation via deep neural networks. In: Proc. CVPR. (2014)
7. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: A deep representation for volumetric shapes. In: Proc. CVPR. (2015)
8. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: Proc. CVPR. (2015)
9. Park, D., Ramanan, D.: Articulated pose estimation with tiny synthetic videos. In: Proc. CHA-LEARN Workshop on Looking at People. (2015)
10. Wei, L., Huang, Q., Ceylan, D., Vouga, E., Li, H.: Dense human body correspondences using convolutional networks. CoRR **abs/1511.05904** (2015)
11. Huang, P., Tejera, M., Collomosse, J., Hilton, A.: Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. ACM Transactions on Graphics (ToG) (2015)
12. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 3810–3818
13. Grauman, K., Shakhnarovich, G., Darrell, T.: A bayesian approach to image-based visual hull reconstruction. In: Proc. CVPR. (2003)
14. Huang, P., Hilton, A., Starck, J.: Shape similarity for 3d video sequences of people. Intl. Journal of Computer Vision (2010)
15. Makadia, A., Daniilidis, K.: Spherical correlation of visual representations for 3d model retrieval. Intl. Journal of Computer Vision (2009)
16. Rasmussen, C.E., Williams, C.: Gaussian Processes for Machine Learning. MIT Press (2006)
17. Lawrence, N.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. Journal of Machine Learning Research **6** (2005) 1783–1817