2016 IEEE 16th International Conference on Data Mining

A Rotation Invariant Latent Factor Model for Moveme Discovery from Static Poses

Matteo Ruggero Ronchi, Joon Sik Kim and Yisong Yue
California Institute of Technology, Pasadena, CA, USA
{mronchi, jkim5, yyue}@caltech.edu

***Abstract*—We tackle the problem of learning a rotation invariant latent factor model when the training data is comprised of lower-dimensional projections of the original feature space. The main goal is the discovery of a set of 3-D bases poses that can characterize the manifold of primitive human motions, or movemes, from a training set of 2-D projected poses obtained from still images taken at various camera angles. The proposed technique for basis discovery is data-driven rather than hand-designed. The learned representation is rotation invariant, and can reconstruct any training instance from multiple viewing angles. We apply our method to modeling human poses in sports (via the Leeds Sports Dataset), and demonstrate the effectiveness of the learned bases in a range of applications such as activity classification, inference of dynamics from a single frame, and synthetic representation of movements.**[1]

I. Introduction

What are the typical ranges of motion for human arms? What types of leg movements tend to correlate with specific shoulder positions? How can we expect the arms to move given the current body pose? Our goal is to address these questions by recovering a set of "bases poses" that summarizes the variability of movements in a given collection of static poses captured from images at various viewing angles.

One of the main difficulties of studying human movement is that it is a priori unrestricted, except for physically imposed joint angle limits which have been studied in medical text books, typically in limited configurations [2]. Furthermore, human movement may be distinguished into movemes, actions, and activities [3] depending on structure, complexity, and duration. Movemes refer to the simplest meaningful pattern of motion: a short, target-oriented trajectory, that cannot be further decomposed ("grasp", "step"). A complex gesture should be composed out of simple movemes: we define an action as a predefined and ordered sequence of movemes ("drink from a glass"), and an activity as a possibly stochastic combination of actions taking place over a stretch of time, ("dine", "read"). Extensive studies have been carried out on human action and activity recognition, however little attention has been paid to movemes since human behaviour is difficult to analyze at such a fine scale of dynamics.

In this paper, our primary goal is to learn a bases space to smoothly capture movemes from a collection of two dimensional images, although our learned representation can also aid in higher level reasoning.

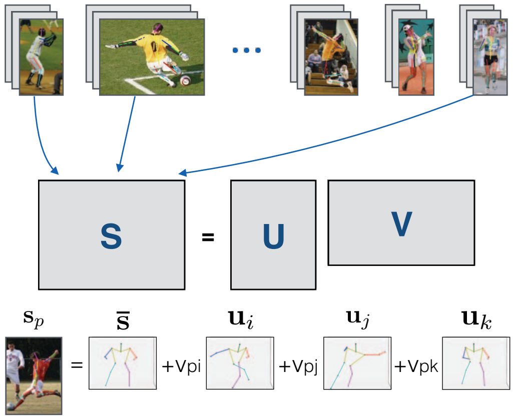

Fig. 1. **Rotation Invariant Moveme Discovery**. Given a collection of static joint locations from images taken at any angle of view we learn a factorization into a bases poses matrix **U** and a coefficient matrix **V**. The learned bases poses **U** are *rotation-invariant* and can be globally applied across a range of viewing angles. A sparse linear combination of the learned bases accurately reconstructs the pose of a human involved in an action at any angle of view.

Static poses extracted from two-dimensional images are the most abundant source of pose information. Thus, finding a bases representation using such training data can prove extremely valuable, given the number of image datasets that are currently being collected with a focus on common activities [4], [5]. However, such images are typically taken from a wide range of viewing angles, and can yield only two-dimensional projections of the underlying three-dimensional pose. Any method that does not directly address these issues will learn a naive representation that fails to provide global three-dimensional bases poses that can capture pose changes due to the true human motion while disregarding those due to a change of the angle of view. We propose a simple but effective rotation invariant latent factor model that can recover a set of three-dimensional bases poses from a training set of two-dimensional projections. Our approach is distinguished from previous approaches by directly incorporating geometric operations in an integrated way, and yields interpretable three dimensional bases poses that can be easily visualized as well as manipulated to express a natural range of human poses (Fig. 1). We applied our approach to model poses in sports activities, since they have characteristic motions and share trajectories of parts of the body, which allows to more easily interpret and evaluate qualitatively the learned movemes.

[1] A longer version of this paper, with an extdended *Models* Section, additional *Experiments* and more detailed interpretations, is published at [1].

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.167

A compact representation such as the proposed one can be used in addition to the feature representation of state of the art methods for activity recognition, favoring both performance and interpretability of results, to predict future dynamics of an action, or morph a pose into another from one frame, by observing the movemes which better describe the pose.

In summary, the main contributions of our paper are:

1. An **unsupervised** method for learning a **rotation-invariant** set of bases poses. We propose a solution to the intrinsically ill-posed problem of going from static poses to movements, without being affected by the angle of view.

2. A demonstration of how the learned bases poses can be used in various applications, including manifold traversal, discriminative classification, and synthesis of movements.

II. Related Work

Human Pose Analysis: There are two main directions of research for human pose analysis. The first one is estimation: given a picture containing a person, the goal is to predict the location of a predefined set of joints of its body, either in the 2-D image [6] or in the 3-D space [7]. Methods for 3-D pose reconstruction build upon the results of 2-D pose estimators by using physical constraints and domain knowledge to infer the true underlying human pose, and are more of interest in this study since they implicitly learn an overcomplete basis for modeling human movement. The second line of investigation uses pose as a form of contextual information that can be combined with objects' category and location in an image to obtain higher performance for activity recognition through a joint learning procedure [8]. From the perspective of pose analysis, the goal of this work is to learn a semantically meaningful representation of pose that can model human motion, be independent of the application, and that can be incorporated with other representations. We are the first to propose a representation that directly treats the problem of rotation-invariance and can be learned only from static poses, which is the most abundant form of data.

Latent Factor Models: We build upon a long line of research in latent factor models. Applications include modeling variations of faces [9], document and text analysis, and behavior patterns in sports [10]. In this regard, our work introduces an approach for learning a latent factor model in a high-dimensional space, when the observed training data are lower-dimensional projections. Our method can be viewed as a form of representation learning, which includes methods such as deep neural networks [11]. One of the benefits of representation learning is the ability to smoothly traverse the representation space, which in our setting translates to learning movemes as transitions between poses.

III. Models

We characterize the challenge of learning only from lower-dimensional projections of the underlying feature space, and present a rotation-invariant latent factor model for dealing with such training data.

A. Basic Notation and Framework

In this paper, we focus on learning from two-dimensional projections of three-dimensional human poses, however, it is straightforward to generalize to other settings. We are given a training set $S = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{n}$ of n two-dimensional poses, where x and y correspond to the image coordinates of the pose joints from the observed viewing angle. Let $\mathbf{S} \in \Re^{2d \times n}$ denote the dataset matrix, where $2d$ is the dimensionality of the projected space (twice the number of joints d for two-dimensional projections). Our goal is to learn a bases poses matrix $\mathbf{U} \in \Re^{2d \times k}$ composed of k latent factors, and a coefficient matrix $\mathbf{V} \in \Re^{k \times n}$, so that every training example can be represented as a linear combination:

$$\mathbf{s}_j = \mathbf{U} \cdot \mathbf{v_j} + \bar{\mathbf{s}}, \tag{1}$$

where $\bar{\mathbf{s}}$ denotes the "mean" pose. Of course, (1) does not deal with rotation invariance and treats the x and y coordinates as having the same semantics across training examples. We present in Sec. III-C a rotation-invariant latent factor model to address this issue and recover a three-dimensional $\mathbf{U} \in \Re^{3d \times k}$.

B. Baselines

To the best of our knowledge, no existing approach tackles the problem of learning a rotation-invariant bases for modeling human movement. Previous work is focused on either learning bases poses only from frontal viewing angles or by extensive manual crafting of a predefined set of poses [7], [12]. As such, we develop our approach by building upon classical baselines such as the SVD.

Singular Value Decomposition: The example in (1) is the most basic form of a latent factor model, and the solution can be recovered via SVD. The basis matrix $\mathbf{U}$ and the coefficient matrix $\mathbf{V}$ respectively correspond (up to a scaling) to the left and right singular vectors of the mean-centered data matrix $\mathbf{S}_c = (\mathbf{S} - \bar{\mathbf{s}})$. However, naively applying the SVD to our setting will result in the bases matrix $\mathbf{U}$ conflating viewing angle rotations with true pose deformations.

Clustered Singular Value Decomposition: If the viewing angle of the training data is available, or a quantized approximation of it, then the basic latent factor model (1) can be instantiated separately for different viewing angles, via:

$$\mathbf{s}_j = \mathbf{U}(a_j) \cdot \mathbf{v_j} + \bar{\mathbf{s}}(a_j), \tag{2}$$

where a_j denotes the viewing angle cluster that example j belongs to. In other words, given p clusters, we learn p separate latent factor models, one per cluster. Intuitively, we expect this method to suffer less conflation between changes in pose due to a viewing angle rotation and true pose deformation, and the more clusters, the less susceptible. The main drawbacks are that: (i) the learned bases representation is not global, and will not be consistent across the clusters since they are learned independently, and (ii) the amount of training data per model is reduced, which can yield a worse representation.

C. Rotation-Invariant Latent Factor Model

Our goal is to develop a latent factor model that can learn a global representation of bases poses across different angles. For simplicity, we restrict ourselves to settings where there are only differences in the pan angle, and assume no variation in the tilt angle (i.e., all horizontal views). To that end, we propose both a 2-D and a 3-D model which can be used depending on the quality and quantity of additional information available at training time. For some applications it may suffice to use the 2-D model, however the 3-D model is generally better able to intrinsically capture rotation-invariance.

We first motivate some of the desirable properties:

- **Unsupervised** – the bases discovery should not be limited to or dependent on images of specific classes of actions.
- **Rotation Invariant** – the learned bases should be composed of movements from a given canonical view (e.g., frontal) and be able to reconstruct poses oriented at any angle. The exact same pose may look different when observed from different camera angles; as such, it is important to disambiguate pose from viewing angle.
- **Sparse** – to encourage interpretability, the learned bases should be sparsely activated for any training instance.
- **Complementary** – our method should be easy to integrate with other modeling approaches, and thus should implement an orthogonal extension of the basic latent factor modeling framework.

General Framework: Our general framework aims to learn a latent factor matrix $\mathbf{U}$, containing the bases poses instantiated globally across all the training data; a coefficient matrix $\mathbf{V}$, whose columns correspond to the weights given to the bases poses to reconstruct all training instance; and a vector θ, containing the angle of view of each training pose.

We can thus model every training example as:

$$\mathbf{s}_j = f(\bar{\mathbf{s}} + \mathbf{U} \cdot \mathbf{v}_j, \theta_j), \tag{3}$$

where $f(\cdot,\cdot)$ is a projection operator of the higher-dimensional model into the two-dimensional space. We train our model via:

$$\mathbf{U}, \mathbf{V}, \theta = \underset{\mathbf{U},\mathbf{V},\theta}{\arg\min}\, \mathcal{L}(\mathbf{U}, \mathbf{V}, \theta), \tag{4}$$

$$\mathcal{L}(\mathbf{U}, \mathbf{V}, \theta) = \mathcal{E}(\mathbf{U}, \mathbf{V}, \theta) + \Omega(\mathbf{U}, \mathbf{V}, \theta), \tag{5}$$

$$\mathcal{E}(\mathbf{U}, \mathbf{V}, \theta) = \sum_j \Big(\mathbf{s}_j - f(\bar{\mathbf{s}} + \mathbf{U} \cdot \mathbf{v}_j, \theta_j)\Big)^2, \tag{6}$$

where $\mathcal{E}$ is the squared reconstruction error over the training instances, and Ω is a model-specific regularizer. The projection operator f and the regularizer Ω are specified separately for the 2-D and 3-D approach. This optimization problem is non-convex, and requires a reasonable initialization in order to converge to a good local optimum.

1) 2-D approach: The 2-D approach, uses the same approach as the clustered SVD baseline and, given a set of p angle clusters, instantiates the projection operator as:

$$f(\bar{\mathbf{s}} + \mathbf{U} \cdot \mathbf{v}_j, \theta_j) = \bar{\mathbf{s}}(a_j) + \mathbf{U}(a_j) \cdot \mathbf{v}_j, \tag{7}$$

a_j denotes the cluster that θ_j belongs to, and a separate rank-k $\mathbf{U}$ is learned for each viewing angle cluster. At this point, (7) looks identical to (2). However, we encourage global consistency via the regularization terms:

$$\Omega(\mathbf{U}, \mathbf{V}, \theta) = R_{reg}(\mathbf{U}, \mathbf{V}, \theta) + R_{spat}(\mathbf{U}, \mathbf{V}, \theta). \tag{8}$$

The first term in (8) is a standard regularizer:

$$R_{reg}(\mathbf{U}, \mathbf{V}, \theta) = \sum_{a=1}^{p} \Big[\lambda_U \|\mathbf{U}(a)\|_F^2 + \lambda_V \|\mathbf{V}(a)\|_1\Big]. \tag{9}$$

We wish to have sparse activations so we regularize $\mathbf{V}$ using L1 norm. Depending on the application, Sec. IV-B, we sometime enforce that $\mathbf{V}$ be non-negative for added interpretability.

The second term in (8) is the spatial regularizer that encourages (or in some cases enforces) consistency across the per-cluster models:

$$R_{spat}(\mathbf{U}, \mathbf{V}, \theta) = \lambda_{spat} \sum_{a,a'} \kappa_{a,a'} \|\mathbf{U}^{(x)}(a) - \mathbf{U}^{(x)}(a')\|_F^2 \tag{10}$$

$$+ \sum_{a,a'} \mathbf{1}\Big(\mathbf{U}^{(y)}(a), \mathbf{U}^{(y)}(a')\Big), \tag{11}$$

$\mathbf{U}^{(x)}$ and $\mathbf{U}^{(y)}$ represent the x and y coordinate portion of the bases poses: e.g. $\mathbf{U}^{(x)} = [\mathbf{U}_{i,-}]$, $i \in X$, where X is the set of indices corresponding to x coordinates in the pose representation. Since we are only modeling variations in the pan angle, the x coordinates can vary across different viewing angles, while the y coordinates should remain constant. As such, the first term in R_{spat}, (10), corresponds to encouraging the $\mathbf{U}^{(x)}(a)$ and $\mathbf{U}^{(x)}(a')$ of different clusters to be similar to each other (with $\kappa_{a,a'}$ controlling the degree of similarity), and the second term, (11), is a $\{0, \infty\}$ indicator function that takes value 0 if the two arguments are identical, and value ∞ if they are not (i.e., it is a hard constraint).

In summary, the spatial regularization term is the main difference between the 2-D latent factor model and the clustered SVD baseline. Global consistency of the per-cluster models is obtained by encouraging similar values in the x coordinates, and enforcing identical y coordinates. In a sense, one can view spatial regularization as a form of multi-task regularization, which enables sharing statistical strength across the clusters. The main limitation of the 2-D model is that the spatial regularization does not incorporate more sophisticated geometric constraints, so the notion of consistency achieved may not align with the true underlying three-dimensional data.

2) 3-D approach: The 3-D model directly learns a three-dimensional representation of the underlying pose space, through a single and global $\mathbf{U} \in \mathfrak{R}^{3d\times k}$ that is inherently three-dimensional, and captures k bases poses.

The projection operator is now defined as:

$$f(\bar{\mathbf{s}} + \mathbf{U} \cdot \mathbf{v}_j, \theta_j) = \Big[\mathbf{Q}(\theta_j)\Big(\bar{\mathbf{s}} + \mathbf{U} \cdot \mathbf{v}_j\Big)\Big]^{(x,y)}, \tag{12}$$

where $\mathbf{Q}(\cdot)$ is the 3-D rotation matrix around the vertical axis:

$$\mathbf{Q}(\theta_j) = \begin{bmatrix} \cos(\theta_j) & 0 & \sin(\theta_j) \\ 0 & 1 & 0 \\ -\sin(\theta_j) & 0 & \cos(\theta_j) \end{bmatrix}, \tag{13}$$

and the superscript $^{(x,y)}$ denotes the projection from the 3-D space of $\mathbf{U}$ to the 2-D space of the dataset annotations, obtained by indexing only the x and y coordinates (the underlying model provides x, y, and z coordinates). The projection operator in (12) allows to compute the two-dimensional projection of any underlying three-dimensional pose at any viewing angle θ_j using standard geometric rules. Spatial regularization is no longer needed, because the rotation operator $\mathbf{Q}$ relates all the viewing angles to a common model, thus the regularizer assumes the standard form:

$$\Omega(\mathbf{U}, \mathbf{V}, \theta) = \lambda_U \|\mathbf{U}\|_F^2 + \lambda_V \|\mathbf{V}\|_1. \tag{14}$$

In summary, the 3-D latent factor model improves upon the 2-D version by learning a global representation that is intrinsically three-dimensional and integrates domain knowledge of how the viewing angle affects pose via geometric projection rules. This results in a more robust method, that does not learn a separate model per viewing angle or rely on the spatial regularization to obtain consistency. The main drawback is that a more complex initialization will be required.

D. Training Details

Initialization: Our approaches require an initial guess of the viewing angle for each training instance, and of the bases poses $\mathbf{U}$. For angle initialization, we show in our experiments that we only need a fairly coarse prediction of the viewing angle (e.g., into quadrants). The 2-D latent factor model bases poses $\mathbf{U}$ are initialized uniformly between -1 and 1, while for the 3-D model we used [7] to recover a 3-D pose estimate of every instance, and set the initial bases poses $\mathbf{U}$ through SVD.

Optimization: We optimize Eq. (4) using alternating stochastic gradient descent, divided in two phases:

- Representation Update: we employ standard stochastic gradient descent to update $\mathbf{U}$ and $\mathbf{V}$ while keeping θ fixed. For the 3-D model, this involves computing how the training data (which are two-dimensional projections) induce a gradient on $\mathbf{U}$ and $\mathbf{V}$ through the rotation $\mathbf{Q}$. Because we employ an L1 regularization penalty, we use the standard soft-thresholding technique [13].
- Angle Update: Once the optimal $\mathbf{U}$ and $\mathbf{V}$ are fixed, we employ standard stochastic gradient descent to update θ.

Convergence and Learning Rates: Three training epochs of 10,000 iterations are usually sufficient for convergence to a good local minimum. Typical values of the learning rate are 1×10^{-4} for $\mathbf{U}$ and $\mathbf{V}$ and 1×10^{-6} for θ. We use a smaller step size in the update of θ, since the curvature of the objective function (4) w.r.t. θ is higher than w.r.t. $\mathbf{U}$ and $\mathbf{V}$.

IV. Experiments

A. Dataset and Additional Annotations

We use the Leeds Sports Dataset (LSP) [14] for our experiments. LSP is composed of 2,000 images containing a single person performing one of eight sports (Athletics, Badminton, Baseball, Gymnastics, Parkour, Soccer, Tennis, Volleyball) annotated with the x, y location and a visibility flag for 14 joints of the human body. We discard "Gymnastics" and "Parkour" from our analysis because they have few examples and the class poses do not vary exclusively along the pan angle (but appear in very unconventional views, i.e. upside-down and horizontal), violating the assumption in Sec. III-C.

We collected high-quality viewing angle annotations for each pose in LSP. Altough these annotations are not necessary for training, we use them to demonstrate the robustness of our model to poor angle initialization, and that it can in fact recover the ground truth value, see Sec. IV-B4. Three annotators evaluated each image and were instructed to provide the direction at which the torso was facing.

B. Empirical Results

We evaluate the learned representation in terms of (i) the performance on supervised learning tasks such as activity classification; (ii) whether it captures enough semantics for meaningful manifold traversal; (iii) the robustness to initialization and the generalization error. Collectively, results suggest that our approach is effective at capturing rotation invariant semantics of the underlying data. A more detailed description and interpretation of the experiments can be found in [1], along with a t-SNE visualization of the manifold of human motion learned with the "lfa3d" method, showing that similar movements are mapped to consistent clusters in nearby positions, regardless of the viewing angle.

1) Activity Recognition: The matrix $\mathbf{V}$ describes each pose in the dataset as a linear combination of the learned latent factors, Sec. III-A. Thus, $\mathbf{v}_j$ can be interpreted as a semantically more meaningful feature representation for j-th data point. For instance, if a lower body basis pose (e.g. Fig. 4) has a high weight, the reconstructed pose is very likely to represent a movement from an activity related to running or kicking.

To test the effectiveness of the learned representation, we used the coefficients in $\mathbf{V}$ as input features for classifying the sport categories in LSP. Fig. 2 shows the results obtained from five fold cross validation. The proposed "lfa3d" model outperforms all other methods by an average accuracy of about 11%. Note that only the weights of the latent factors reconstructing a pose are being used to discriminate between the activities, without the aid of visual cues from the image. It is thus surprising that "lfa3d" achieves an average 39% accuracy, when a random guess would merely give 16.7%.

2) Action Dynamics Inference: If the latent factor model captures the semantics of the data, then poses that occur in chronological order within a given action should lie in a monotonic sequence within the learned space. A quantitative

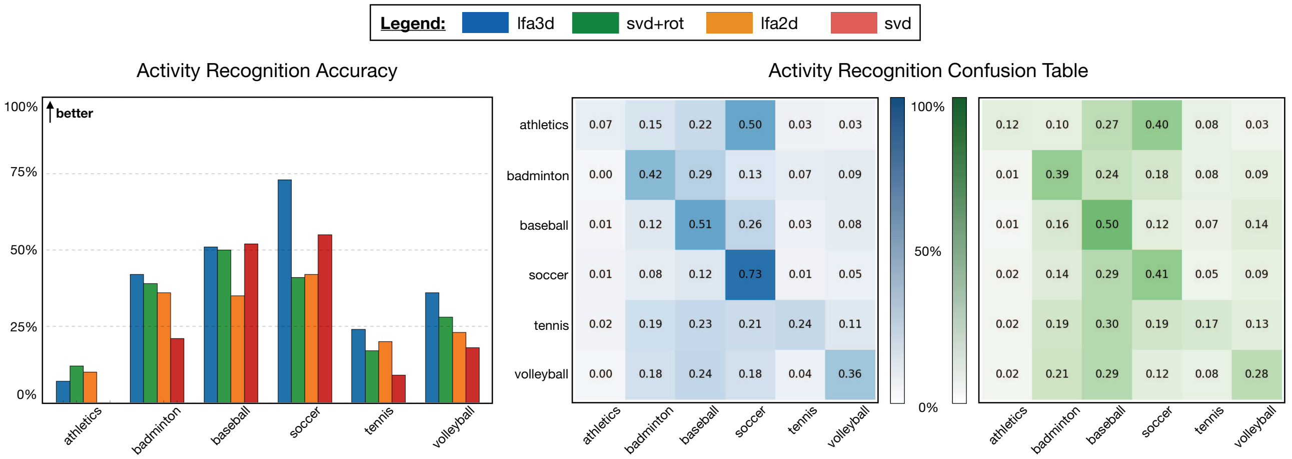

Fig. 2. **Activity Recognition Performance**. (Left) The activity classification accuracy across the sports in LSP for the following methods: "svd" – baseline, "svd+rot" – clustered version of the baseline, "lfa2d" – 2-D latent factor model with spatial regularization, "lfa3d" – full 3-D latent factor model. (Right) The confusion tables for the best two performing methods, "lfa3d" and "svd+rot". Full details in Sec. IV-B1.

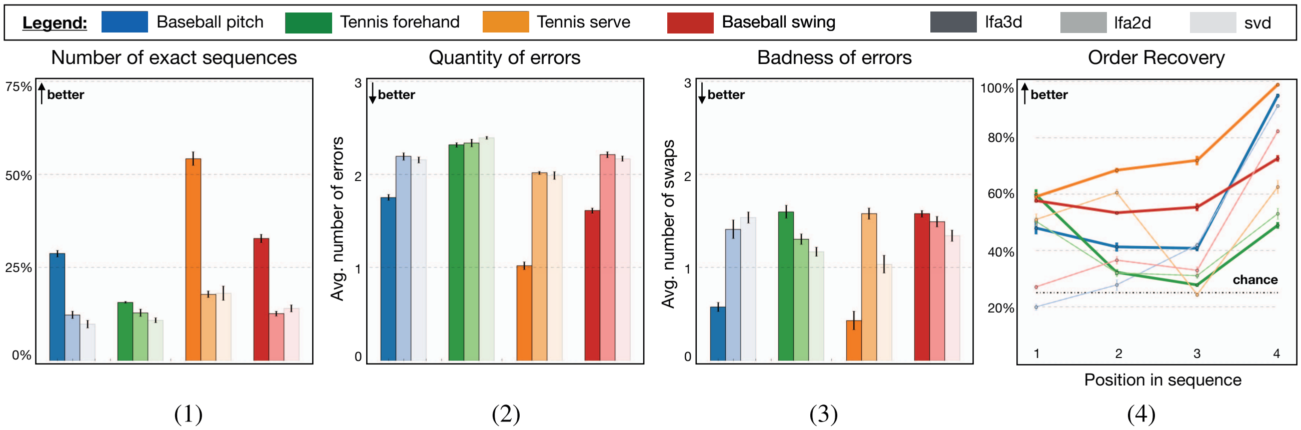

Fig. 3. **Action Dynamics Inference Performance**. We compare the methods "svd", "lfa2d", and "lfa3d" in the task of reordering shuffled sequences of images sampled from four different sport actions. The color scheme represents actions, the methods are plotted with a different transparency value. The performance is described in terms of: (1) number of sequences exactly reordered; (2) average number of errors contained in a sequence; (3) average number of swaps needed to obtain the correct sequence; (4) accuracy per position in the sequence – shown only for the best two methods ("lfa3d" - dark marker, "lfa2d" - light marker). Full details in Sec. IV-B2.

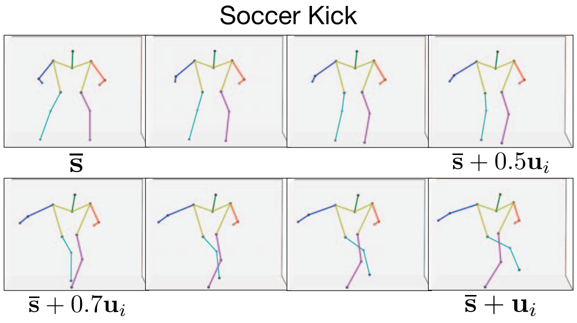
Soccer Kick
$\bar{\mathbf{s}}$
$\bar{\mathbf{s}} + 0.5\mathbf{u}_i$
$\bar{\mathbf{s}} + 0.7\mathbf{u}_i$
$\bar{\mathbf{s}} + \mathbf{u}_i$

Fig. 4. **Learned Movemes Visualization**. A latent factor, encoding a moveme from the learned bases poses matrix $\mathbf{U}$, easily interpretable as a "soccer kick". The sequence is obtained by adding an increasing fraction of the basis (30% to 50% on the first row, 70% to 100% on the second row) to the mean pose of the dataset. Full details in Sec. IV-B3.

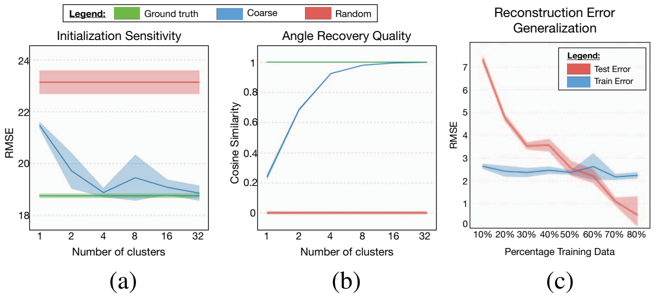

Fig. 5. **Initialization and Generalization Analysis**. Sensitivity to initialization of (a) the RMSE, and (b) the Cosine Similarity between the learned and ground truth angles. A coarse initialization, within the correct quadrant, yields performances similar to ground truth. (c) Reconstruction error for poses not in the training set. Full details in Sec. IV-B4, IV-B5.

measure of the quality of the representation can be obtained by observing how well the order of poses in a same action is preserved. To find the sequence in which a set of poses lies in the manifold, we compute a linear traversal through the representation space. This ordering should hold regardless of the angle of view of the input instances. In this experiment, Fig. 3, we shuffled 1,000 sequences of four images for four sport actions, and verified how precisely could the underlying chronological sequence be recovered; example sequences can be found in [1]. The "lfa3d" model has significantly better outcomes: it correctly reorders more than twice the sequences overall (1,314 against 555 of "lfa2d") averages 1.6 errors, and is the only algorithm with an average number of swaps <1.

3) Moveme Visualization: The "lfa3d" method can be used to recover and synthesize realistic human motions from static joint locations in images. The underlying idea, is that models of human motion can be successfully learned from observations of poses of people performing various actions, as opposed to deriving mathematical principles which define control laws (e.g. inverse kinematics). The most significant movemes contained in the training set are captured by the bases poses matrix $\mathbf{U}$ and encoded in the form of a displacement from the mean pose. Each column of $\mathbf{U}$ corresponds to a latent factor that describes some of the movement variability present in the data. Fig. 4 reports the motion of a learned latent factor (additional motions are presented in [1]) by adding an increasing portion of the learned moveme to the mean pose of the data (top-left). We verify empirically that two parameters mainly affect the correspondence between an action and a latent factor (moveme purity): the number of latent factors, and the constraints put on the coefficients of $\mathbf{V}$.

4) Angle Recovery: The "lfa3d" method learns a rotation invariant representation by treating the angle of view of each pose as a variable which is optimized through SGD (Sec. III-C2), and requires an initial guess for each training instance. We investigate how sensitive the model is to initialization, and how close the recovered angle of view is to the ground truth. Fig. 5 shows the Root Mean Squared Error (RMSE) and cosine similarity with ground truth, for three initialization methods: (1) "random", between 0 and 2π; (2) "coarse", coarsening into discrete buckets (e.g., 4 clusters indicates that we only know the viewing angle quadrant during initialization); and (3) "ground-truth". As the number of clusters increases, both evaluation metrics improve significantly for "coarse" initialization. These results suggest that using very simple heuristics to predict the viewing angle quadrant of a pose is sufficient to obtain optimal performance.

5) Generalization Behaviour: To verify the generalization quality of the learned bases poses we trained the "lfa3d" model on a subset of the dataset and measured the RMSE on the remaining part, for an increasingly larger portion of the data. As reported in Fig. 5, the RMSE over the training set is approximately constant, while the test set RMSE decreases significantly. This indicates that the learned latent factors can successfully reconstruct poses of unseen data, and is not tied uniquely to the specific image collection it was learned from.

V. Conclusion and Future Directions

We introduced a model for learning the primitive movements underlying human actions (movemes) from a set of static 2-D poses from images taken at various angles of view. The set of bases poses is rotation-invariant and learned through a latent matrix factorization that accounts for geometric properties inherent to viewing angle variability. The approach can be trained efficiently, requires modest effort to identify a reasonable initialization, and generalizes well on unseen data.

We investigated the practical use of the learned representation for applications such as activity recognition and inference of action dynamics, observing significantly better performance compared to conventional baselines that do not account for variability of viewing angles. We used the bases poses for synthetic generation of movements, and explored how specific poses are mapped into the manifold of human motion.

One desirable property of our algorithm is that it is complementary to existing latent factor, pose estimation and feature extraction approaches, and may be used in combination with them to yield a better overall rotation-invariant representation.

Possible interesting future direction of investigation would be to: use the proposed model in a semi-supervised setting where there is some availability of true 3-D data along with a large collection of 2-D joint locations; learn to morph actions and synthesize *unseen* actions from the set of extracted movemes; infer the location of occluded or missing joints based on the position of the visible ones.

References

[1] M. R. Ronchi, J. S. Kim, and Y. Yue, "A Rotation Invariant Latent Factor Model for Moveme Discovery from Static Poses," *ArXiv e-prints*, 2016.

[2] H. Hatze, "A three-dimensional multivariate model of passive human joint torques and articular boundaries," *Clinical Biomechanics*, 1997.

[3] D. J. Anderson and P. Perona, "Toward a science of computational ethology," *Neuron*, vol. 84, no. 1, pp. 18–31, 2014.

[4] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] M. R. Ronchi and P. Perona, "Describing common human visual actions in images," in *Proceedings of the British Machine Vision Conference (BMVC 2015)*. BMVA Press, September 2015, pp. 52.1–52.12.

[6] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in Neural Information Processing Systems*, 2014, pp. 1736–1744.

[7] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1446–1455.

[8] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *ACCV 2014*. Springer, 2014.

[9] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*. IEEE, 1991, pp. 586–591.

[10] Y. Yue, P. Lucey, P. Carr, A. Bialkowski, and I. Matthews, "Learning fine-grained spatial models for dynamic sports play prediction," in *IEEE International Conference on Data Mining (ICDM)*, December 2013.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[12] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *Computer Vision–ECCV 2012*.

[13] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal*, 2009.

[14] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proceedings of the British Machine Vision Conference*, 2010, doi:10.5244/C.24.12.