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.

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 basis pose matrix $\mathbf{U}$ and a coefficient matrix $\mathbf{V}$. The learned bases poses in $\mathbf{U}$ 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, also for poses not contained in the training set.

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 summarize 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 for a limited number of configurations [1], [2]. Furthermore, human movement may be distinguished into movemes, actions, and activities [3], [4] 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, e.g. "reach", "grasp", "step", "kick". A complex gesture should be composed out of simple movemes: we define an action as a predefined and ordered sequence of movemes, such as "drink from a glass", or "open a door". An activity is a (possibly stochastic) combination of actions taking place over a stretch of time with a typical and yet variable structure, e.g. "dine", "read". Extensive studies have been carried out on human action and activity recognition [5], [6], however little attention has been paid to movemes since human behaviour is difficult to analyze
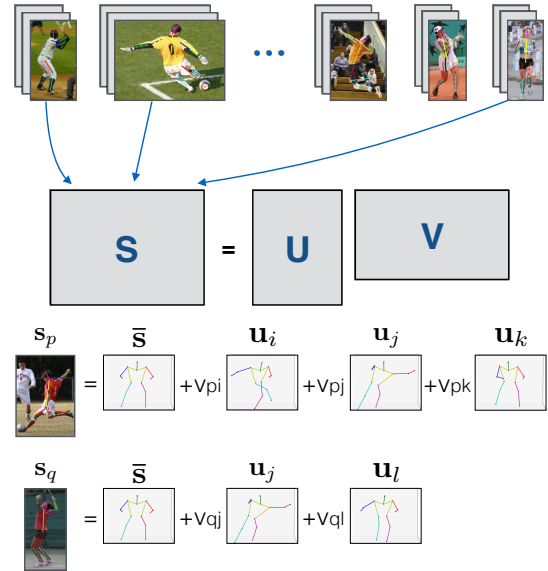
at such a fine scale of dynamics.[1] In this paper, our primary goal is to learn a basis space to smoothly capture movemes from a collection of two dimensional images, although our learned representation can also aid in higher level reasoning.

Static poses extracted from two-dimensional images are the most abundant source of pose information. Thus, finding a basis representation using such training data can prove extremely valuable, given the number of image datasets (as opposed to video or mo-cap data) that are currently being collected with a focus on common activities [8]–[10]. 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

---

[1]The extent in time and complexity of human motion is not directly observable in still images but requires videos of humans involved in activities which cannot be recorded extensively without legal or ethical issues, as opposed to fly or mouse behaviour which is very well documented [7].

address these issues will learn a naive representation that fails to provide a set of 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.

In this paper, 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 latent factor modeling 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 (as depicted in Fig. 1). We applied our approach in a case study on modeling poses that arise in sports activities, since they have very characteristic and recognizable motions and typically share trajectories of parts of the body (e.g., tennis serve and volleyball strike), which allows to more easily interpret and evaluate qualitatively the learned movemes.

Our study is not purely academic, we have four applications in mind; in this paper we carry out a quantitative and qualitative analysis for two of them, and leave the study of the latter to future work. **Activity recognition**; 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 the performance [11], and the interpretability of results. **Action dynamics inference**; modifying the weights of the learned bases poses is analogous to moving along a line in the high-dimensional space of human poses (either 2-D or 3-D). This allows to predict the future dynamic of an action [12], or morph a pose into another from a single frame, by observing the dynamics of the movemes which better describe the captured pose. **Computer graphics animation**; many animation systems are still based on *key-framing* and *in-betweening* [13]: master animators draw the key frames of a sequence to be animated and assistant animators complete the intermediate frames by inferring the movements occurring between the keys. Knowing the movemes underlying human actions would provide an automated method for interpolating between key frames, resulting in a faster and simplified animation pipeline. **3-D pose estimation**; a sparse overcomplete dictionary of human poses has been used effectively for the reconstruction of 3-D human pose given its 2-D joint locations from a single frame image [14]–[16]. Our technique would allow to identify the most suited pose bases for a given collection of images without any experimenter bias, or the need of curating the angle of view of the images in the training set.

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 [17], [18] or in the 3-D space [14]–[16]. Methods for 3-D pose reconstruction build upon the results of 2-D pose estimators by using mechanisms based on physical constraints and domain knowledge to infer the true underlying human pose observed in an image, and are more of interest in this study since they implicitly learn an overcomplete basis for modeling human movement. However, such methods typically predefine the dictionary of actions, use additional data in the training phase (such as mo-cap), and do not treat explicitly the problem of varying angles of view. In contrast, our goal is to learn a low-rank manifold of 3-D poses consistent across multiple viewing angles, given only two-dimensional data.

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 [19]–[21]. Our approach can as well be used as a feature representation for improved activity recognition.

From the perspective of pose analysis, the goal of this work is to learn a semantically meaningful representation of human pose that can model human motion. This representation should be independent of the application domain, and flexible, allowing it to be incorporated with other representations. Other people investigated this problem: it is known that dynamic information can be recovered from static images of humans engaged in activities [22], and similar representations for action recognition have been learned using video data [23], [24]. 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 we believe is important since it is the most abundant form of data.

Latent Factor Models and Representation Learning: We build upon a long line of research in latent factor models, first popularized for collaborative filtering problems in content recommendation [25]. Applications include modeling variations of faces [26], document and text analysis [27], and behavior patterns in sports [28], amongst many others. Latent factor models are variants of matrix and tensor factorization, which can easily incorporate missing values or other types of constraints. 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 is complementary to and can be integrated with other latent factor modeling approaches.

Our approach can be viewed as a form of representation learning, which includes methods such as deep neural networks and dictionary learning [29], [30]. One of the benefits of representation learning is the ability to smoothly traverse the representation space [31], which in our setting translates to learning movemes as transitions between poses.

III. MODELS

We develop our approach by building from the classical singular value decomposition. 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, see Fig. 2. 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 [14], [16]. As such, we develop our approach by building upon classical baselines such as the SVD, which we briefly describe here.

Singular Value Decomposition: The example in (1) is the most basic form of a latent factor model. When the training objective is to minimize the squared reconstruction error of the training data, then the solution can be recovered via SVD, also used for eigenfaces [26]. The bases 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:

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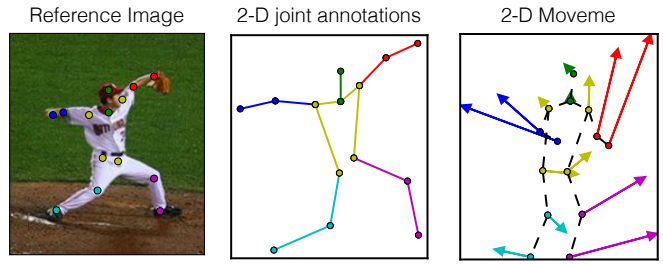

Fig. 2. **Moveme Representation**. The joint annotations from an image in LSP [32], and their displacement from the mean pose, which we use to encode movemes.

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.
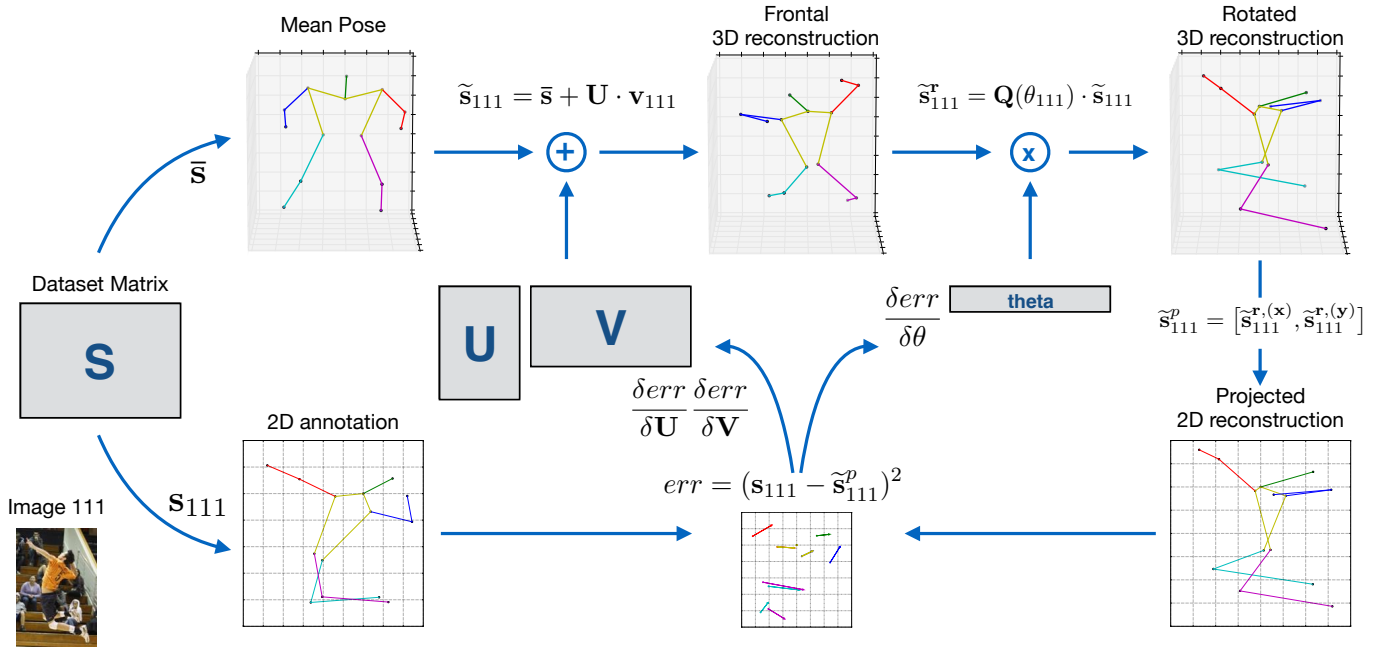
Fig. 3. **LFA3-D Method Pipeline**. Bases poses $\mathbf{U}$, coefficient matrix $\mathbf{V}$ and angles of view θ are initialized and updated through alternate stochastic gradient descent. Each iteration consists of the following steps: (1) a sparse linear combination of the current bases poses with coefficients from $\mathbf{V}$ is added to the dataset mean pose to obtain a frontal 3-D reconstruction of the true pose; (2) the 3-D reconstruction is rotated by the current estimate of the angle of view θ_j for that pose; (3) the 3-D pose is projected to the 2-D space where it is compared to the ground truth; (4) the gradient update step is computed to minimize the root mean square error *wrt.* to quantities $\mathbf{U}$, $\mathbf{V}$ and θ.

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:

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

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

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

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 between the per-cluster models 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 used to prevent overfitting:

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

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:

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

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

$\mathbf{U}^{(x)}$ 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:

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

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

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

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:

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

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 the bases poses $\mathbf{U}$. For angle initialization, we show in our experiments (Sec. IV-B4) 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 use an off-the-shelf pose estimator [16] and initialize $\mathbf{U}$ as the left singular vectors of the mean centered 3-D pose data, obtained through SVD.

Optimization: For both models, 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 [33].
- Angle Update: Once the optimal $\mathbf{U}$ and $\mathbf{V}$ are fixed, we employ standard stochastic gradient descent to update θ.

Fig. 3 provides an overview of the steps for the 3-D approach.

Convergence and Learning Rates: Three training epochs of 10000 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) [32] for our experiments. LSP is composed of 2000 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. Example images and annotations are shown in Fig. 1, 2 and 8. Sports activities are particularly well suited for this study, as they present characteristic motions that share trajectories of parts of the body, that allow investigating basis pose sharing across sports. As part of preprocessing, we normalize all the poses in the dataset by modifying each bone to have the average bone length computed over all the training instances [15]. 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. Generalizing the framework, to incorporate a wider variability of the viewing angles, is an interesting future direction.

We collected high-quality viewing angle annotations for each pose in LSP. Although 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[2]. The standard deviation in the reported angle of view averaged over the whole dataset is 12 degrees, and more than half of the images have a deviation of less than 10 degrees, showing a very high annotator agreement for the task.

---

[2]The angle annotations for LSP, annotator agreement statistics, and details about the Amazon Mechanical Turk GUI are available at the project page [34].
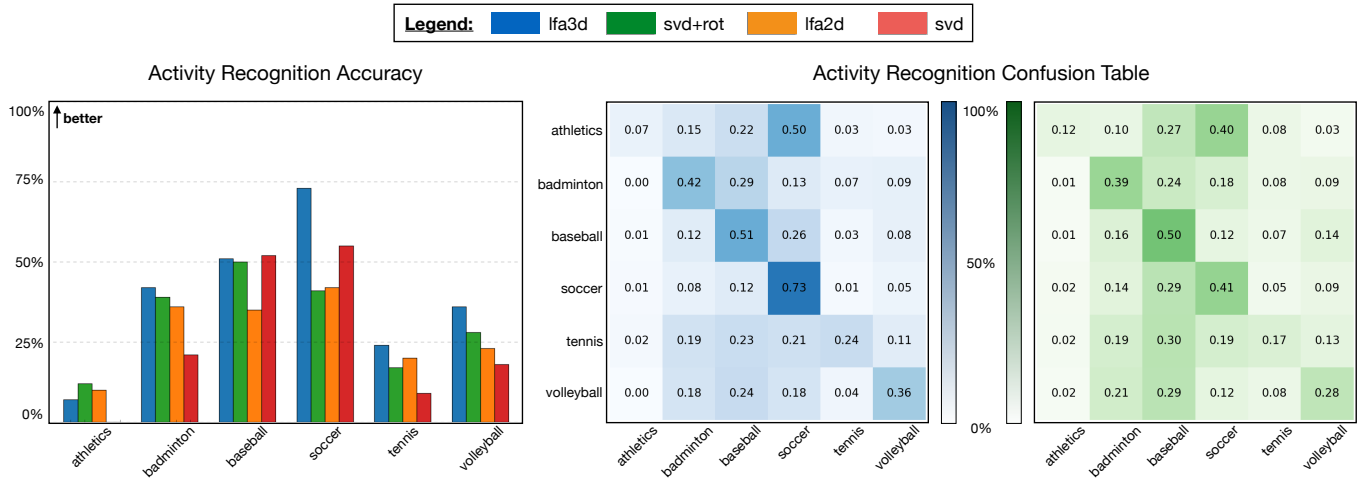
Fig. 4. **Activity Recognition Performance**. (Left) The activity classification accuracy across the sports in LSP for the following methods: "svd" – baseline, "svd+rot" – clustered version of the baseline, "lfa2d" – 2-D latent factor model with spatial regularization, "lfa3d" – full 3-D latent factor model. (Right) The confusion tables for the best two performing methods, "lfa3d" and "svd+rot". Full details in Sec. IV-B1.

B. Empirical Results

We analyze the flexibility and usefulness of the proposed model in a variety of application domains and experiments. In particular, we evaluate (i) the performance of the learned representation for supervised learning tasks such as activity classification; (ii) whether the learned representation captures enough semantics for meaningful manifold traversal and visualization; and (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.

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. 6 top row) has a high weight, the reconstructed pose is very likely to represent a movement from an activity related to running, or kicking.

A natural way to test the effectiveness of the learned representation is to use it for supervised learning tasks. To that end, we used the coefficients in $\mathbf{V}$ as input features for classifying the sport categories in LSP.

Fig. 4 shows the results obtained from five-fold cross validation. The proposed 3-D latent factor model ("lfa3d") outperforms all other methods by an average accuracy of about 11%. The 2-D model ("lfa2d") performs slightly worse than the clustered SVD baseline ("svd+rot"), but both show more than a 5% average improvement over the "svd" baseline. The two most challenging activities are "athletics", which does not posses characterizing movements; and "tennis", whose movemes are shared and thus confused with multiple other sports, "badminton" and "baseball" above all. We also report the full classification confusion tables in Fig. 4. 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%. Finally, the obtained feature representation is complementary to other representations, such as the hidden layer activations of a convolutional neural network [35], and we wish to investigate in future work the performance obtained by their combination.

2) Action Dynamics Inference & Manifold Traversal: Every pose in the training set belongs to a movement of the body corresponding to a complex trajectory in the manifold of human motion. 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 measure of the quality of the representation can be obtained by observing how well the order of poses belonging to a same action is preserved. One straightforward way to find the sequence in which a set of poses lies in the manifold, is to look at the coefficient of their projection along the "total least squares" line fit [36] of the corresponding columns in the matrix $\mathbf{V}$. In other words, we are computing a linear traversal through the representation space. Furthermore, this ordering should hold regardless of the angle of view of the input instances.

In this experiment, we shuffled 1000 sequences of four images for four sport actions ("baseball pitch", "tennis forehand", "tennis serve", "baseball swing"), and verified how precisely could the underlying chronological sequence be recovered. The analysis is repeated five times to obtain standard deviations, and performance is measured in terms of three metrics: (1) what percentage of the 1000 sequences is exactly reordered; (2) how many poses are wrongly positioned; and (3) how bad are the reordering mistakes, computed as the number of swaps necessary to correct a sequence.

Fig. 5 shows the results for the latent factor models "lfa2d", "lfa3d" and for the "svd" baseline. It is not possible to study
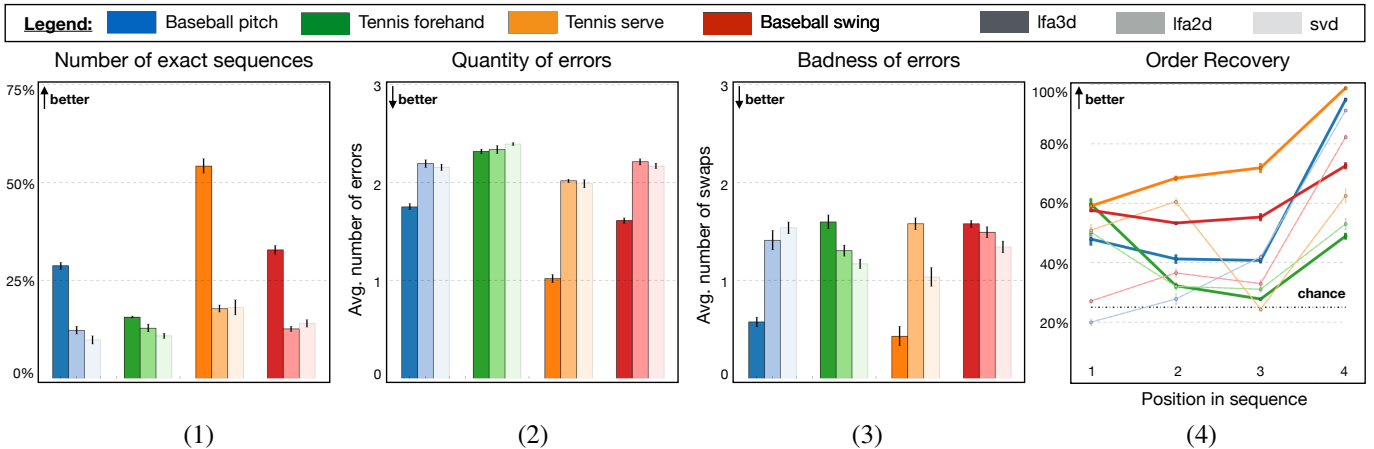
Fig. 5. **Action Dynamics Inference Performance**. We compare the methods "svd", "lfa2d", and "lfa3d" in the task of reordering shuffled sequences of images sampled from four different sport actions. The color scheme represents actions, the methods are plotted with a different transparency value. The 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). Example sequences in Tab. I. Full details in Sec. IV-B2.

the performance of the clustered baseline "svd+rot" since it does not learn a global matrix U, thus the coefficients in V are not comparable across different viewing angles.

The "lfa3d" model has significantly better outcomes compared to "lfa2d" and "svd", which perform similarly. Specifically, "lfa3d" correctly reorders more than twice the sequences overall (1314 against 555 of "lfa2d") averages 1.6 errors, and is the only algorithm to require an average number of swaps smaller than 1. Fig. 5-(4) shows the per-position accuracy.

An example sequence for "tennis serve" is shown in Tab. I. Only the "lfa3d" method recovers the order correctly; note how the images are all taken from different viewing angles.

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 U and encoded in the form of a displacement from the mean pose. Each column of U corresponds to a latent factor that describes some of the movement variability present in the data.

Fig. 6 reports the motion described by three latent factors: the rows show the pose obtained by adding an increasing portion of the learned moveme (from 30% - second column, to 100% - last column) to the mean pose of the data (first column). Two are easily interpretable, "soccer kick" and "tennis forehand", while one is not as well defined, "volleyball strike / tennis serve". The movemes differentiate very quickly, as early as 30% of the final movement is added.

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

TABLE I. We use the coefficients in V to order chronologically four images sampled from a tennis serve. For each method, we report the number of images out of position and swaps necessary to obtain the correct order.

Method	Reordered Sequence	Errors	Swaps
lfa3d		0	0
lfa2d		2	1
svd		2	3

constraints put on the coefficients of V. We obtain the best visualizations by approximately matching the number of latent factors with the number of recognizable actions contained in the dataset (10 for this experiment), and constraining the coefficients of V to be between 0 and 1.

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 gradient descent (Sec. III-C2 and Fig. 3), and requires an initial guess for each training instance. We investigate how sensitive is the model to initialization, and how close is the recovered angle of view to the ground truth. Fig. 7(a) 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".
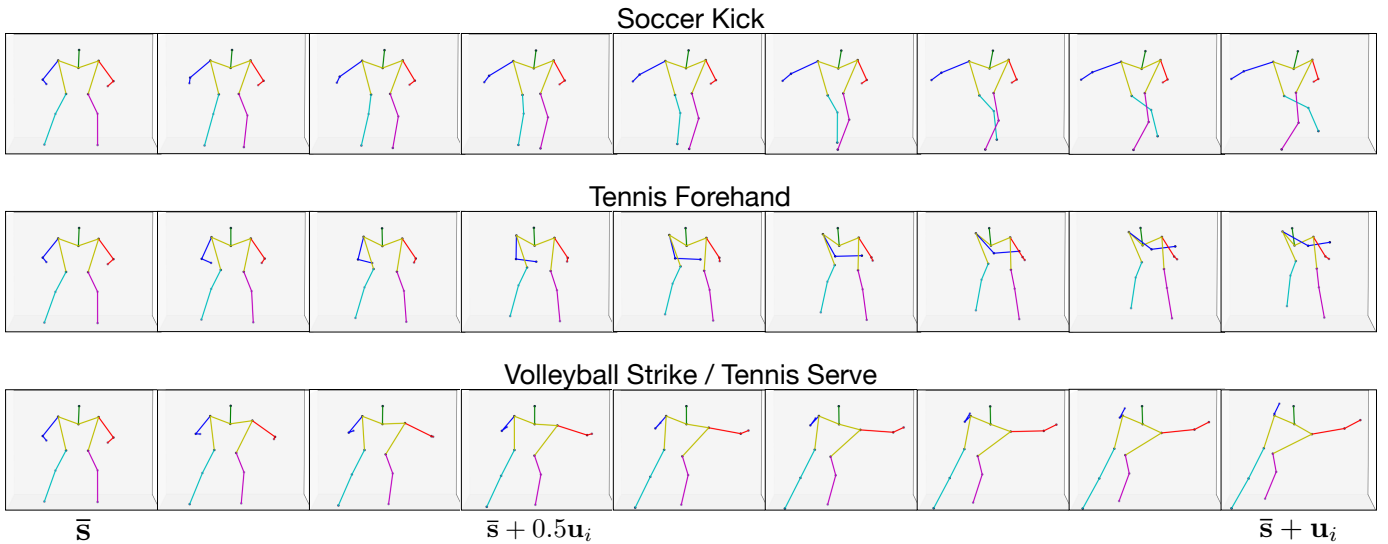
Soccer Kick

Tennis Forehand

Volleyball Strike / Tennis Serve

$$\bar{\mathbf{s}} \qquad \bar{\mathbf{s}} + 0.5\mathbf{u}_i \qquad \bar{\mathbf{s}} + \mathbf{u}_i$$

Fig. 6. **Learned Movemes Visualization**. Three latent factors, encoding movemes, from the learned bases poses matrix $\mathbf{U}$; two are easily interpretable ("soccer kick", "tennis forehand") and one is not as well defined ("volleyball strike / tennis serve"). The sequences are obtained by adding an increasing fraction of the basis to the mean pose of the dataset and differentiate very clearly, as early as 30% of the final movement, as visible in the second column. Full details in Sec. IV-B3.

As the number of clusters increases, we see that performance remains constant for "random" and "ground truth", while both evaluation metrics improve significantly for "coarse" initialization. For instance, using just four clusters, "coarse" initialization obtains almost minimal RMSE and perfect cosine similarity. 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: A desirable property of the obtained model is to be able to reconstruct with low error poses that are not contained in the training set, so the representation is not tied uniquely to the specific image collection it was learned from. 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. We repeated the experiment five times to obtain standard deviations.

As reported in Fig. 7(b), the RMSE over the training set is approximately constant, while the test set RMSE decreases significantly when going from 10% to 80% of the data used in training. This indicates that the learned latent factors can successfully reconstruct poses of unseen data.

6) Manifold Visualization: Fig. 8 visualizes an embedding of the manifold of human motion learned with the "lfa3d" method. Each pose in LSP is mapped in the human motion space through the coefficients of the corresponding column of $\mathbf{V}$ and then projected in two-dimensions using t-SNE [37].

Poses describing similar movements are mapped to nearby positions and form consistent clusters, whose relative distance depends on which latent factors are used to reconstruct the contained poses. Upper body movements are mapped closely
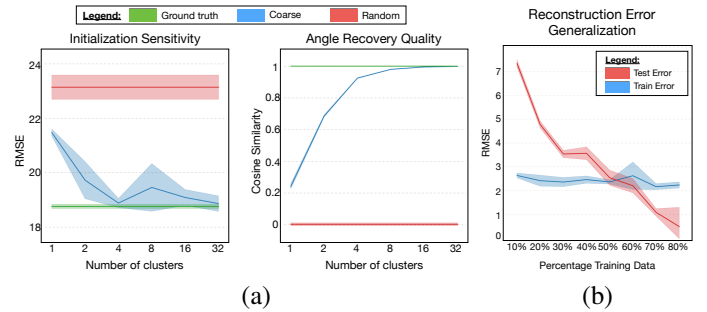

Fig. 7. **(a) Angle Recovery and (b) Generalization Performance**. (a) Sensitivity *wrt.* the initial value of the angle of view of the training poses of (Left) the Root Mean Squared Error and (Right) the Cosine Similarity between the learned and ground truth angles. A coarse initialization, within the correct quadrant of the true value, yields performances similar to ground truth. (b) The reconstruction error for poses not contained in the training set *wrt.* the percentage of data used in the training set. Full details in Sec. IV-B4, and Sec. IV-B5.

in the lower right corner, while lower body movements appear at the opposite end of the embedding. The mapping in the manifold is not affected by the direction each pose is facing, as nearby elements may have very different angle of view, confirming that the learned representation is rotation invariant. In Fig. 9, we show the heatmaps obtained from the activations of two latent factors from Fig. 6, overlaid on top of the t-SNE mapping of Fig. 8. To compute the heatmaps, we extract the coefficients for the "soccer kick" and "volleyball strike" latent factors from each column of $\mathbf{V}$ corresponding to a location in the embedding, and plot their value after normalization[3].

---

[3]To better depict the high-level trends, we enhance the contrast using a power of 1.5 and employ Gaussian smoothing.

Fig. 8. **Human Motion Manifold Visualization**. t-SNE embedding of the poses contained in the LSP dataset. Images, instead of poses, are shown for interpretability purpose. The type of body movement, and the influence of the learned bases poses determine the location in the manifold:"tennis serve" and "volleyball block" appear close in the manifold, while "running" is at the opposite end of the embedding. The angle of view does not affect the location in the manifold, as nearby poses may have very different angle of view. Full details in Sec. IV-B6.

Fig. 9. **Learned Movemes Heat-maps**. Activation strength of the learned "volley strike" and "soccer kick" bases poses from Fig. 6 (third and first row) in the t-SNE embedding. The heat-maps are consistent with Fig. 8 in which movements of the upper and lower body are respectively mapped to the low-right and high-left corner.

Clearly, the epicentrum of the "volleyball strike" basis pose is located where volleyball-like poses appear in the t-SNE plot (lower-right corner). Noticeable upward arm movements are not as present in many other sports, hence the low intensity of the activation in the rest of the map. Conversely, the "soccer kick" basis pose is mostly dominant in the top-left area and the heatmap is diffused, consistent with the observation that most poses contain some movement of the legs.

V. Conclusion and Future Directions

In this paper, we proposed a model for learning the primitive movements underlying human actions (movemes) from a set of static 2-D poses obtained from images taken at various angles of view. The bases poses are rotation-invariant and learned through a modified latent matrix factorization that intrinsically accounts for geometric properties inherent to viewing angle variability. The approach can be trained efficiently, requires modest effort to identify a reasonable initialization, and yields very good generalization 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 to different parts of the high-dimensional 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.

An interesting future direction of investigation would be to use the proposed model in a semi-supervised setting where there is some availability of true three-dimensional data along with a large collection of two-dimensional joint locations.

Other possible extensions of our work are: learning to morph actions and synthesize *unseen* actions from the set of extracted movemes; inferring the location of occluded or missing joints based on the position of the visible ones; applying these techniques to large-scale datasets [38] in conjunction with fine grained annotations of the performed actions [9], [10] to gain new insights on the structure, complexity, and duration of human behaviour.

REFERENCES

[1] E. S. Grood, S. F. Stowers, and F. R. Noyes, "Limits of movement in the human knee. effect of sectioning the posterior cruciate ligament and posterolateral structures." *J Bone Joint Surg Am*, vol. 70, no. 1, pp. 88–97, 1988.

[2] H. Hatze, "A three-dimensional multivariate model of passive human joint torques and articular boundaries," *Clinical Biomechanics*, vol. 12, no. 2, pp. 128–135, 1997.

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

[4] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 568–574.

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[6] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[7] S. C. Hoyer, A. Eckart, A. Herrel, T. Zars, S. A. Fischer, S. L. Hardie, and M. Heisenberg, "Octopamine in male aggression of drosophila," *Current Biology*, vol. 18, no. 3, pp. 159–167, 2008.

[8] 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)*, June 2014.

[9] 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.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016. [Online]. Available: http://arxiv.org/abs/1602.07332

[11] A. Yao, J. Gall, G. Fanelli, and L. J. Van Gool, "Does human action recognition benefit from pose estimation?." in *BMVC*, vol. 3, 2011, p. 6.

[12] D. Fouhey and C. Zitnick, "Predicting object dynamics in scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2019–2026.

[13] R. Parent, *Computer animation: algorithms and techniques*. Newnes, 2012.

[14] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 573–586.

[15] X. Fan, K. Zheng, Y. Zhou, and S. Wang, "Pose locality constrained representation for 3d human pose reconstruction," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 174–188.

[16] 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.

[17] X. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[18] 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.

[19] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3177–3184.

[20] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *Computer Vision–ACCV 2014*. Springer, 2014, pp. 428–443.

[21] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 17–24.

[22] Z. Kourtzi and N. Kanwisher, "Activation in human mt/mst by static images with implied motion," *Journal of cognitive neuroscience*, vol. 12, no. 1, pp. 48–55, 2000.

[23] T. Kim, G. Shakhnarovich, and R. Urtasun, "Sparse coding for learning interpretable spatio-temporal primitives," in *Advances in neural information processing systems*, 2010, pp. 1117–1125.

[24] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2650–2657.

[25] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.

[26] 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.

[27] S. T. Dumais, "Latent semantic analysis," *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.

[28] 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.

[29] 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.

[30] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[31] J. R. Gardner, M. J. Kusner, Y. Li, P. Upchurch, K. Q. Weinberger, and J. E. Hopcroft, "Deep manifold traversal: Changing labels with convolutional features," *arXiv preprint arXiv:1511.06421*, 2015.

[32] 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.

[33] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[34] M. R. Ronchi, J. S. Kim, and Y. Yue, "A rotation invariant latent factor model for moveme discovery from static poses," http://www.vision.caltech.edu/~mronchi/projects/RotationInvariantMovemes/, 2016.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[36] P. P. de Groen, "An introduction to total least squares," *arXiv preprint math/9805076*, 1998.

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.