Predicting Dynamical Evolution of Human Activities from a Single Image

Suhas Lohit[1], Ankan Bansal[2], Nitesh Shroff[3], Jaishanker Pillai[4], Pavan Turaga[1] and Rama Chellappa[2]

[1]Geometric Media Lab, Arizona State University, Tempe
[2]Center for Automation Research (UMIACS), University of Maryland, College Park
[3]Zoox Inc., and [4]Google Research

Abstract

A human pose often conveys not only the configuration of the body parts, but also implicit predictive information about the ensuing motion. This dynamic information can benefit vision applications which lack explicit motion cues. The human visual system can easily perceive the dynamic information in still images. However, computational algorithms to infer and utilize it in computer vision applications are limited. In this paper, we propose a probabilistic framework to infer the dynamic information associated with a human pose. The inference problem is posed as a non-parametric density estimation problem on a non-Euclidean manifold of linear dynamical models. Since direct modeling is intractable, we develop a data driven approach, estimating the density for the test sample under consideration. Statistical inference on the estimated density provides us with quantities of interest like the most probable future motion of the human and the amount of motion information conveyed by a pose. Our experiments demonstrate that the extracted motion information benefits numerous applications in computer vision. In particular, the predicted future motion is useful for activity recognition, human trajectory synthesis, and motion prediction.

1. Introduction

As traditional vision problems like people tracking [46] and activity recognition [1] are based on video as input, motion cues play an important part in these applications. However, with the availability of personal photo collections, images from web-sources, human activity in still images is gaining importance. These image-based applications do not have explicit motion cues, and are currently limited to using just the appearance cues [34, 45]. This leads to an interesting question: Can implicit motion cues be extracted from still images of humans, and used to aid visual analysis?

Extensive studies in psychology have shown that information present in the posture of the human body plays a vital role in biological motion perception [4, 23]. Experiments of Hirai and Hiraki [17] demonstrated that destroying the body structure led to significant reduction in motion perception in humans when compared to destroying the temporal structure of motion. Furthermore, humans can easily anticipate the future motion of actors from their current body configuration [22].

Dynamic information in human poses can aid computer vision systems in multiple ways. Similar to biological systems, vision systems can utilize this information to efficiently predict the future motion of human users. In robotic applications like "assistance to manipulation", robots often assist humans or manipulate the same object as humans. In such applications, accurate prediction of human motion can improve robotic performance, as empirically verified by Jarrasse *et al*. [19]. Dynamic information in poses can also improve activity recognition from still images and aid the synthesis of realistic human motion. The latter is useful in applications involving humanoid robots and animation. Japanese Manga images in Figure 1 is a case in point, which illustrates a technique from the arts, pioneered by Hokusai, of conveying motion using physically unstable human configurations [26]. This is in contrast to other approaches such as showing motion via streak-lines, which are not intrinsic to the depicted human.

Contributions: Motivated by the above discussion, we develop a computational model to infer the "next move" from still images. Our goal is to predict the future motion of a human given a single pose and quantify the extent to which it is constrained by a given pose. We emphasize that the input to our algorithm is just a single image containing a human, and the goal is to predict the motion of the human and the type of action performed. We demonstrate the usefulness of the estimated dynamic information in a variety of vision applications like human motion prediction and activity recognition.

Figure 1. Database of 45 Hokusai Manga Images. The functional Magnetic Resonance Imaging (fMRI) studies by Osaka *et al.* [26] illustrated that the dancer images on the left in unstable poses activated the motion sensitive visual cortex in humans, indicating that humans can perceive the implied motion in these images. However, the priest images on the right in stable poses elicited low responses of implied motion in humans. We use this dataset to experimentally validate the proposed computational model.

Organization of the paper: A brief review of related work is presented in Section 2. The proposed framework for extracting dynamic information in human pose is described in Section 3. We empirically evaluate the proposed technique in Section 4 and present conclusions in Section 5.

2. Related Work

Recently, several algorithms for action recognition from still images have been developed. Thurau and Hlavác [34] recognized human actions by representing actions as a histogram of pose primitives, and using histogram matching for recognition. Ikizler *et al.* [18] represented the human pose using histogram of rectangular regions and used SVMs for classification. [9] used oriented rectangular patches extracted from the human silhouette to represent the action, followed by histogram matching for recognition. Human pose in the query image was considered as a latent variable in [44]. The latent SVM model was used for recognizing activities in this work. However, these techniques are often applicable only for simple actions, since complex activities cannot always be captured by a single pose. Nevertheless, even poses belonging to complex activities often provide information about the local motion trajectory. For instance, consider the pose π_2 in Figure 2. While it is easy to infer that the person is bending down, it is difficult to predict the ensuing activity (for example sitting down or picking up a ball). In this work, we focus on estimating this motion information associated with the human pose in still images.

Another line of research is motion estimation from still images of natural scenes. Roth and Black [28] learned the prior probability of motion fields from still images of natural scenes using an MRF model. Their experiments demonstrated that the learned motion prior captures the rich spatial structure found in natural scenes, and can also improve motion estimation accuracy in test videos. Liu *et. al* [25] proposed SIFT flow, a method to densely align scene images by matching densely sampled pixel-wise SIFT features, while preserving continuity. Motion of pixels in query images were then predicted by transferring SIFT flow from similar training images. Yuen and Torralba [47] estimated the probability density of local motion trajectories in a non-parametric manner at each pixel location, and used samples from the density to estimate the motion trajectories in query scene images. These methods capture only the local structure of the scene, and not the influence of the global scene on the ensuing motion. Hence, they are not directly applicable to human motion prediction, where future motion is dependent on the entire human pose. On the other hand, we directly model the relationship between the human pose and the future motion of the human body in this work.

More recently, several works use deep learning to generate future video frames based on past frames. These include works by Vondrick et al. [39] which attempts to generate video frames using a CNN with a purely data-driven approach. However this fares poorly due to the very large space of predictions. Vondrick and Torralba [40] instead propose to predict transformations that are applied to the input frames, in order to generate the future frames. However, although superior to the previous work, the quality of frames thus generated is still poor. Srivastava et al. [33] showed that LSTMs can be effective for generating entire future frames. For human actions, instead of generating future frames directly, an intermediate step of predicting future poses can be used to improve results considerably. This was shown recently by Villegas et al. [37] and by Walker et al. [41]. In this work, we propose to use relevant examples from a large training set in order to predict human motion from a single image. Similar approaches have been successful in applications like image super-resolution [11] and image inpainting [15]. Another recent paper that is similar in spirit to our paper is that of Bansal et al. [3] which proposes conditional image-synthesis using a two-step pipeline – a CNN first generates an intermediate image which is then refined based on nearest neighbors in the training dataset. The refinement step produces more diverse images and increases interpretability.

3. Dynamic inference from a human pose

Before developing a computational model, we first analyze the physical evidence for the existence of dynamic information in this section. Starting at a particular pose, the future motion of the human body is constrained by numerous factors. The mechanics of body joints prevent arbitrary motion of the body. Laws of physics, like gravity and momentum, also limit future movements of the human. Furthermore, every realistic pose is part of a human activity with a well-defined objective. These constraints on the future motion of the human body are responsible for the dynamic information associated with a particular pose. Furthermore, the set of possible future motions vary

widely between different human poses.As the input feature, we choose the simple HOG-based model [34], representing the human pose using the HOG features extracted from the bounding box. This avoids the need for training models for pose estimation, can generalize to new poses and is robust to errors in the estimated pose parameters. However, the core framework will generalize to any other pose representation.

Given a human pose, there is a set of possible trajectories (in the chosen pose-space) originating from it, and the exact future motion is uncertain. To capture this uncertainty, we develop a probabilistic framework, estimating the conditional probability distribution of subsequent human motion given a pose. Once this distribution is obtained, one can compute useful statistics such as its mode and entropy. Given a single pose, the mode of the conditional distribution gives us the most probable temporal evolution of poses. The entropy of this distribution measures the uncertainty in these future sequences. The work of Kerzel [21] shows that this uncertainty (unpredictability) provides a measure of the amount of dynamic information perceived by humans in a pose. The higher the predictability of motion from a pose, the higher the dynamic information it conveys.

To develop a probabilistic model, we first need to define the space of predictions. Firstly, from a stable pose the set of possible human motions that can follow is extremely large. Further, even for predictable poses where the set of future motions is potentially constrained, there is an equivalence class of future motions differing only in the rate of execution. Hence, we need a representation of motion that is invariant to the rate of execution. Considering the above, we first model human activities as a sequence of movements called action segments, separated by "ballistic" boundaries [38]. These movements are natural units of human actions, typically comprising an initial acceleration of limbs towards a target followed by deceleration to stop the movement. Figure 2 shows a simple illustration of the ballistic boundaries. Here, the ballistic boundaries highlighted in red separate the "picking up" action into two action segments, namely the "bending down" action segment and the "getting up" action segment. Vitaladevuni *et al*. [38] have developed computational models to automatically extract ballistic motion boundaries from videos. By viewing actions as separated by ballistic motion boundaries, we can restrict the scope of the motion prediction problem to predicting statistics over future action segments, which are shorter in duration. Also, since ballistic boundaries are robust to the rate of execution, the estimated statistics are robust to rate as well.

We now introduce the notation and elements of our approach. Let π_i represent the i^{th} pose and $\Pi = \{\pi_i, i = 1, \dots M\}$ be the set of all possible human poses. Similarly, let ϕ_i represent the i^{th} action segment and $\Phi = \{\phi_i, i = 1 \dots N\}$ be the set of all possible action segments. Any action α is a temporally ordered sequence of action segments $\left[\phi_{\alpha_1}, \dots \phi_{\alpha_{t(\alpha)}}\right]$, where each action segment ϕ_k is itself a temporally ordered sequence of individual poses $\left[\pi_{k_1}, \dots \pi_{k_{t(k)}}\right]$. This action α consists of two action segments $[\phi_1, \phi_2]$. Action segment ϕ_1 is a temporally ordered sequence of poses $[\pi_1, \pi_2, \pi_3, \pi_4]$. Similarly, segment ϕ_2 is a temporally ordered sequence of poses $[\pi_4, \pi_5, \pi_6]$.

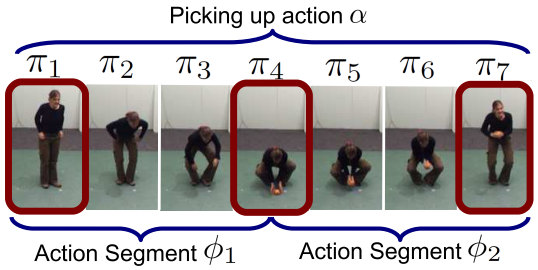

Figure 2. Illustration of ballistic boundaries for the "picking up" action. The three ballistic boundaries π_1, π_4 and π_7, highlighted in red, divide the action α into two action segments ϕ_1 and ϕ_2.

Let $\mathcal{P}(\phi|\pi)$ denote the conditional probability that a given pose π occurred in an action segment ϕ . As discussed earlier, the uncertainty in the temporal evolution of poses starting from π is low, if it has high dynamic information. In an information-theoretic framework, this uncertainty can be measured by the entropy $\mathcal{H}(\phi|\pi)$ of the conditional distribution of an action segment given a pose.

$$\mathcal{H}(\phi|\pi) = -\int_{\phi \in \Phi} \mathcal{P}(\phi|\pi) \log(\mathcal{P}(\phi|\pi)) d\phi \tag{1}$$

This motivates our measure, Degree of Dynamic Information (DDI) of a pose, which can be computed as

$$\text{DDI}(\pi) = \exp[-\mathcal{H}(\phi|\pi)] \tag{2}$$

where the negative exponent captures the inverse relationship between uncertainty in the temporal evolution of poses starting from π and the amount of dynamic information in π. Another piece of valuable information that can be immediately obtained from $\mathcal{P}(\phi|\pi)$ is the most probable action segment $\hat{\phi}$ that contains the pose π.

$$\hat{\phi}(\pi) = \arg\max_{\phi \in \Phi} \mathcal{P}(\phi|\pi) \tag{3}$$

Similarly, given a start pose π_s and an end pose π_e, we can obtain the most probable pose trajectory as

$$\hat{\phi}(\pi_s, \pi_e) = \arg\max_{\substack{\phi \in \Phi \\ \phi = [\pi_s, \dots, \pi_e]}} \mathcal{P}(\phi|\pi_s, \pi_e) \tag{4}$$

Having defined the two terms using $\mathcal{P}(\phi|\pi)$, the question now turns to estimating this density. Explicitly modeling this density and estimating its parameters from finite training data is extremely difficult and prone to overfitting

due to the large variations in humans poses and future motions in unconstrained settings. Hence, we adopt the data-driven approach, which has become very popular in recent years [25, 14, 24]. This approach advocates transferring information from a rich training database to the specific query under consideration, instead of learning a general function applicable to all queries. Such methods have shown significant promise in solving otherwise difficult tasks such as scene alignment [25], geo-localization [14], scene completion [16], scene parsing [24] and object matching [31].

Given a test pose π_s, we estimate $\mathcal{P}(\phi|\pi_s)$ directly from the training data. This estimate is then used to compute the amount of associated dynamic information $\text{DDI}(\pi_s)$ and the most probable action segment $\hat{\phi}(\pi_s)$. We explain this approach in detail below.

3.1. Estimation of Conditional Distribution

Instead of developing a functional form for $\mathcal{P}(\phi|\pi_s)$, we compute this probability whenever we encounter a test pose π_s. Our training data consist of videos of human actions. Let $\mathcal{D}$ denote the database of all the poses which are extracted from these videos. By applying the temporal segmentation algorithm of Vitaladevuni *et al.* [38], these videos are divided into action segments separated by ballistic boundaries. Given a test pose π_s, we find all the instances of the pose in the database $\mathcal{D}$ and denote this set by $\mathcal{N}_{\pi_s}$. In our experiments, nearest neighbors of the test pose π_s in the database $\mathcal{D}$ are used to form the set $\mathcal{N}_{\pi_s}$. Note that every pose $\pi \in \mathcal{D}$ is a part of an action segment $\phi \in \Phi$. This implies that every pose $\pi_r \in \mathcal{N}_{\pi_s}$ has an associated action segment $\phi(\pi_r)$. Let $\mathcal{N}_{\phi(\pi_s)}$ be the set of action segments corresponding to the poses in $\mathcal{N}_{\pi_s}$. $\mathcal{N}_{\phi(\pi_s)} = \{\phi(\pi), \pi \in \mathcal{N}_{\pi_s}\}$ can be considered as samples from the density $\mathcal{P}(\phi|\pi_s)$. Hence, sample-based density estimation techniques can be adopted to estimate $\mathcal{P}(\phi|\pi_s)$ given $\mathcal{N}_{\phi(\pi_s)}$. However, such techniques cannot be applied directly on the space of action segments Φ due to two reasons. First of all, action segments can differ in the number of frames. Hence, a direct representation in terms of the associated pixels lead to vectors of different dimensionality. Secondly, this direct representation in terms of the associated pixels is high dimensional. Learning models from higher dimensional data is often impractical, and has lead to the development of alternate low dimensional representations for the data [13]. Hence we adopt a parametric approach, where the action segments are compactly represented by a low dimensional dynamical model.

Modeling Action Segments: In this work, we employ the linear dynamical system (LDS) [32], a popular model in computer vision. For an action segment ϕ, the LDS model is described by

$$\begin{aligned} z_\phi(t+1) &= A_\phi z_\phi(t) + v_\phi(t), v_\phi(t) \sim N(0, \Xi) \\ y_\phi(t) &= C_\phi z_\phi(t) + w_\phi(t), w_\phi(t) \sim N(0, \Theta) \end{aligned} \quad (5)$$

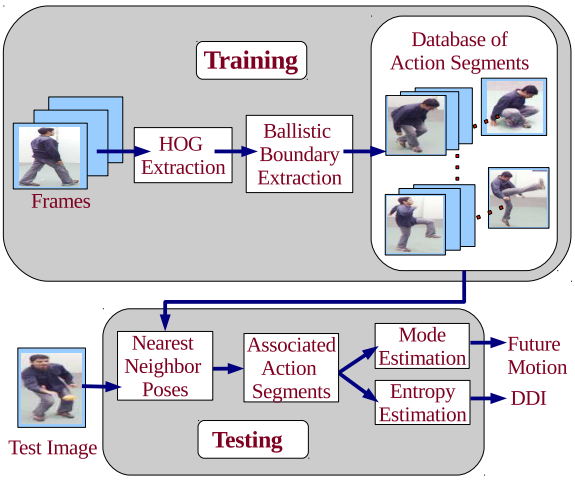

Figure 3. Block diagram demonstrating the various steps in the proposed method.

where $z_\phi(t) \in \mathbb{R}^p$ is the hidden state vector for the t^{th} frame in the action segment ϕ, $y_\phi(t) \in \mathbb{R}^d$ are the features extracted from t^{th} frame, $A_\phi \in \mathbb{R}^{p\times p}$ is the transition matrix, $C_\phi \in \mathbb{R}^{d\times p}$ is the measurement matrix. $v_\phi(t)$ and $w_\phi(t)$ are the noise components, which are modeled as Gaussian with mean zero and covariances Ξ and Θ respectively. A_ϕ is constrained to have eigen vectors inside the unit circle, while C_ϕ is constrained to be orthonormal. Hence, the parameters of the LDS model, namely (A_ϕ, C_ϕ) do not lie on the Euclidean space. For comparison of actions, a commonly used distance metric is the subspace angles between the column spaces of the corresponding observability matrices. The 'observability' matrix $\hat{\Omega}_\phi$ of an action segment ϕ is given by $\Omega_\phi^\top = \left[C_\phi^\top, (C_\phi A_\phi)^\top, \ldots, (C_\phi A_\phi^{m-1})^\top, \ldots\right]$. It is an infinite dimensional matrix, which can be approximated by the finite matrix $\hat{\Omega}_\phi^\top = \left[C_\phi^\top, (C_\phi A_\phi)^\top, \ldots, (C_\phi A_\phi^{m-1})^\top\right]$. Note that $\hat{\Omega}_\phi \in \mathbb{R}^{n\times p}$, where $n = md$. Hence the column space of $\hat{\Omega}_\phi$ is a p-dimensional subspace in $\mathbb{R}^n$, which constitute the Grassmann manifold $\mathcal{G}_{n,p}$. For notational simplicity, we denote the observability matrices $\hat{\Omega}_\phi$, $\hat{\Omega}_{\phi_i}$ and $\hat{\Omega}_{\phi_j}$ by Ω, Ω_i and Ω_j respectively. Then, a natural metric $\zeta^2(\Omega_i, \Omega_j)$ between action segments ϕ_i and ϕ_j is given by [35]

$$\zeta^2(\Omega_i, \Omega_j) = p - tr(\Omega_j^T \Omega_i \Omega_i^T \Omega_j). \quad (6)$$

Density Estimation on the Grassmann Manifold: Using $\mathcal{N}_{\phi(\pi_s)}$, the set of samples from $\mathcal{P}(\phi|\pi_s)$, we now estimate the conditional density using non parametric density estimation techniques [8], as

$$\hat{\mathcal{P}}(\phi|\pi_s) = c_1 \sum_{\phi_i \in \mathcal{N}_{\phi(\pi_s)}} \Psi(M^{-\frac{1}{2}}(I_d - \Omega_i^\top \Omega \Omega^\top \Omega_i) M^{-\frac{1}{2}}) \quad (7)$$

where $\Psi(T) = \exp(\mathrm{tr}(-T))$ for $T \in \mathbb{R}^{p\times p}$, $\mathrm{tr}(.)$ is the matrix trace operator, $M \in \mathbb{R}^{p\times p}$ is a smoothing matrix and c_1 is the normalization factor to ensure that the probability density integrates to unity.

3.2. Statistical Inference on the Estimated Density

Having formulated the conditional density for the action segment given the test pose, we now estimate statistical information from it. The block diagram of the proposed method in shown in Figure 3.

Mode Estimation: Given a pose π_s, the likely future motion can be predicted by finding the most probable action segment $\phi^*(\pi_s)$, which is the mode of the distribution $\mathcal{P}(\phi|\pi_s)$. Non-parametric techniques have been recently developed for mode seeking on analytic manifolds [36, 5]. In particular, Cetingul and Vidal [5] computes the mode on the Grassmann manifold using iterative optimization. It intrinsically locates the modes of the distribution via consecutive evaluations of a mapping. For the Grassmann manifold, these evaluations constitute an efficient gradient ascent scheme, which avoids the computation of expensive exponential mappings. However, this algorithm will only compute the LDS parameters of the most probable action segment. It is not possible to generate the frames of the action segment from the LDS parameters. Hence, in applications where a valid action segment with high probability of occurrence is required, a more efficient scheme is to directly select the action segment with the highest conditional density from $\mathcal{N}_{\phi(\pi_s)}$.

$$\hat{\phi}(\pi_s) = \underset{\phi_i \in \mathcal{N}_{\phi(\pi_s)}}{\arg\max}\ \hat{\mathcal{P}}(\phi_i|\pi_s) \tag{8}$$

By similar analysis, we can also obtain $\hat{\phi}(\pi_s, \pi_e)$, the most probable pose trajectory given a start pose π_s and an end pose π_e, by using the samples from $\mathcal{N}_{\phi(\pi_s,\pi_e)}$. Here, $\mathcal{N}_{\phi(\pi_s,\pi_e)}$ denote the set of training action segments, whose start and end poses are nearest neighbors of π_s and π_e respectively.

$$\hat{\phi}(\pi_s, \pi_e) = \underset{\phi_i \in \mathcal{N}_{\phi(\pi_s,\pi_e)}}{\arg\max}\ \hat{\mathcal{P}}(\phi_i|\pi_s, \pi_e) \tag{9}$$

where $\hat{\mathcal{P}}(\phi_i|\pi_s, \pi_e)$ is given by $\hat{\mathcal{P}}(\phi|\pi_s, \pi_e) = c_1 \sum_{\phi_i \in \mathcal{N}_{\phi(\pi_s,\pi_e)}} \Psi(M^{-\frac{1}{2}}(I_d - \Omega_i^\top \Omega \Omega^\top \Omega_i) M^{-\frac{1}{2}})$ with $c_1, \Psi(\cdot)$ and M denoting similar quantities as in (7).

Entropy Estimation: To estimate the entropy of $\mathcal{P}(\phi|\pi_s)$ from the samples $\mathcal{N}_{\phi(\pi_s)}$, we use the resubstitution estimate of entropy [2] as follows

$$\hat{\mathcal{H}}(\phi|\pi_s) = -\frac{1}{|\mathcal{N}_{\phi(\pi_s)}|} \sum_{\phi_i \in \mathcal{N}_{\phi(\pi_s)}} \log \hat{\mathcal{P}}(\phi_i|\pi_s) \tag{10}$$

where $\hat{\mathcal{P}}(\phi_i|\pi_s)$ is obtained from (7). Under mild conditions, this estimate has been proved to be consistent in the first and second order means [2].

Figure 4. Motion prediction using the proposed method. Note that the predicted motion is performed by a different subject, since there is no overlap between training and testing subjects.

4. Experiments

In this section, we empirically evaluate the proposed method on action datasets of varying complexity, namely the Weizmann activity dataset [12] with clean background and fixed view point, INRIA XMAS (IXMAS) dataset [43] where actors freely change their orientation and the UCF Sports Activity dataset [27] with large changes in scene and view points.

The dynamic information associated with a given human pose can benefit several computer vision problems like motion prediction from still images, and semi-supervised still image action recognition. For human motion prediction from still images, we represent the future poses in terms of a sequence of images of humans, as shown in Figure 4.The proposed method for predicting action segments can act as a natural way of propagating labels from the labeled training images to the unlabeled video data. For each labeled training pose, we find the most probable action segments from the unlabeled video data, as explained in Section 3.2. If the original training poses are discriminative, the retrieved action segments will belong to the same action. Hence, we add these action segments as additional training samples. Additionally, one could use DDI to propagate labels from just the informative training poses. To evaluate the method under large variations in training and testing conditions, we perform a cross dataset experiment using unlabeled videos from the Weizmann dataset and test images from the CMU action dataset [20]. The poses are represented by HOG [10] features, and action segments by the finite observability matrix $\Omega_m^\top$ in the LDS model. Given a test pose π_s, $\mathcal{N}_{\pi_s}$ is created by identifying the k nearest neighbors in the HOG feature space from the gallery. Unless specified, we fixed k in all our experiments to the average number of repetitions of actions in the unlabeled videos, which is roughly the number of subjects in the unlabeled videos. Our experiments suggest that the proposed method works well over a wide range of k. The bin size and cell size of the HOG features are both set to 8, with 2×2 cells forming a block.

4.1. Perceptual Evaluation on Manga Images

In this section, we estimate the amount of dynamic information in the Hokusai Manga image database, which we

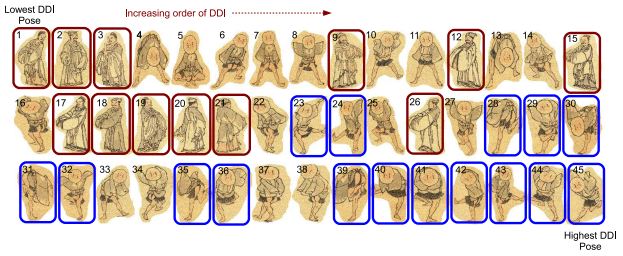

Figure 5. The priest and dancer images in the Hokusai Manga collection are displayed in the increasing order of their DDI, with the indices in the sorted order indicated in the top left of each pose. The priest images are marked in red, and the dancer images having the most unstable poses, where the human is standing on a single leg are marked in blue. Observe that most of the priest images (in red) have lower DDI values, while most of the dancer images in unstable poses (in blue) have higher DDI values, reflecting the the perceptual results in [26].

compiled from web-sources. This database consists of 45 images belonging to two groups namely the priests and the dancers. The same set of images had been used by Osaka *et al.* [26] in their experiments, which reported that the unstable poses in the dancer images activated the motion sensitive regions of the visual cortex, while the priest images did not. This indicates that the dancer images have higher dynamic information compared to the priest images.

Since the Manga images have wide variations in human poses, we use the SFU skating dataset [42] for training.For each Manga test image, we do a simple thresholding to obtain a binary image and extract the HOG features. Using the SFU training data, we obtain the DDI values for each Manga image. The Manga images are then sorted in increasing order of DDI and are displayed in Figure 5. The priest images are highlighted in red. As can be observed, most priest images have low DDI values indicating low amounts of implied motion. Furthermore, among the dancer images, the most unstable poses are the ones where the human is standing on one leg. Such images are highlighted in blue. Based on the studies in [26], such unstable poses should have higher implied motion. These images come towards the end of the sorted order in Figure 5, indicating that the DDI values are higher in them. Thus, most of the stable poses have lower DDI values, and most of the unstable ones have higher DDI values, thereby empirically verifying that the proposed measure is perceptually meaningful.

4.2. Human Motion Prediction from still images

We predict motion given a single pose on the IXMAS dataset. We used the first nine subjects in the first view as training data and predicted future motion for each pose of the last subject. The predicted motion of some of the test poses are shown in Figure 4. We can observe that the predicted motion mostly agrees with ones expected by humans.

To evaluate the prediction accuracy, we used the motion prediction error, which is defined as the difference between the true action segments for each test frame and the predicted action segment. We use the distance metric between action segments defined in (6). We plot this error for the proposed method for different values of k in Figure 6. The baseline method (NN-Based) consists of using the mean of the k retrieved action segments as the predicted motion. Using the first nearest neighbor as the prediction motion, the prediction error is 0.47. The proposed method decreases this prediction error considerably achieving an error of 0.39 using 6 nearest neighbors. Also, the simple baseline of averaging the retrieved nearest neighbor action segments leads to higher prediction error for higher values of k. We attribute the improvement in performance to the following. Due to errors in pose matching, the nearest neighbor poses and their associated motion are often erroneous. These erroneous motion normally form outliers and do not contribute to the most probable motion. Since mean is not robust to outliers, averaging the retrieved action segments lead to poor performance. However, the mode is not sensitive to outliers. Hence, by finding the mode of the nearest neighbor action segments, the proposed method improves the robustness of the motion prediction algorithm to errors in pose matching.

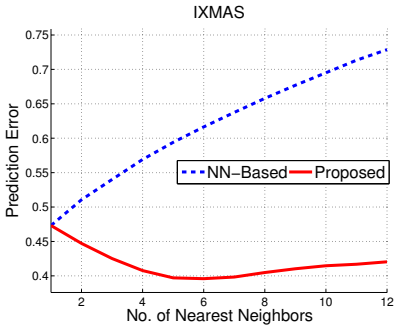

Figure 6. Motion prediction error in IXMAS dataset using the nearest neighbor-based and the proposed methods. Due to outliers in the nearest neighbor poses, the NN-Based method leads to lower performance with more nearest neighbors. However, since the proposed method of mode computation is insensitive to outliers, the motion prediction error is reduced with more nearest neighbors by the proposed method.

4.3. Semi-supervised single-image action recognition

In this section, we evaluate the label propagation technique for semi-supervised activity recognition from a single image. We used the UCF Sports Activity dataset in our experiments. We considered nine out of the thirteen actions, avoiding the classes differing only in motion. Action classes which differ only in their motion signature cannot be distinguished in still images, even by humans. The classes used in our experiments are shown in Figure 7. We used 2 subjects for training, 2 for testing and 8 as unlabeled data with

no overlap. We chose 8 images at random of the 2 training subjects to form the training data. No labels or temporal segmentation is assumed for the unlabeled data. We used HOG features for representing human poses and the nearest neighbor classifier for activity recognition, similar to the approach introduced in [34].

We compared the proposed method of label propagation with the nearest neighbor classifier using the labeled data alone (supervised algorithm) and three popular semi-supervised algorithms namely Self-Training [6], Semi-Supervised SVM (S3VM) [30] and Single View CoTraining [7]. In Self-Training, the classifier trained on the labeled data is applied on the unlabeled data and the L (fixed as 20 in our experiments) most confident images are added to the training set as additional labeled data, using the predicted labels. Test samples are classified using this extended training set. We used one-versus-all classification for multiclass classification. We used the Multiple Switching algorithm in [30], which iteratively labels the unlabeled data and switches the labels to reduce the optimization cost. The Single View CoTraining algorithm automatically splits the feature vectors into two views, and uses the most confident samples in one view to retrain the other view. In the proposed method for label propagation, for each labeled image, we added the k most probable action segments from the unlabeled data into the training set. We used $k = 8$ in our experiments, since unlabeled data contained each action roughly 8 times, performed by each of the 8 subjects. Recognition of test samples were done as before using the extended training set.

Method	Accuracy (%)
Supervised	49.3
Self-Training	51.9
Semi-Supervised SVM	51.7
Single View CoTraining	53.5
Proposed Method	**57.9**

Table 1. Activity Recognition accuracy on the UCF dataset.

The recognition accuracies using 8 action segments are shown in Table 1 . We include the corresponding confusion matrices in Figure 7. The proposed method provides a significant improvement of 8.6%, compared to the supervised algorithm. Also confusion with wrong classes is considerably reduced. We show some of the test images and the nearest neighbors obtained by the supervised algorithm and the proposed method in Figure 8. Comparison between nearest neighbor poses added by Self-Training and those added by the proposed approach for the 'Diving' action can be seen in Figure 9. We also plot the variation in accuracy with the number of action segments added in Figure 10. As can be observed, the accuracy increases with action segments till 9 (close to 8, the average number of repetitions in the unlabeled data) and then falls gradually.

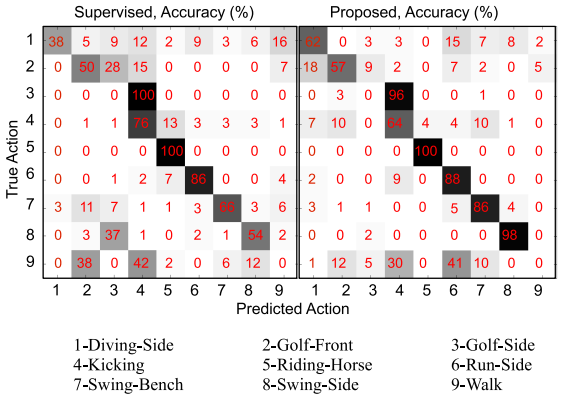

Figure 7. Confusion matrices for action recognition on UCF dataset show significant improvements. In the proposed method, confusion remains mainly between Golf Side and Kicking which have similar leg poses (legs far apart), and among walk, run and kicking, which differ mainly in the rate of execution of the action.

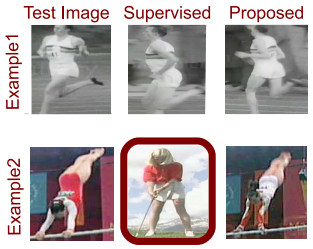

Figure 8. For each test image, the nearest neighbors obtained using the supervised method and the proposed method are shown. Erroneous results are encircled in red.

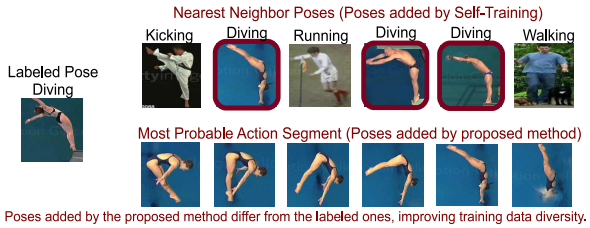

Figure 9. Illustration of the proposed label propagation approach for semi-supervised action recognition, for the diving action. The correctly retrieved nearest neighbor poses are in red. While some of the nearest neighbors belong to incorrect activities due to errors in pose matching, the most probable action segment belongs to the correct class. Furthermore, the poses added by the proposed method are clearly different from the test pose. Hence, the training set is greatly enriched by the proposed label propagation method.

4.4. Cross-dataset dynamic inference

To evaluate the robustness of our method, we consider the scenario where the test pose whose motion is to be inferred, is significantly different from the unlabeled videos available for learning the conditional density. Specifically, we picked poses from the CMU dataset [20] and learned

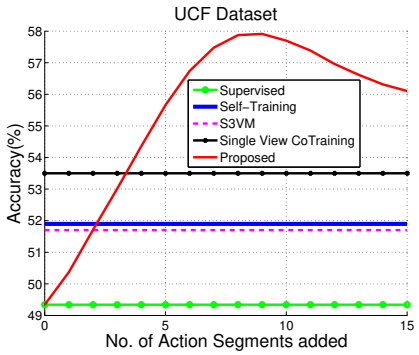

Figure 10. Variation of recognition accuracy with the number of action segments added per training image.

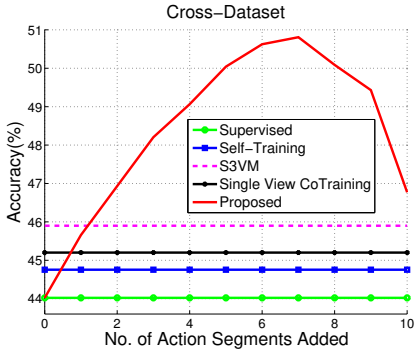

Figure 11. Variation of recognition accuracy with the number of nearest neighbors in the CMU cross-dataset experiment.

conditional density using the videos from the Weizmann dataset. We then propagated action segments from the Weizmann dataset into the training set as explained before. Test poses in the CMU dataset were recognized using this extended training set. Out of the four actions in the CMU dataset which are also present in the Weizmann dataset, we use "jumping jack", "one handed wave" and "pickup" for our experiment. We avoid the fourth action, namely, the "two handed wave", since it closely resembles jumping jack in still images. The entire Weizmann dataset is used for learning the conditional density, without any labeling or temporal segmentation.

Method	Accuracy (%)
Supervised	44.0
Self Training	44.8
Semi-Supervised SVM	45.9
Single View CoTraining	45.2
Proposed Method	**50.5**

Table 2. Activity Recognition accuracy on the CMU dataset.

We picked one image per action for training from the CMU dataset and tested on frames from the remaining videos. For each training image, we added the most probable action segments for the Weizmann dataset. To reduce the cross-dataset variations, before recognition, we learned a Partial Least Squares(PLS) subspace, using the training samples from the CMU dataset and then added action segments from the Weizmann dataset. PLS-based latent spaces have been used to handle cross-dataset and cross-model recognition [29]. We observed the method to be robust to the subspace dimension and chose half the original feature dimension in our experiments. We present the recognition accuracies in Table 2, and plot the performance with varying number of nearest neighbors (k) in Figure 11.

5. Conclusion

In this paper, we proposed methods to model the implicit motion information contained in a human pose. We introduced a probabilistic framework to infer this dynamic information, by posing this inference as a density estimation problem on non-Euclidean manifolds. Direct modeling of the density is difficult and prone to error due to the large variation in human poses. Hence, we have developed a non parametric data-driven approach for estimating the density and the associated statistics. Utilizing the proposed model, we predicted the most probable future sequence of poses and the amount of dynamic information conveyed by a given image. We showed the utility of the proposed framework in human motion prediction and activity recognition. Future work include exploring other useful statistics within the proposed framework. We will evaluate the increase in probability as more frames are added. This will enable us to determine the minimal number of frames needed to detect an activity at a pre-specified probability of detection. We will also study the influence of scene and object context on the dynamic information conveyed by human poses.

Acknowledgements

SL and PT were supported by NSF grant 1452613 and ARO grant W911NF-17-1-0293.

The work of AB and RC was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of IARPA, DOI/IBC or the U.S. Government.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), 2011.

[2] I. Ahmad and P. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory*, 22(3), 1976.

[3] A. Bansal, Y. Sheikh, and D. Ramanan. Pixelnn: Example-based image synthesis. *arXiv*, 2017.

[4] L. Battelli, P. Cavanagh, and I. M. Thornton. Perception of biological motion in parietal patients. *Neuropsychologia*, 41(13), 2003.

[5] E. Cetingul, H. and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *International Conference on Computer Vision and Pattern Recognition*, 2009.

[6] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, 2006.

[7] M. Chen, K. Q. Weinberger, and Y. Chen. Automatic feature decomposition for single view co-training. In *International Conference on Machine Learning*, 2011.

[8] Y. Chikuse. *Statistics on Special Manifolds*. Springer-Verlag, 2003.

[9] N. Cinbis, R. Cinbis, and S. Sclaroff. Learning actions from web. In *International Conference on Computer Vision*, 2009.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision and Pattern Recognition*, 2005.

[11] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12), 2007.

[13] S. Hauberg and K. S. Pedersen. Predicting articulated human motion from spatial processes. *International Journal of Computer Vision*, 94(3), 2011.

[14] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

[15] J. Hays and A. A. Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007.

[16] J. Hays and A. A. Efros. Scene completion using millions of photographs. *Communications of the ACM*, 51(10), 2008.

[17] M. Hirai and K. Hiraki. The relative importance of spatial versus temporal structure in the perception of biological motion: An event-related potential study. *Cognition*, 99(1), 2006.

[18] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *International Conference on Pattern Recognition*, 2008.

[19] N. Jarrasse, J. Paik, and G. Morel. Can human motion prediction increase transparency? In *International Conference on Robotics and Automation*, 2008.

[20] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *International Conference on Computer Vision*, 2007.

[21] D. Kerzel. A matter of design: No representational momentum without predictability. *Visual Cognition*, 9(1-2), 2002.

[22] Z. Kourtzi and N. Kanwisher. Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience*, 12(1), 2000.

[23] J. Lange, K. Georg, and M. Lappe. Visual perception of biological motion by form: a template-matching analysis. *Journal of vision*, 6(8), 2006.

[24] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *International Conference on Computer Vision and Pattern Recognition*, 2009.

[25] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5), 2011.

[26] N. Osaka, D. Matsuyoshi, T. Ikeda, and O. M. Implied motion because of instability in hokusai manga activates the human motion-sensitive extrastriate visual cortex: an fmri study of the impact of visual art. *Neuroreport*, 21(4), 2010.

[27] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

[28] S. Roth and M. J. Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1), 2007.

[29] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

[30] V. Sindhwani and S. S. Keerthi. Large scale semi-supervised linear svms. In *International SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[31] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, 2003.

[32] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. In *International Conference on Computer Vision*, 2001.

[33] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[34] C. Thurau and V. Hlavác. Pose primitive based human action recognition in videos or still images. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

[35] P. Turaga, A. Veeraraghavan, and R. Chellappa. Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

[36] O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3d motion estimation via mode finding on lie groups. In *International Conference on Computer Vision*, 2005.

[37] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. 2017.

[38] S. N. P. Vitaladevuni, V. Kellokumpu, and L. S. Davis. Action recognition using ballistic dynamics. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

[39] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *The Neural Information Processing Systems*, pages 613–621, 2016.

[40] C. Vondrick and A. Torralba. Generating the future with adversarial transformers.

[41] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3352–3361. IEEE, 2017.

[42] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *International Conference on Computer Vision and Pattern Recognition*, 2006.

[43] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(3), 2006.

[44] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

[45] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

[46] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4), 2006.

[47] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *European Conference of Computer Vision*, 2010.