AMDO 2012

Combining Skeletal Pose with Local Motion for Human Activity Recognition

Ran Xu[1], Priyanshu Agarwal[2], Suren Kumar[2],
Venkat N. Krovi[2], and Jason J. Corso[1]

[1]Computer Science and Engineering, [2]Mechanical and Aerospace Engineering,
State University of New York at Buffalo, NY, USA
{rxu2,priyansh,surenkum,vkrovi,jcorso}@buffalo.edu

Abstract. Recent work in human activity recognition has focused on bottom-up approaches that rely on spatiotemporal features, both dense and sparse. In contrast, articulated motion, which naturally incorporates explicit human action information, has not been heavily studied; a fact likely due to the inherent challenge in modeling and inferring articulated human motion from video. However, recent developments in data-driven human pose estimation have made it plausible. In this paper, we extend these developments with a new middle-level representation called *dynamic pose* that couples the local motion information directly and independently with human skeletal pose, and present an appropriate distance function on the dynamic poses. We demonstrate the representative power of dynamic pose over raw skeletal pose in an activity recognition setting, using simple codebook matching and support vector machines as the classifier. Our results conclusively demonstrate that dynamic pose is a more powerful representation of human action than skeletal pose.

Keywords: Human Pose, Activity Recognition, Dynamic Pose

1 Introduction

Bottom-up methods focusing on space-time motion have dominated the activity recognition literature for nearly a decade, e.g., [1–3], and have demonstrated good performance on challenging and realistic data sets like UCF Sports [4]. Although human activity is essentially articulated space-time motion, these methods avoid any need to explicitly model the articulated motion and rather focus on low-level processing to indirectly model the articulated space-time motion. Examples include local space-time interest points (STIP) [1], dense 3D gradient histograms (HOG) [2], and point trajectories [3], among many. More recent efforts have focused on mid-level representations that build on top of these elements, such as Niebles et al. [5] who model the local trajectories of STIP points and Gaidon et al. [6] who learn a time-series kernel to explicitly model repetitive motion in activities. All of these methods have limited transparency from a semantic point-of-view and rely on large amounts of available training data.

Alternatively, it seems reasonable to develop more semantically rich representations, such as those that more explicitly use articulated human models,

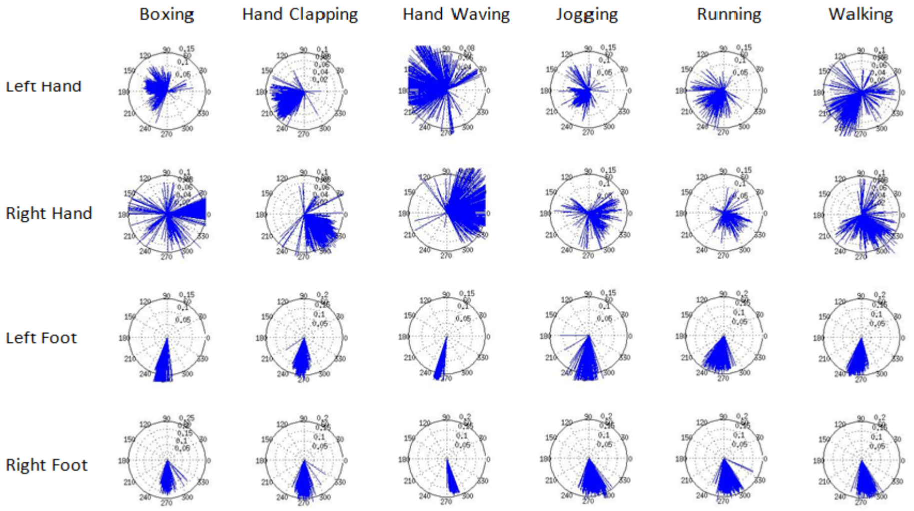


Fig. 1. Polar histograms of limb-extrema points in human pose for the six actions in the KTH data set [14]. Note the large similarity for each of the limb-extrema histograms (the rows) across the different actions. These data suggest that pose alone may not be suitable for human activity recognition.

to overcome these issues of transparency and scalability. However, fewer works have directly attempted to use human pose for activity recognition, e.g., [7–9], likely due to the challenging, unsolved nature of the pose estimation problem itself. Recent developments in pose estimation based on data-driven discriminative methods, such as Yang and Ramanan [10] who build a deformable parts model [11] and Bourdev and Malik [12] who learn *poselets* that are tightly coupled in 3D pose-space and local appearance-space, have paved the way for a reinvestigation into the suitability of pose for activity recognition. There has been limited success, yet, in the literature exploiting these better-performing pose estimation methods. Our early experimental evidence implies that pose alone may be insufficient to discriminate some actions. Figure 1, for example, contains polar histograms of limb-extrema for a variety of actions; in many cases the polar histograms across different actions are indistinguishable. Yao et al. [13] also evaluate whether pose estimation helps action recognition by randomly selecting appearance or pose feature in a random forest framework; they found no improvement after combination.

In this paper, we explore a unification of these independent research trends—motion- and pose-based human activity recognition—that addresses the limitations of each separate approach, to some degree. Our main contribution is a new representation called *dynamic pose*. The basic idea is to couple local motion information directly and independently with each *skeletal pose* keypoint. For example, the actions of sitting down in a chair and standing up from the

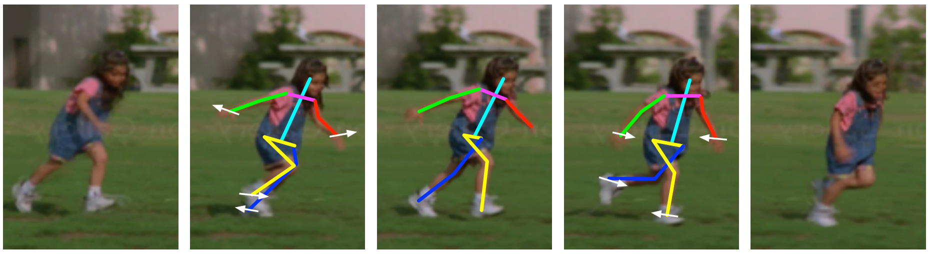


Fig. 2. Dynamic pose illustration: the girl is running to the right. The poses in the second and fourth frames are not very different, but when the joints are augmented with motion information, the distinction between them increases. In the second frame, her arms are expanding and in the fourth, they are contracting.

chair are distinct but the set of poses are nearly identical; however, the set of dynamic poses are indeed distinct and make the two actions distinguishable. We give a visual example of dynamic pose in Figure 2; in this example, we highlight the different stages of running, which have the arms contracting and expanding in one period of the activity giving two separate dynamic poses where only one skeletal pose would be apparent.

We adopt the state of the art pose estimation work of Yang and Ramanan [10] to compute a 13-point *skeletal pose* (i.e., one point for the head, one for the left shoulder, and so on). Then, at each of these 13-points, we compute the motion in a small cube around the point using a histogram of oriented space-time gradients (HoG3D) [2]. We apply the dynamic pose work in an activity recognition setting and propose a novel distance function on the dynamic poses to do so. Our experimental results conclusively demonstrate that dynamic pose outperforms skeletal pose on two benchmarks (UCF Sports [4] and KTH [14]).

Related Work. Some methods avoid the need to explicitly compute human pose and yet maintain a rich description of the underlying activity through templates. Exemplary methods along these lines are the space-time shapes [15], local optical-flow templates [16], the Action MACH represent that unifies many example templates into one based on spatiotemporal regularity [4], and the motion-orientation template representation in the action spotting framework [17]. These methods show good performance in some scenarios, but their ability to generalize to arbitrary settings is not clear, largely due to the difficulty in selecting the templates, which is frequently a manual process.

There has been some recent work pushing in the general direction of combining elements of human pose with local motion, but no work we are aware of couples a full skeletal pose with local motion. In contrast, the closest works we are aware of, [18–20], instead use bottom-up part-regions and/or foreground-silhouettes. Tran et al. [18] represent motion of body parts in a sparse quantized polar space as the activity descriptor, but discard the pose/part structure.

Brendel and Todorovic [19] build a codebook jointly on spatial-region appearance features (2D HOG) and motion features tied to these regions. They ultimately use a Viterbi algorithm on the sequence of codebook elements for activity recognition. The key idea is that the shape of the moving region—in this application it is primarily the human torso—will give information about the underlying activity. However, the work does not go far enough as to directly couple the motion information with full human pose and is hence limited in its direct semantic transparency. Lin et al. [20] attend to the moving human and separate the moving human foreground from the background, which they call *shape*. They couple dense motion features in the attended shape region of focus. Actions are then represented as sequences of prototypical shape-motion elements. Our proposed dynamic pose based model clearly differs from these approaches by incorporating local motion directly with full skeletal human pose, leveraging on impressive recent developments in human pose estimation [10].

2 Dynamic Pose for Human Activity Recognition

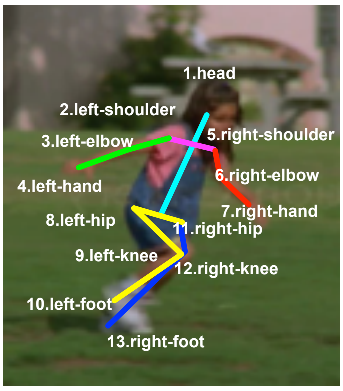


Fig. 3. Explanation of the 13 points on the skeletal pose. Section 2.1 explains how the local motion at each of these points is used to enhance the description of the pose for our dynamic pose representation.

Human pose is the core of our representation for activity recognition. Skeletal pose is represented by 13 joint points, as depicted in the Fig. 3. We use Yang and Ramanan's [10] articulated pose estimation method to extract the initial pose; their method outputs a bounding box for each human parts and we reduce these to the desired 13 joint points.

We define a local coordinate space for the skeletal pose to allow for scale-invariant inter-pose comparisons. Considering that human action can be represented as pose points' stretch and rotation relative to torso, as well as the whole body movement, we normalize the pose points by eliminating the scale variance and whole body movement. Denote the location of the 13 pose points as $L = \{l_1, ..., l_{13}\}$, in the original image coordinate space $l_i = \{x_i, y_i\}$ (refer to the indices in Fig. 3 to look up specific joint point identities in the following discussion). We anchor the scale-normalization using the extracted head point (x_1, y_1), as we have observed it to be the most stable of the extracted points with the method in use [10]. And, we normalize the spatial scale based on the maximum of the left lower or right lower leg; denote this as $d = \max\left(||l_{10} - l_9||, ||l_{13} - l_{12}||\right)$. The scale-normalized skeletal pose P is hence

$$p_i = \left(\frac{x_i - x_1}{d}, \frac{y_i - y_1}{d}\right) . \tag{1}$$

where $i = \{2, \ldots, 13\}$. At last we normalize the 24-dimensional pose vector norm to be 1. In the following sections, we introduce the dynamic pose formulation and then describe how we use dynamic pose for activity recognition.

2.1 Dynamic Pose

Although skeletal pose constrains the set of plausible activities a human may be engaging in, our experiments in looking at statistics of joint locations for different activities suggest that pose alone may not be sufficient for good activity recognition (see Figure 1 and Section 3). We hence extend the skeletal pose to incorporate local motion of the joint points, which we expect to add a richness to the pose-based representation for better descriptiveness. For example, jogging and running have similar skeletal poses, but pose with local motion information (e.g. magnitudes of local motion at feet) better encodes their differences.

To capture the local motion information of each skeletal joint point, we compute the histogram of oriented 3D gradients (HoG3D) [2] in the neighborhood around the point. HoG3D has demonstrated strong performance in activity and event classification as it encodes the statistics of local space-time articulation, giving a sense of the *texture* of the motion in the video. The local gradients are computed at multiple spatial and temporal scales in the neighboring vicinity of the joint point and binned according to their orientation and magnitude. Specifically, we define the scales to include $15 - 60$ pixels in each spatial direction and $5 - 20$ frames in time. Ultimately, at each joint-point, the numerous multiscale HoG3D vectors are summarized by a single local motion histogram; a codebook (150 entries) is built over the HoG3D vectors (separately for each joint-point) and then a histogram over the multiscale HoG3D vector-indices is calculated.

We now discuss computing the distance between two dynamic poses. The direct distance of the combined skeletal and local-motion distance is not plausible—for example, one can envision a case where two skeletal poses are quite different, but the local motions of the points are similar. In this contrived case, we expect the two dynamic poses to remain different. In other words, we seek a distance that is constrained by the skeletal pose and incorporates the local motion information only when needed. When two joint points have spatial distance smaller than some threshold, we compute the distance by comparing the histogram of HoG3D descriptor in that joint point; and when the spatial distance is larger than the threshold, we give a maximum distance instead.

Define the threshold that indicates small spatial distance as γ and the maximum distance value between the local motion features for a large spatial distance as β when we calculate the distance of two skeletal poses p and q (specific values for these parameters are discussed in the experiments). Let $d_i(p, q)$ define some appropriate distance function on joint point i in skeletal poses p and q; plausible options are Euclidean distance and cosine distance (since the poses are normalized). At each joint point i for pose p, denote the local space-time HoG3D histograms as $h_p(i)$. The distance $D(p, q)$ between two dynamic poses is

$$\delta(i) = \left\{ \begin{array}{ll} 1 - \min\left(h_p(i), h_q(i)\right) & \text{if } d_i(p,q) < \gamma \\ \beta & \text{if } d_i(p,q) \geq \gamma \end{array} \right\} \ ,$$

$$D(p,q) = \sum_{i=1}^{12} \delta(i) \ . \tag{2}$$

We discuss parameter settings in Section 3. The distance function is capable of clustering similar dynamic poses together, and separating different dynamic pose with similar spatial configuration, because the local motion histogram can characterize both joint motion orientation and speed.

2.2 Codebook-Based Dynamic Pose for Activity Recognition

To apply our dynamic pose to activity recognition, we use a bag-of-features approach. Incorporating the local motion information with the pose affords this simpler classifier than say a tracking-based one, which may be susceptible to noise in the frame-to-frame pose estimates. For skeletal pose, we construct a k-means codebook of 1000 visual words from the full set of 24-dimensional skeletal pose data. We use a similar technique to generate a 1000 word dynamic pose codebook, using the specified distance function in Eq. (2) instead of the standard Euclidean distance.

For classification we use many one-versus-one histogram intersection kernel SVMs [21]. Given labeled training data $\{(y_i, \mathbf{x_i})\}_{\mathbf{i=1}}^{\mathbf{N}}$, $x_i \in R^d$, where d is equal to the size of the codebook and N is the number of training data. For vectors $\mathbf{x_1}$ and $\mathbf{x_2}$, the histogram intersection kernel is expressed as:

$$k(\mathbf{x_1}, \mathbf{x_2}) = \sum_{i=1}^{d} \min(x_1(i), x_2(i)) \ . \tag{3}$$

Since we adopt a one-versus-one strategy, for a classification with c classes, $c(c-1)/2$ SVMs are trained to distinguish the samples of one class from another. Suppose we reduce the multi-class classification to binary classification, with $y_i \in \{-1, +1\}$. We minimize equation (4) in order to find a hyperplane which best separates the data.

$$\tau(\mathbf{w}, \xi) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=i}^{N} \xi_i \ , \tag{4}$$

$$\text{subject to:} \quad y_i((\mathbf{w} \cdot \mathbf{x_i}) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \ . \tag{5}$$

where $C > 0$ is the trade-off between regularization and constraint violation. In the dual formulation we maximize:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j k(\mathbf{x_1}, \mathbf{x_2}) \ , \tag{6}$$

$$\text{subject to:} \quad 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum \alpha_i y_i = 0 \ . \tag{7}$$

For a given SVM, suppose we have m support vectors $\mathbf{x_l} : l \in 1, 2, ...m$, for each histogram vector $\mathbf{x_i}$, m kernel computations are needed to score it:

$$v(\mathbf{x_i}) = \sum_{l=1}^{m} (\alpha_l y_l k(\mathbf{x_i}, \mathbf{x_l})) \ . \tag{8}$$

The final classification of a video $\mathbf{x_i}$ is the selected as the positive class in the one-versus-one SVM with the highest score, $v(\mathbf{x_i})$.

Table 1. Performance comparison between skeletal pose and dynamic pose on two standard benchmark datasets.

Method	KTH	UCF-Sports
BoP	76.39%	71.33%
BoDP	**91.2%**	**81.33%**

3 Experimental Evaluation

We test our algorithm on two benchmarks: KTH [14] and UCF-Sports [4]. The KTH dataset consists of six actions (Boxing, Hand-clapping, Hand-waving, Jogging, Running and Walking) performed about four times each by 25 subjects, for a total of 2396 sequences, including both indoor and outdoor scenes under varying scale. We follow the standard experimental setting described in [14], using person 02, 03, 05, 06, 07, 09, 10 and 22 as testing data and the other 16 people as training data. The UCF Sports dataset consists of ten sports actions (Diving, Golf-Swing, Kicking, Lifting, Riding Horse, Running, Skateboarding, Swing-Bench, Swing-SideAngle and Walk), totally 150 videos in unconstrained environments from wide range of scenes and viewpoints. We apply leave-one-out scheme for training and testing on UCF Sports.

We test using both Bag of Pose (BoP) and our Bag of Dynamic Pose methods (BoDP). As described in Section 2.2, we construct two 1000-dimensional codebooks from 10000 randomly sampled training features of skeletal pose as well as dynamic pose. As for the parameters that we use in the process of training codebook and encoding, we empirically set small distance threshold γ as 0.02 and max distance threshold β as 1.5. Table 1 summarizes the recognition accuracy of dynamic pose and skeletal pose on both benchmark datasets; the results demonstrate that dynamic pose, as a middle-level representation that incorporate both human pose skeletal and local motion information, is effective to represent articulated human activity.

Fig. 4 shows the visualization of dynamic pose codebook for the KTH dataset (scale-normalized skeletal poses are displayed only for simplicity); the ten samples displayed are the ten codebook centroids with the most support from the data set. We have observed that the first codebook centroid and the 9th one look very similar in the spatial coordinate, so we have inspected the video. We find they are corresponding to the 220th frame of video `person23_handclapping_d4_uncomp.avi` and the 134th frame of video `person25_handclapping_d4_uncomp.avi`. Fig. 5 shows the sequences of frames around the 1st and 9th codebook centroids. It is clear that, although the two canonical poses have great spatial similarity as depicted in the frames with pose skeleton on human, the motion is in the opposite direction. This visualization of codebook echoes our argument that dynamic pose is capable of capturing local motion information, which will definitely contribute to distinguishing different human action; it thus tends to improve classification of different human activities as our experimental evaluation will now show.

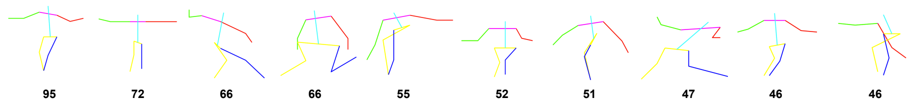


Fig. 4. The top ten canonical dynamic poses in the learned codebook. The poses are drawn after normalization and without any rendering of the local motion information at each pose point. The number of samples from the training set are given for each example.

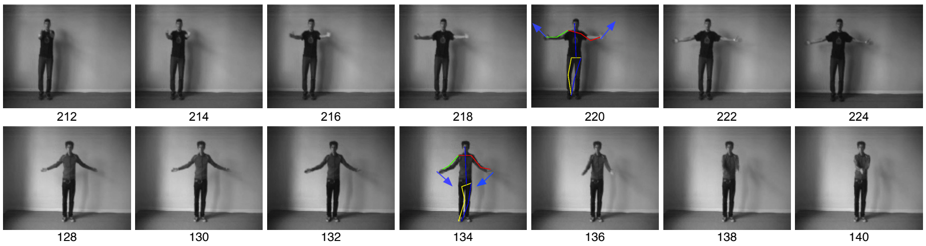
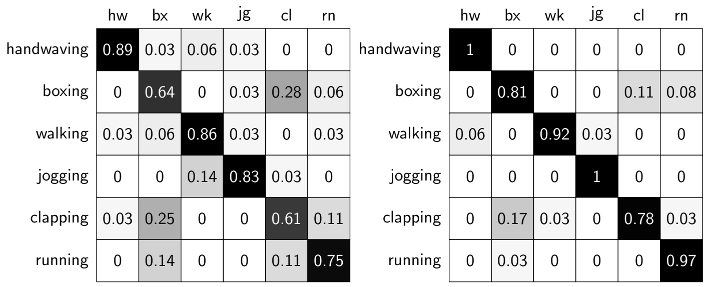


Fig. 5. The 1st and 9th dynamic pose codebook centroids visualized in video sequence. The first row corresponds to the 1st centroid: the canonical pose in 220th frame of video person23_handclapping_d4_uncomp.avi, and the second row corresponds to the 9th centroid: the canonical pose in 134th frame of video person25_handclapping_d4_uncomp.avi. The number below the video sequence is frame number, and the arrow indicates the direction of hand clapping.

For the KTH dataset, Fig. 6 shows the confusion matrices of BoP and BoDP, both classified by intersection kernel SVM over 1000 dimensional histograms. BoDP clearly outperforms BoP in every action class. Specifically, with BoP, 14% of jogging is misclassified as walking, whereas with BoDP jogging achieves 100% accuracy; another example is boxing, BoDP reduces 28% misclassification with hand clapping to 11%. These findings indicate that dynamic pose does capture enough articulated local motion information to distinguish spatially similar skeletal poses, e.g., similar feet position distribution among walking and jogging, and similar hand position distribution among boxing and hand-clapping. The overall accuracy increases from 76.4% to 91.2%. BoDP, as a middle level representation of articulated human action, has already achieves results comparable to the state-of-the-art in terms of simple actions.

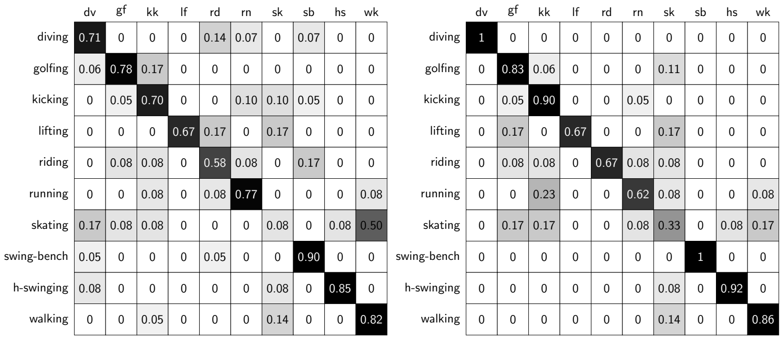
For UCF-Sports data set, the accuracy of BoP and BoDP are 71.33% and 81.33%, respectively. The confusion matrices in Fig. 7 show that, specifically, the classification result of the action "diving" increased from 71% to 100%, totally distinguish from "riding-horse", "running" and "skating"; and the classification accuracy of "kicking" increases from 70% to 90%, which shows that dynamic pose helps distinguish it from "running", "skating" and "swing-bench". The



	hw	bx	wk	jg	cl	rn
handwaving	0.89	0.03	0.06	0.03	0	0
boxing	0	0.64	0	0.03	0.28	0.06
walking	0.03	0.06	0.86	0.03	0	0.03
jogging	0	0	0.14	0.83	0.03	0
clapping	0.03	0.25	0	0	0.61	0.11
running	0	0.14	0	0	0.11	0.75

	hw	bx	wk	jg	cl	rn
handwaving	1	0	0	0	0	0
boxing	0	0.81	0	0	0.11	0.08
walking	0.06	0	0.92	0.03	0	0
jogging	0	0	0	1	0	0
clapping	0	0.17	0.03	0	0.78	0.03
running	0	0.03	0	0	0	0.97

Fig. 6. Confusion Matrix Comparison over BoP(Left) and BoDP(Right)



	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	0.71	0	0	0	0.14	0.07	0	0.07	0	0
golfing	0.06	0.78	0.17	0	0	0	0	0	0	0
kicking	0	0.05	0.70	0	0	0.10	0.10	0.05	0	0
lifting	0	0	0	0.67	0.17	0	0.17	0	0	0
riding	0	0.08	0.08	0	0.58	0.08	0	0.17	0	0
running	0	0	0.08	0	0.08	0.77	0	0	0	0.08
skating	0.17	0.08	0.08	0	0	0	0.08	0	0.08	0.50
swing-bench	0.05	0	0	0	0.05	0	0	0.90	0	0
h-swinging	0.08	0	0	0	0	0	0.08	0	0.85	0
walking	0	0	0.05	0	0	0	0.14	0	0	0.82

	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	1	0	0	0	0	0	0	0	0	0
golfing	0	0.83	0.06	0	0	0	0.11	0	0	0
kicking	0	0.05	0.90	0	0	0.05	0	0	0	0
lifting	0	0.17	0	0.67	0	0	0.17	0	0	0
riding	0	0.08	0.08	0	0.67	0.08	0.08	0	0	0
running	0	0	0.23	0	0	0.62	0.08	0	0	0.08
skating	0	0.17	0.17	0	0	0.08	0.33	0	0.08	0.17
swing-bench	0	0	0	0	0	0	0	1	0	0
h-swinging	0	0	0	0	0	0	0.08	0	0.92	0
walking	0	0	0	0	0	0	0.14	0	0	0.86

Fig. 7. Confusion Matrix Comparison over BoP(Left) and BoDP(Right) on UCF-Sports Dataset

experiments demonstrate that dynamic pose is also effective in dealing with the complex articulations in UCF Sports.

4 Conclusion and Future Work

In conclusion, we propose a new middle level representation of articulated human action—dynamic pose—that adds local motion information to skeletal joint points. The basic premise behind dynamic pose is that skeletal pose alone is insufficient for distinguishing certain human actions, those which have similar spatial distributions of limb points over the course of an action. We have implemented our representation in an activity recognition setting using bag of features with kernel intersection SVM as the base classifier. Our experiments conclusively indicate that dynamic pose is a capable middle-level representation of articulated human motion. In the future, we plan to combine our dynamic pose with global context.

Acknowledgements This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), and the DARPA Mind's Eye program (W911NF-10-2-0062).

References

1. Laptev, I.: On space-time interest points. IJCV (2005)
2. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
3. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR. (2011) 3169–3176
4. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008)
5. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV. (2010)
6. Gaidon, A., Harchaoui, Z., Schmid, C.: A time series kernel for action recognition. In: BMVC. (2011)
7. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: ICCV. (2007)
8. Ramanan, D., Forsyth, D.A.: Automatic annotation of everyday movements. In: NIPS. (2003)
9. Shakhnarovich, G., Viola, P., Darrell, T.: Fast Pose Estimation with Parameter-Sensitive Hashing. In: ICCV. (2003)
10. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI **32** (2010) 1627–1645
12. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)
13. Yao, A., Gall, J., Fanelli, G., Gool, L.V.: Does human action recognition benefit from pose estimation? In: BMVC. (2011)
14. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR. (2004)
15. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. TPAMI **29**(12) (2007) 2247–2253
16. Essa, I., Pentland, A.: Coding, analysis, interpretation and recognition of facial expressions. TPAMI **19**(7) (1997) 757–763
17. Derpanis, K.G., Sizintsev, M., Cannons, K., Wildes, R.P.: Efficient action spotting based on a spacetime oriented structure representation. In: CVPR. (2010)
18. Tran, K.N., Kakadiaris, I.A., Shah, S.K.: Modeling motion of body parts for action recognition. In: BMVC. (2011)
19. Brendel, W., Todorovic, S.: Activities as time series of human postures. In: ECCV. (2010)
20. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: ICCV. (2009)
21. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR. (2008)