An approach to pose-based action recognition

Chunyu Wang[1], Yizhou Wang[1], and Alan L. Yuille[2]

[1]Nat'l Engineering Lab for Video Technology, Key Lab. of Machine Perception (MoE), Sch'l of EECS, Peking University, Beijing, 100871, China

{wangchunyu, Yizhou.Wang}@pku.edu.cn

[2]Department of Statistics, University of California, Los Angeles (UCLA), USA

yuille@stat.ucla.edu

Abstract

We address action recognition in videos by modeling the spatial-temporal structures of human poses. We start by improving a state of the art method for estimating human joint locations from videos. More precisely, we obtain the K-best estimations output by the existing method and incorporate additional segmentation cues and temporal constraints to select the "best" one. Then we group the estimated joints into five body parts (e.g. the left arm) and apply data mining techniques to obtain a representation for the spatial-temporal structures of human actions. This representation captures the spatial configurations of body parts in one frame (by spatial-part-sets) as well as the body part movements(by temporal-part-sets) which are characteristic of human actions. It is interpretable, compact, and also robust to errors on joint estimations. Experimental results first show that our approach is able to localize body joints more accurately than existing methods. Next we show that it outperforms state of the art action recognizers on the UCF sport, the Keck Gesture and the MSR-Action3D datasets.

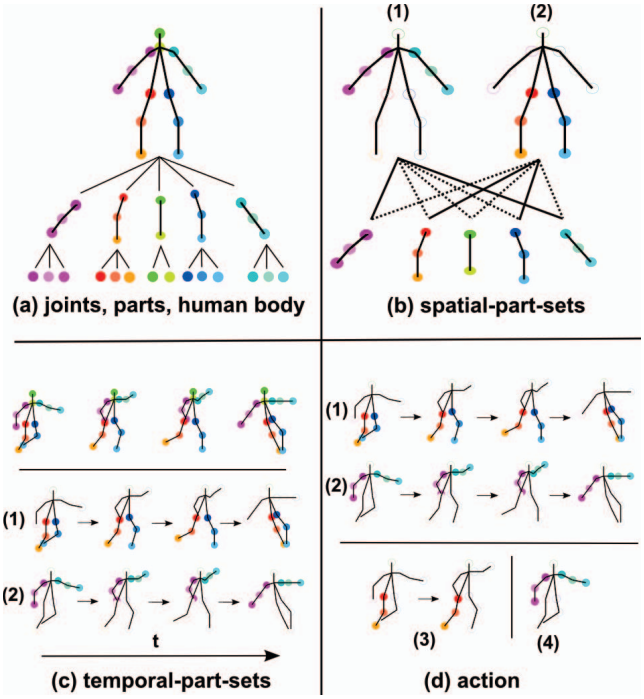

Figure 1. Proposed action representation. **(a)**A pose is composed of 14 joints at the bottom layer, which are grouped into five body parts in the layer above; **(b)**shows two spatial-part-sets which combine frequently co-occurring configurations of body parts in an action class. **(c)**temporal-part-sets are co-occurring sequences of evolving body parts. (e.g. evolving left and right legs compose a temporal-part-set(1)). **(d)**action is represented by a set of spatial-part-sets(4) and temporal-part-sets(1-3).

1. Introduction

Action recognition is a widely studied topic in computer vision. It has many important applications such as video surveillance, human-computer interaction and video retrieval. Despite great research efforts, it is far from being a solved problem; the challenges are due to intra-class variation, occlusion, and other factors.

Recent action recognition systems rely on low-level and mid-level features such as local space-time interest points (e.g. [14][19]) and dense point trajectories (e.g. [20]). Despite encouraging results on several datasets, they have limited discriminative power in handling large and complex data because of the limited semantics they represent [18].

Representing actions by global templates (e.g. [7][2][1]) has also been explored. Efros et al. [7] compare optical flow based features against templates stored in databases

and classify them by the k-nearest-neighbor classifier. The features are computed from figure centric videos obtained by tracking, which are sometimes unreliable. Bobick et al. [2] construct motion templates by computing motion energy/history images. Blank et al. [1] represent actions as space-time shapes and extract space-time features such as local space-time saliency. These types of methods lack the flexibility to handle challenging cases such as dynamic backgrounds, camera movement and intra-class appearance variation, which limit their performance on real videos.

An alternative line of work represent actions by sequences of poses in time (e.g. [4][26]), where poses refer to spatial configurations of body joints. These representations conform to studies of how human understand actions [3]. Some of these work use poses obtained by motion capture systems [4][26]. However, pose-based action recognition can be very hard because of the difficulty to estimate high quality poses from action videos, except in special cases (e.g., static cameras and simple backgrounds).

In this paper we present a novel pose-based action recognition approach which is effective on some challenging videos. We first extend a state of the art method [27] to estimate human poses from action videos. Given a video, we first obtain best-K pose estimations for each frame using the method of [27], then we infer the best poses by incorporating segmentation and temporal constraints for all frames in the video. We experimentally show that this extension localizes body joints more accurately.

To represent human actions, we first group the estimated joints into five body parts (e.g. left arm, see Figure 1.a). We then apply data mining techniques in the spatial domain to obtain sets of distinctive co-occurring spatial configurations(poses) of body parts, which we call spatial-part-sets. Similarly, in the temporal domain, we obtain sets of distinctive co-occurring pose sequences of body parts, which we call temporal-part-sets (e.g. the left arm going up is usually coupled with the right arm going up in "lifting" actions). These part-sets are obtained using an efficient contrast mining algorithm [6]. For test videos, we first detect these part-sets from the estimated poses then represent the videos by histograms of the detected part-sets. We classify the videos into actions using support vector machines (SVMs)[5].

To summarize, the proposed representation has three advantages. (i) It is interpretable, because we decompose poses into parts, guided by human body anatomy, and represent actions by the temporal movements of these parts. This high interpretability enables us to efficiently spot why and where the model may fail. (ii) It is compact. Only 14 joint locations are encoded for each frame. This has advantages, compared to high dimensional models (e.g. bag of low-level features), because it helps prevent overfitting when training action classifiers. (iii) It is robust to variations, because our part-sets are local and partially ambiguous joint locations have limited influence to the final representation. This boosts action recognition performance compared with holistic pose features. We demonstrate these advantages by showing that our proposed method outperforms state of the art action recognizers on the UCF sport, the Keck Gesture and the MSR-Action3D datasets.

The paper is organized as follows. Section 2 reviews the related work. Section 3, 4 introduces pose estimation and action representation, respectively. Section 5 shows experiment results. Conclusion is in section 6.

2. Related Work

We briefly review the pose-based action recognition methods in literature. In [4][26], body joints are obtained by motion capture systems or segmentation. Then, the joints are tracked over time and the resulting trajectories are used as input to the classifiers. Xu et al[25] propose to automatically estimate joint locations from videos, and use joint locations coupled with motion features for action recognition. Modest joint estimation can degrade the action recognition performance as shown in experiments.

Given the difficulty of pose estimation, some approaches adopt implicit poses. For example, Ijuzker et al. [10] extract oriented rectangular patches from images and compute spatial histograms of oriented rectangles as features. Maji et al. [16] use "poselet" activation vector to implicitly capture human poses. However, implicit pose representations are difficult to relate to body parts, and so are it is hard to model meaningful body part movements in actions.

Turning to feature learning algorithms, the strategy of combining frequently co-occurring primitive features into larger compound features has been extensively explored (e.g. [22][9][21]). Data mining techniques such as *Contrast Mining* [6] have been adopted to fulfill the task. However, people typically use low-level features such as optical flow [22], and corners [9] instead of high-level poses. Our work is most related to [21] which groups joint locations into actionlet ensembles. But our work differs from [21] in two respects. First, we do not train SVMs for individual joints because they may carry insufficient discriminative information. Instead, we use body parts as building blocks as they are more meaningful and compact. Secondly, we model spatial pose structures as well as temporal pose evolutions, which are neglected in [21].

3. Pose Estimation in Videos

We now extend a state of the art image-based pose estimation method [27] to video sequences. Our extension can localize joints more accurately, which is important for achieving good action recognition performance. We first briefly describe the initial frame-based model in section 3.1, then present the details of our extension in section 3.2.

3.1. Initial Frame-based Pose Estimation

A pose P is represented by 14 joints J_i: head, neck, (left/right)-hand/elbow/shoulder/hip/knee/foot. The joint J_i is described by its label l_i (e.g. neck), location (x_i, y_i), scale s_i, appearance f_i, and type m_i (defined by the orientation of the joint), i.e. $J_i = (l_i, (x_i, y_i), s_i, f_i, m_i)$. The score for a particular configuration P in image I is defined by:

$$S(I,P) = c(m) + \sum_{J_i \in P} \omega_i \cdot f(I, J_i) + \sum_{i,j \in E} \omega_{ij} \cdot u(J_i, J_j) \quad (1)$$

where $c(m)$ captures the compatibility of joint types; the appearance $f(I, J_i)$ is defined by HoG features extracted for joint J_i; the edge set E defines connected joints, and $\omega_{ij} \cdot u(J_i, J_j)$ captures the deformation cost of connected joints. The deformation feature u is defined by $u(J_i, J_j) = [dx, dx^2, dy, dy^2]$, where $dx = x_i - x_j$. The weights ω are learned from training data. The inference can be efficiently performed by dynamic programming. Please see [27] for more details of this approach.

The estimation results of the model are not perfect. The reasons are as follows. Firstly, the learnt kinematic constraints tend to bias estimations to dominating poses in training data, which decreases estimation accuracy for rare poses. Secondly, for computational reasons, some important high-order constraints are ignored which may induce the "double-counting" problem (where two limbs cover the same image region). However, looking at 15-best poses returned by the model for each frame, we observe a high probability that the "correct" pose is among them. This motivates us to extend this initial model to automatically infer the correct pose from the K-best poses, using temporal constraints in videos. Similar observations have been made in recent work [12]. We differ from [12] by exploiting richer temporal cues in videos to fulfill the task.

3.2. Video-based Pose Estimation

The inputs to our model are the K-best poses of each frame I^t returned by[27]: $\{P_j^t | j = 1...K, t = 1...L\}$. Our model selects the "best" poses $(P_{j_1}^1, ..., P_{j_L}^L)$ for the L frames by maximizing the energy function E_P:

$$j^* = \underset{(j_1,...,j_L)}{\operatorname{argmax}} E_P(I^1, ..., I^L, P_{j_1}^1, ..., P_{j_L}^L)$$
$$E_P = \sum_{i=1}^{L} \phi(P_{j_i}^i, I^i) + \sum_{i=1}^{L-1} \psi(P_{j_i}^i, P_{j_{i+1}}^{i+1}, I^i, I^{i+1}) \quad (2)$$

Where $\phi(P_{j_i}^i, I^i)$ is a unary term that measures the likelihood of the pose and $\psi(P_{j_i}^i, P_{j_{i+1}}^{i+1}, I^i, I^{i+1})$ is a pairwise term that measures the appearance, and location consistency of the joints in consecutive frames.

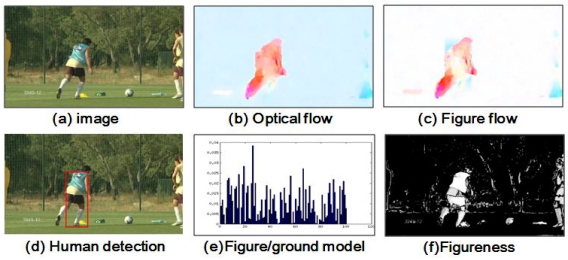

Figure 2. Steps for computing figure/ground color models.

3.2.1 Unary Term

A pose P essentially segments a frame into figure/ground pixel sets I_F/I_B. Hence we compute the unary term by explaining all pixels in the two sets. In particular, we group the 14 joints of pose P into five body parts(head, left/right arm, left/right leg) by human anatomy, i.e. $P = \{p_1, ..., p_5\}$, $p_j = \{J_{j_k} | k = 1...z_j\}$. z_j is the number of joints in part p_j. Each joint J_i covers a rectangular image region I_{J_i} centered at (x_i, y_i) with side length s_i; accordingly, each part p_j covers image regions $I_{p_j} = \cup_{J_i \in p_j} I_{J_i}$; image regions covered by the five body parts constitute figure regions $I_F = \cup_{i=1}^5 I_{p_i}$, and the remaining regions constitute the ground regions $I_B = I - I_F$. We measure the plausibility of pose P by "explaining" every pixel in I_F and I_B with pre-learnt figure/ground color distributions K_F and K_B:

$$\phi(P, I) = \prod_{x \in I_F} K_F(x) \cdot \prod_{x \in I_B} K_B(x) \quad (3)$$

We automatically learn the figure/ground distributions K_F and K_B for each video. Essentially, we create a rough figure/ground segmentation of the frames in the video, from which we learn the figure/ground color distributions(color histogram). We propose two approaches to detect figure regions. We first apply a human detector[8] on each frame to detect humans as figure regions (see Figure 2.d). However, the human detector cannot detect humans in challenging pose. Hence, we also use optical flow to detect moving figures (see Figure 2.b-c). We assume the motion field M contains figure motion F and camera motion C, i.e. $M = F + C$. Without loss of generality, we assume that the majority of the observed motion is caused by camera motion. Since the camera motion is rigid, C is low rank. We recover F and C from M by rank minimization using the method described in [23]. We consider regions whose figure motion F are larger than a threshold as figure regions. See Figure 2.c. We learn figure color distributions K_F from figure pixels detected by the human detector and by optical flow. Similarly, ground color distribution is learnt from remaining pixels of the video.

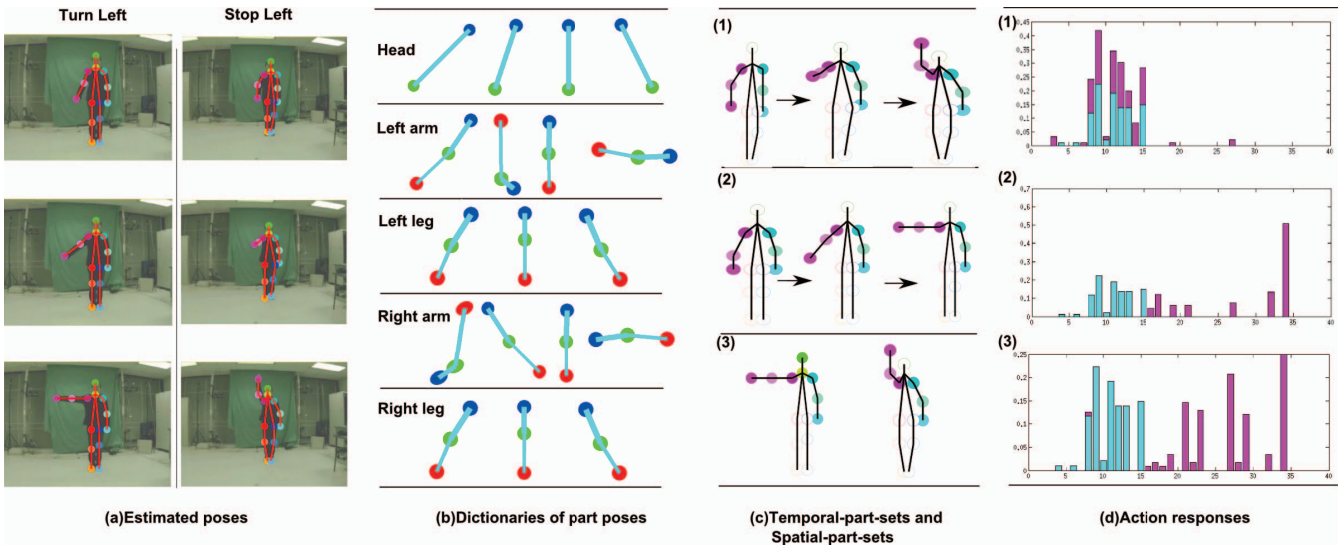

Figure 3. Overall framework for action representation. **(a)**we start by estimating poses for videos of the two action classes, i.e. *turn left*(left-column) and *stop-left*(right column). **(b)**then we cluster poses of each body part in training data and construct a part pose dictionary as described in Section 4.1. The blue, green and red dots in arms are the joints of shoulders, elbows and hands. Similarly, they are the joints of hip, knee and foot for legs. **(c)**we extract temporal-part-sets(1-2) and spatial-part-sets(3) for the two action classes as described in Section 4.2-4.3. **(d)**we finally represent actions by histograms of spatial- and temporal- part-sets. (1)shows two histograms of two different humans performing the same action. The histograms are similar despite intra-class variations. (2)and(3) show the histograms of *turn left* vs. *stop-left*, and *turn left* vs. *stop-both*, respectively. The histograms differ a lot although they share portions of poses.

3.2.2 Temporal Consistency

$\psi(P^i, P^{i+1}, I^i, I^{i+1})$ captures appearance and location coherence of the joints in consecutive frames. We measure the appearance coherence by computing Kullback-Leibler divergence of the corresponding joints' color distributions:

$$E_a(P^i, P^{i+1}) = -\sum_{k=1}^{5} \sum_{J \in p_k} KL(f_J^i, f_J^{i+1}) \qquad (4)$$

f_J^i is the color histogram computed for the rectangular image region around joint J^i. For location coherence, we compute the Euclidean distance (discretized into 10 bins) between the joints in consecutive frames:

$$E_l(P^i, P^{i+1}) = -\sum_{k=1}^{5} \sum_{J \in p_k} d((x_J^i, y_J^i), (x_J^{i+1}, y_J^{i+1})) \qquad (5)$$

Finally we define ψ as the sum of E_a and E_l.

3.2.3 Inference

The global optimum of the model can be efficiently inferred by dynamic programming because of its chain structure(in time). In implementation, we first obtain the 15-best poses by[27] for each frame in the video. Then we identify the best poses for all frames by maximizing the energy function(see equation 2).

4. Action Representation

We next extract representative spatial/temporal pose structures from body poses for representing actions. For spatial pose structures, we pursue sets of frequently co-occurring spatial configurations of body parts in a single frame, which we call the spatial-part-set, $sp_i = \{p_{j_1}, ..., p_{j_{n_i}}\}$. For temporal pose structures, we pursue sets of frequently co-occurring body part sequences $al_i = (p_{j_1}, ..., p_{j_{m_i}})$, which we call temporal-part-sets, $tp_i = \{al_{k_1}, ..., al_{k_{l_i}}\}$. Note that body part sequence al_i captures the temporal pose evolution of a single body part (e.g. left arm going up). We represent actions by histograms of activating spatial-part-sets and temporal-part-sets. See Figure 3 for the overall framework of the action representation.

4.1. Body Part

A body part p_i is composed of z_i joint locations $p_i = (x_1^i, y_1^i, ..., x_{z_i}^i, y_{z_i}^i)$. We normalize p_i to eliminate the influence of scale and translation. We first anchor p_i by the head location (x_1^1, y_1^1) as it is the most stable joint to estimate. Then we normalize its scale by head length d, $p_i = \frac{p_i - (x_1^1, y_1^1)}{d}$.

We learn a dictionary of pose templates $V_i = \{v_i^1, v_i^2, ..., v_i^{k_i}\}$, for each body part by clustering the poses of training data. k_i is the dictionary size. Each template pose represents a certain spatial configuration of body parts(See Figure 3.b). We quantize all body part poses p_i

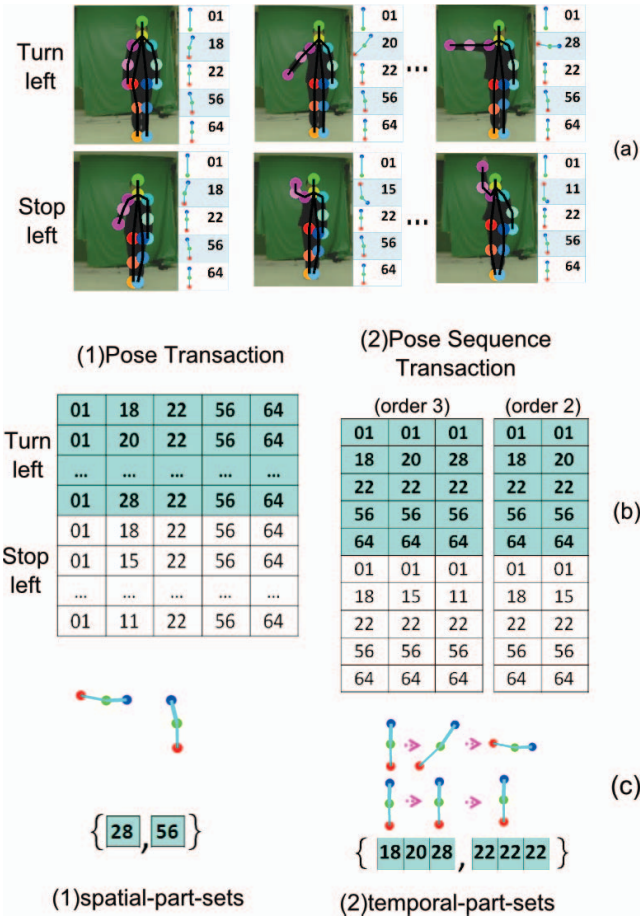

Figure 4. Spatial-part-sets and temporal-part-sets pursued by contrast mining techniques. **(a)**shows estimated poses for videos of *turn-left* and *stop-left* actions. The numbers in the right of each figure are indexes of quantized parts in the dictionaries. **(b)**shows two transaction databases for mining spatial-part-sets(1) and temporal-part-sets(2) respectively. Each row in(1) is a transaction composed by five indexes of quantized body parts. Each row in(2) is an item, i.e. sub-sequences of body parts of order three(left) and two(right). All items in one video(e.g. top five rows) compose a transaction. **(c)**.(1) shows one pursued *spatial-part-set* which is a typical configuration of "turn-left" action. (c).(2)shows one typical *temporal-part-set* of "turn-left" action.

by the dictionaries to consider pose variations. Quantized poses are then represented by the five indexes of the templates in the dictionaries.

4.2. Spatial-part-sets

We propose *spatial-part-sets* to capture spatial configurations of multiple body parts: $sp_i = \{p_{j_1}, ..., p_{j_{n_i}}\}, 1 \leq n_i \leq 5$. See Figure1.b for an example. The compound spatial-part-sets are more discriminative than single body parts. The ideal *spatial-part-sets* are those which occur frequently in one action class but rarely in other classes (and hence have both representative and discriminative power). We obtain sets of *spatial-part-sets* for each action class using *Contrast Mining* techniques[6].

We use the notation from [6] to give a mathematical definition of *contrast mining*. Let $I = \{i_1, i_2, ..., i_N\}$ be a set of N items. A *transaction* T is defined as a subset of I. The transaction database D contains a set of transactions. A subset S of I is called a *k-itemset* if $||S|| = k$. If $S \subseteq T$, we say the transaction T contains the itemset S. The support of S in a transaction database D is defined to be $\rho_S^D = \frac{count_D(S)}{||D||}$, where $count_D(S)$ is the number of transactions in D containing S. The growth rate of an itemset S from one dataset D_+ to the other dataset D_- is defined as:

$$T_S^{D_+ \to D_-} = \begin{cases} 0 & \text{if } \rho_S^{D_-} = \rho_S^{D_+} = 0 \\ \infty & \text{if } \rho_S^{D_-} \neq 0, \rho_S^{D_+} = 0 \\ \frac{\rho_S^{D_-}}{\rho_S^{D_+}} & \text{if } \rho_S^{D_-} \neq 0, \rho_S^{D_+} \neq 0 \end{cases} \quad (6)$$

An itemset is said to be a η-emerging itemset from D_+ to D_- if $T_S^{D_+ \to D_-} > \eta$.

We now relate the notations in contrast mining to our problem of mining spatial-part-sets. Recall that the poses are quantized and represented by the five indexes of pose templates. Each pose template is considered as an item. Hence the union of the five dictionaries V composes the item set, $V = V_1 \cup V_2 ... \cup V_5$. A pose P represented by five pose templates is a transaction. All poses in the training data constitute the transaction database D(See Figure 4.b). We now mine η-emerging itemsets, i.e. spatial-part-sets, from one action class to the others. See Figure 4 for an illustration of the mining process.

We pursue sets of *spatial-part-sets* for each pair of action classes y_1 and y_2. We first use transactions of class y_1 as positive data D_+, and transactions of y_2 as negative data D_-. The itemsets, whose support rates for D_+ and growth rates from D_- to D_+ are above a threshold, are selected. Then we use y_2 as positive data and y_1 as negative data and repeat the above process to get another set of itemsets. We combine the two sets as *spatial-part-sets*. We need to specify two threshold parameters, i.e. the support rate ρ and the growth rate η. By increasing the support rate, we guarantee the representative power of the spatial-part-sets for the positive action class. By increasing the growth rate, we guarantee the spatial-part-sets' discriminative power. The mining task can be efficiently solved by [6].

4.3. Temporal-part-sets

We propose temporal-part-sets to capture joint pose evolution of multiple body parts. We denote pose sequences of body parts as $al_i = (p_{j_1}, ..., p_{j_{n_i}})$, where n_i is the order of the sequence. We mine a set of frequently co-occurring pose sequences, which we call temporal-part-sets,

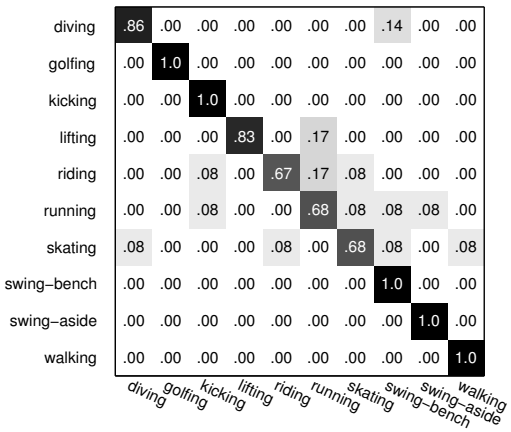

Figure 5. The confusion matrix of our proposed approach on the UCF Sport Dataset.

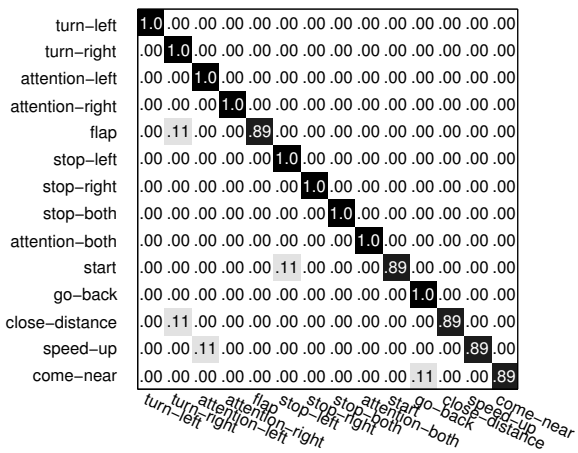

Figure 6. The confusion matrix of our proposed approach on the Keck Gesture Dataset(training subset).

$tp_i = \{al_{j_1}, ..., al_{j_{n_j}}\}$ (e.g.left arming going up is usually coupled with right arm going up in "lifting" action). We also use *contrast mining* to mine *temporal-part-sets*.

In implementation, for each of the five pose sequences $(p_i^1, ..., p_i^L)$ of a training video with L frames, we generate a set of sub-sequences of order n, i.e. $\{(p_i^k, ..., p_i^{k+n-1})|1 \leq k \leq L-n+1\}$. We set $n = \{2, 3, ..., L\}$. Each sub-sequence is considered as an item, all the sub-sequences of the video compose a transaction, and the transactions of all videos compose the transaction database. We mine a set of co-occurring sub-sequences for each pair of action classes as spatial-part-sets mining. See Figure 4 for illustration of the mining process.

4.4. Classification of Actions

We use the bag-of-words model to leverage spatial-part-sets and temporal-part-sets for action recognition. In the off-line mode, we pursue a set of part-sets for each pair of action classes. Then, for an input video, we first estimate poses and then quantize them using the proposed method. We count the presence of part-sets in the quantized poses and form a histogram as the video's features(see Figure 3.d). We train one-vs-one intersection kernel SVMs for each pair of classes. In the classification stage, we apply the learnt multiple one-vs-one SVMs on the test video and assign it the label with maximum votes.

5. Experiments

We evaluate our approach on three datasets: the UCF sport [17], the Keck Gesture [11] and the MSR-Action3D [15]. We compare it with two baselines and the state of the art methods. For the UCF sport and Keck Gesture datasets, we estimate poses from videos by our proposed approach. We report performance for both pose estimation and action recognition. For the MSR-Action3D dataset, we bypass pose estimation and use the provided 3D poses (because the video frames are not provided) to recognize actions. We also evaluate our approach's robustness to ambiguous poses by perturbing the joint locations in MSR-Action3D.

5.1. Datasets

The UCF sport dataset [17] contains 150 videos of ten human actions. We use the standard leave-one-out criteria for evaluation. The Keck gesture dataset [11] contains 14 different gesture classes. There are three or four persons performing the same gestures in the training and testing datasets. We use leave-one-person strategy as [11]. The MSR-Action3D dataset [15] contains 20 actions, with each action performed three times by ten subjects. We use the cross-subject test setting as in [15][21]

5.2. Comparison to two baselines

We compare our proposed representation with two baselines: holistic pose features and local body part based features. A holistic pose feature is a concatenated vector of 14 joint locations. We cluster holistic pose features using the k-means algorithm and obtain a prototype dictionary of size 600. We describe each video with "bag-of-holistic pose features" and use intersection kernel SVM classifier.

For local body part based features, we compute a separate pose dictionary for each body part, extract "bag-of-body part" features, and concatenate them into a high dimensional vector. We use the intersection kernel SVM classifier. We set dictionary sizes to $(8, 25, 25, 25, 25)$ for the five body parts by cross validation. The approach is not sensitive to dictionary sizes. Generally, the performance improves as dictionary size increases but begins to degrade after exceeding a threshold.

For our *part-sets* based representation, the support rate and growth rate are set to 10% and 3 by cross validation.

Table 1. Comparison of our method with the state-of-the-arts on the UCF sport dataset. Sadanand et al [18] achieves the highest recognition rate. But they manually constructed their *Action bank* which makes it incomparable to automatic methods like ours.

Approach	Year	Accuracy(%)
Kovashka[13]	2010	87.27
Wang[20]	2011	88.20
Wu[24]	2011	91.30
Sadanand[18]	2012	95.00
Our method		**90.00**

Table 2. Comparison of our method with the state-of-the-arts on the MSR-Action3D dataset.

Approach	Year	Accuracy(%)
Li[15]	2010	74.70
Wang[21]	2012	88.20
Our method		**90.22**

We obtain 1700, 513, 1630 spatial-part-sets and temporal-part-sets in total for the UCF sport, Keck Gesture and MSR-Action3D datasets respectively. We use intersection kernel SVM classifiers as described in section 4.4.

On the UCF sport dataset, the holistic pose features and the local body part based features get 69.33% and 78.67% accuracy respectively. Our approach achieves 90% accuracy which is a big improvement over the baselines. Figure 5 shows the confusion matrix of our approach.

On the Keck Gesture dataset, the holistic pose features achieve 76.30%/72.30% accuracy on training/testing subsets, respectively, while local body part based features get 87.30%/81.50%. Our approach achieves highest accuracy with 97.62% and 93.40%. The confusion matrix (see Figure 6) on the Keck Gesture dataset shows that our approach can correctly differentiate almost all classes even for *go back* and *come near*. Note that *go back* and *come near* actions have very similar poses but in reverse temporal order. Temporal-part-sets play an important role here.

On MSR-Action3D, the three methods get 64.84%, 71.22% and 90.22% respectively. Our approach is shown to boost the performance for both 2D and 3D poses.

5.3. Comparison to state-of-the-art performance

Table 1 summarizes the state-of-the-art performance on the UCF sport dataset. We outperform [13]and [20], and achieve comparable performance to [24]. Sadanand's action bank [18] achieves the highest recognition rate. But their action bank is constructed by manually selecting frames from training data, so it is not appropriate to compare our fully automatic method to their's.

To our best knowledge, the best results on the Keck Gesture dataset are 95.24% and 91.07% [11] accuracy on training and testing subsets respectively. We outperform it with 97.62% and 93.40%. In particular, the most confusing classes in [11], i.e. *come near* and *go back*, are well handled in our representation. See Figure 6.

Table 3. Comparison of action recognition using poses estimated by [27] and by our method. The numbers in bold are the results of our method and numbers above are the results of [27].

Dataset	Holistic	Body Part	Our approach
UCF sport	60.67%	70.00%	85.33%
	69.33%	**78.67%**	**90.00%**
Keck Gesture	56.35%	72.22%	82.54%
	76.30%	**87.30%**	**97.62%**

For the MSR-Action3D dataset, we achieve 90.22% accuracy and outperform the state of the arts [21] by about 2%. See Table 2 for results on this dataset. The results validate our representation's applicability on 3D poses which can be easily obtained by depth sensors.

5.4. Evaluation for Pose Estimation

We also evaluated the proposed pose estimation method. We annotated 352 frames, which are randomly sampled from the Keck Gesture dataset, for evaluation. We use standard evaluation protocol based on the probability of a correct pose(PCP)[27], which measures the percentage of correctly localized body parts. These experiments gave performance of 88.07% for [27] and 92.39% for our method.

We also evaluated the pose estimation method in the context of action recognition. Table 3 compares the action recognition accuracy using poses obtained by different pose estimation methods. The tables shows that using the poses obtained by our method (which are more accurate) does improve the action recognition performance compared to using the poses obtained by[27]

5.5. Evaluation on Model Robustness

Under challenging situations – such as cluttered background and video quality degradation – the estimation of joint locations can be very ambiguous. We evaluated the robustness of our proposed method as follows. We synthesized a set of data by randomly perturbing up to 20% of the 3D joint locations in the MSR-Action3D dataset. Figure 7 shows the experiment results. The performance of *holistic pose features* drop dramatically as the perturbation gets severe, which is expected since the accuracy of joint locations has large impact on holistic pose features. However, our method outperforms the two baseline methods even with perturbations of more than 10% of the joint locations.

6. Conclusion

We proposed a novel action representation based on human poses. The poses were obtained by extending an existing state-of-the-art pose estimation algorithm. We apply

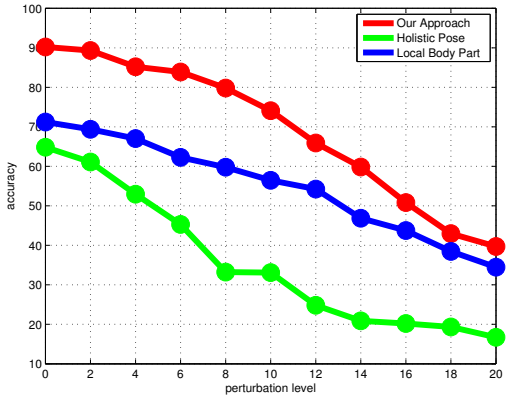

Figure 7. Recognition accuracy on MSR-Action3D when $0\% - 20\%$ of joint locations are perturbed.

data mining techniques to mine spatial-temporal pose structures for action representation. We obtained state-of-the-art results on three datasets. Another advantage of our method is that it is interpretable, compact and computationally efficient. In future work, we intend to exploit this interpretability and extend our method to more challenging data, including dealing with significant occlusions and missing parts, and to the recovery of poses in three-dimensions.

Acknowledgements: We'd like to thank for the support from the following research grants 973-2009CB320904, NSFC-61272027, NSFC-61231010, NSFC-61121002, and ARO Proposal Number 62250-CS.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402. IEEE, 2005.

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI, IEEE Transactions on*, 23(3):257–267, 2001.

[3] I. Bülthoff, H. Bülthoff, P. Sinha, et al. Top-down influences on stereoscopic depth-perception. *Nature neuroscience*, 1(3):254–257, 1998.

[4] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630. IEEE, 1995.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on IST*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM SIGKDD*, pages 43–52. ACM, 1999.

[7] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733. IEEE, 2003.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, pages 1–8. IEEE, 2008.

[9] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. *ECCV*, pages 222–233, 2008.

[10] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. *Human Motion–Understanding, Modeling, Capture and Animation*, pages 271–284, 2007.

[11] Z. Jiang, Z. Lin, and L. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *PAMI*, 34(3):533–547, 2012.

[12] V. Kazemi and J. Sullivan. Using richer models for articulated pose estimation of footballers.

[13] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood feature for human action recognition. In *CVPR*, pages 2046–2053. IEEE, 2010.

[14] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[15] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR workshop*, pages 9–14. IEEE, 2010.

[16] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, pages 3177–3184. IEEE, 2011.

[17] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1 –8, june 2008.

[18] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, pages 1234–1241. IEEE, 2012.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, volume 3, pages 32–36. IEEE, 2004.

[20] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176. IEEE, 2011.

[21] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297. IEEE, 2012.

[22] L. Wang, Y. Wang, T. Jiang, and W. Gao. Instantly telling what happens in a video sequence using simple features. In *CVPR*, pages 3257–3264. IEEE, 2011.

[23] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *ICCV*, pages 1419–1426. IEEE, 2011.

[24] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, pages 489–496. IEEE, 2011.

[25] R. Xu, P. Agarwal, S. Kumar, V. Krovi, and J. Corso. Combining skeletal pose with local motion for human activity recognition. *Articulated Motion and Deformable Objects*, pages 114–123, 2012.

[26] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. In *ICCV*, pages 120–127. IEEE, 1998.

[27] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE, 2011.