Learning Discriminative Activated Simplices for Action Recognition

Chenxu Luo,[1] Chang Ma,[1] Chunyu Wang,[2] Yizhou Wang,[1]
[1]Nat'l Eng. Lab. for Video Technology, Cooperative Medianet Innovation Center
Key Lab. of Machine Perception (MoE), Sch'l of EECS, Peking University, Beijing, 100871, China
[2]Microsoft Research

Abstract

We address the task of action recognition from a sequence of 3D human poses. This is a challenging task firstly because the poses of the same class could have large intra-class variations either caused by inaccurate 3D pose estimation or various performing styles. Also different actions, e.g., walking vs. jogging, may share similar poses which makes the representation not discriminative to differentiate the actions. To solve the problems, we propose a novel representation for 3D poses by a mixture of Discriminative Activated Simplices (DAS). Each DAS consists of a few bases and represent pose data by their convex combinations. The discriminative power of DAS is firstly realized by learning discriminative bases across classes with a block diagonal constraint enforced on the basis coefficient matrix. Secondly, the DAS provides tight characterization of the pose manifolds thus reducing the chance of generating overlapped DAS between similar classes. We justify the power of the model on benchmark datasets and witness consistent performance improvements.

Recent advances in human pose estimation from both color (Toshev and Szegedy 2014; Chen and Yuille 2015; Cherian et al. 2014; Wang et al. 2014) and depth images (Shotton et al. 2013; Girshick et al. 2011) enpower pose-based action recognition closer to practical applications.

However, there are still some challenges requiring further study. The first challenge is the large intra-class variations. This is mostly because: (i) people may perform the same action in very different styles, (ii) and the estimated 3D poses are sometimes inaccurate, e.g., due to occlusions and degraded image qualities. The second critical challenge is the small inter-class distance, *i.e.*, some actions share very similar poses which make them very hard to differentiate.

There have been many works aiming to improve the representation's discriminative power. Some works, e.g., (Yao et al. 2011; Wang et al. 2012b) attempt to solve the problem by using more effective pose features to represent actions. Another line of research, instead, focuses on developing or using powerful classifiers, e.g., SVMs, multiple kernel learning and data mining techniques, to map pose features to a new space where classification is easier. However, in spite of the promising results on some datasets, there is risk of over-fitting to noisy/small training data when either features or classifiers are too complex.

There is a recent work (Wang et al. 2016) which proposes a novel generative representation called *activated simplices* (AS). The AS model represents poses by a mixture of activated simplices where each activated simplex (structure) consists of a few bases whose convex combinations tightly characterize a portion of the pose manifold. The structures are representative to cover the variations of the poses, and also at the same time, discriminative to separate different classes. The discriminativeness is implicitly realized by requiring the learned AS structure to be tight. In other words, the learned AS structure for a class can and can only represent the poses of this class well. For classification, they learn an AS model for each action class independently, and then classify a new pose sequence by projecting its poses to the learned models of all classes respectively and choosing the class with the smallest projection error. This simple classifier outperforms the existing state-of-the-arts with a large margin on three popular benchmark datasets.

However, one limitation of the method (Wang et al. 2016) is that the AS models are independently learned for each action class without considering the data of other classes. So if two action classes share similar poses, then the learned models will probably have large overlap and its discriminative power can be weakened.

Our work improves over the AS model by fortifying its discriminative power. Firstly, instead of learning the basis dictionary for each class independently, we propose to jointly learn a large structured basis dictionary for all classes together. By enforcing a block diagonal constraint on the basis coefficient matrix, the poses from a specific class are encouraged to activate a specific subset of the bases (the bases of its own class) in the dictionary and are suppressed to activate others. This results in a distinctive representation in between different classes. Secondly, considering that some classes may share poses (e.g. neutral poses), a small shared basis dictionary is simultaneously constructed which improves the compactness of the representation. With the learned discriminative activated simplices (DAS), we perform classification/recognition by a simple nearest neighbor classifier which will be disucssed in detail later.

We validate the discriminative power of the proposed DAS model by performing action recognition on three popular benchmarks and a larger dataset composed ourselves. The results show that the DAS model outperforms all the state-of-the-arts on all the datasets even using a simple nearest neighbor based classifier.

Related Work

We review the related work on 3D action recognition and the general discriminative dictionary learning strategies which are not necessarily applied for the task of action recognition.

3D Action Recognition

We classify the existing works on 3D action recognition into three categories based on the adopted representation (features). The first class of works, e.g., (Oreifej and Liu 2013; Wang et al. 2012a; Li, Zhang, and Liu 2010), takes raw depth maps as inputs. Since depth maps are high dimensional, it is critical to sample or extract compact features from them to prevent from over-fitting. For example, Omar et al. (Oreifej and Liu 2013) consider a depth sequence as a surface in the 4D space and compute a histogram of surface normal orientations as a global descriptor. Wang et al. (Wang et al. 2012a) propose to uniformly sample subvolumes from the 4D space and compute "random occupancy patterns" which mainly capture the shape cues of human body. These features are usually fast to compute and some of them have achieved promising performance on existing datasets.

Another line of works represents actions by a set of 3D joint locations, e.g.(Xia, Chen, and Aggarwal 2012; Wang et al. 2012b; Luo, Wang, and Qi 2014). Xia et al. (Xia, Chen, and Aggarwal 2012) partition the 3D space into several bins and quantize each joint into several nearest bins with weighted votes. Then they collect the votes in each bin, form a histogram descriptor and extract discriminative features of the histogram by linear discriminant analysis. Wang et al. (Wang et al. 2012b) compute pairwise joint position features and use data mining techniques to select the joints that are most relevant to class labels.

The third line of works groups body joints into a set of rigid body parts, e.g.(Wang, Wang, and Yuille 2013; Vemulapalli, Arrate, and Chellappa 2014). Wang et al. (Wang, Wang, and Yuille 2013) propose a spatial-temporal-part model to capture the differentiable configurations of body parts for classification. By breaking holistic human poses into parts, the representation is robust to situations where some body joints are inaccurate. A similar idea of learning mid-level action representations was explored in (Wang et al. 2015) for video-based action recognition. Vemulapalli et al. (Vemulapalli, Arrate, and Chellappa 2014) model the geometric relationships between body parts using rotations and translations which is a member of special Euclidean group. Human actions are treated as curves in Lie groups. They also use SVMs as classifiers.

Discriminative Dictionary Learning

We briefly review the most related work in discriminative dictionary learning due to space limit. Jiang et al. (Jiang, Lin, and Davis 2011) associate label information with each basis to enforce discriminability in sparse codes during sparse dictionary learning. Specifically, data from the same class are encouraged to have similar sparse codes. This is successfully applied to the face classification task. Ramirez et al. (Ramirez, Sprechmann, and Sapiro 2010) jointly learn a basis dictionary for each class and encourage that the bases of different classes are as independent as possible. Guo et al. (Guo, Jiang, and Davis 2012) extend (Jiang, Lin, and Davis 2011) to the case where only pairwise "same" or "different" labels are available when learning sparse dictionaries. Specifically, a penalty term is added to the model where data having the same label are encouraged to be similar in terms of sparse codes. Mairal et al. (Mairal et al. 2008) learn a separate sparse dictionary for each class with the constraint that it is "good" at reconstructing this class, and at the same time "bad" for other classes. Our work is most similar to (Jiang, Lin, and Davis 2011) in the sense that we learn the activated simplices by encouraging that the poses of the same class activating similar bases. But it differs from (Jiang, Lin, and Davis 2011) in two aspects: (i) our learned dictionary has two parts— the first part contains the class specific bases and the second part is composed of the shared bases; This can enhance the model's discriminative power as well as its compactness. (ii) We use the label consistent cues in a different way from [19] which directly adds a linear classification error term in the dictionary learning. In contrast, we propose to use block-diagonal constraints on the basis coefficient matrix, which is more consistent with the inherent generative nature of the AS representation, so that the pursued bases better account for the structure of the data distribution.

Activated Simplices

We first briefly summarize the original AS learning method to lay the background. Let $y \in \mathbb{R}^d$ denote a data point and $\{y^{(1)}, y^{(2)}, \cdots, y^{(n)}\}$ denote the training dataset on which we aim to learn a dictionary Z to accurately represent the data. The dictionary Z consists of a set of k bases $Z = \{z_1, \cdots, z_k\}$ and represents any data point y by a linear combination of the bases $y = \sum_{i=1}^{k} z_i \alpha_i$ (or $y = Z\alpha$ in matrix form). The basis coefficient vector $\alpha = (\alpha_1, \cdots, \alpha_k)$ is optimized by minimizing the following reconstruction error: $\alpha^* = \text{argmin}_\alpha \|y - \sum_{i=1}^{k} z_i \alpha_i\|_2$. In order to provide a bounded representation, the authors in (Wang et al. 2016) also enforce convex constraints on the coefficient vector: $\alpha \geq 0$ and $\|\alpha\|_1 = 1$. Putting all these together, learning the basis dictionary Z from the training dataset equals to solving the following optimization problem:

$$\begin{aligned} &\min_{z,a} \sum_{\mu=1}^{n} \|y^\mu - \sum_{i=1}^{k} z_i \alpha_i^\mu\|_2^2 \\ &s.t. \sum_{i=1}^{k} \alpha_i^\mu = 1, \quad \alpha_i^\mu \geq 0, \quad \text{for all } i \text{ and } \mu \\ &\quad \|z_i\|_2 \leq 1, \quad \text{for all } i \end{aligned} \tag{1}$$

The authors theoretically explained that by learning the bases according to equation (1), they finally obtain a mixture of bounded structures (activated simplices) to represent the training dataset. Each simplex is responsible for representing a subset of the training data and each datum is represented by the same simplex.

After learning the bases, they extract a mixture of activated simplices by inspecting the coefficient matrix A of training data. Intuitively, the algorithm groups the bases of each class into several clusters with the requirement that the bases in the same cluster were co-activated by a sufficient datapoints. This grouping strategy gives a much tighter data representation by limiting basis co-activations. We refer the readers to the original paper for more information.

So the final representation for each class is a set of activated simplices $\mathbb{S} = \{\Delta_1, \cdots, \Delta_m\}$. Each simplex consists of several bases $\Delta_i = \{z_1^i, \cdots, z_{m_i}^i\}$ where m_i is the number of bases in this simplex. The model represents a pose y by its closest simplex $\Delta_{a(y)} : \Delta_{a(y)} = \text{argmin}_{\Delta_a \in \mathbb{S}} Dist(\Delta_a, y)$ where $Dist(\Delta_a, y)$ is the representation error of using simplex Δ_a to represent y. The representation error is computed as $Dist(\Delta_a, y) = \min_a \|y - \sum_{i=1}^{m_a} \alpha_i z_i^a\|$ with the non-negative and sum to one constraints on α.

For the action recognition task, the authors in (Wang et al. 2016) first learn independent dictionary for each of the C action classes. Then they project a testing pose sequence onto each basis dictionary (simplex dictionary more precisely; we will come to this later) respectively and obtain the class with the smallest reconstruction error.

Discriminative Dictionary Learning

The above model (Wang et al. 2016) achieves promising results on three popular datasets. However, we argue (and validate by experiments) that this independent learning strategy weakens the model's discriminative power especially for actions having similar poses. We solve the problem by proposing a discriminative dictionary strategy.

We denote $\mathbf{Y} = \{y^{(1)}, y^{(2)}, \cdots, y^{(n)}\}$ as the training data and $\mathbf{Z}^{(i)}$ the dictionary of class i. Suppose $Z^{(i)}$ has k_i bases: $Z^{(i)} = \{z_1^{(i)}, \cdots, z_{(k_i)}^{(i)}\}$. The dictionary of all classes forms the whole dictionary set $\mathbf{Z} = [Z^{(1)}, \cdots, Z^{(C)}]$. Instead of learning the dictionary of each class individually, we learn the dictionary for all the classes jointly. In order to attain a discriminative dictionary, we encourage each pose to be represented by the bases of its own class and discourage it from activating bases of other classes. We achieve it by adding a structured penalty term $\mathbf{Q}$. The cost function is:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{A}} \|\mathbf{Y} - \mathbf{Z}\mathbf{A}\|_F^2 + \|\mathbf{Q} \circ \mathbf{A}\|_F^2 \\ &s.t. \quad \|\mathbf{a}_i\|_1 = 1, \quad \mathbf{a}_i \succeq 0, \quad \forall i \\ &\qquad\; \|\mathbf{z}_i\|_2 \leq 1, \quad \forall i, \end{aligned} \tag{2}$$

where $\mathbf{z}_i$ is the i-th column vector of $\mathbf{Z}$, which represents a base vector, and $\mathbf{a}_i$ is the i-th column vector of $\mathbf{A}$, which correspond to the coefficient to represent a certain data in the training set. $\mathbf{A}$ is denoted as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \cdots & \mathbf{A}_{1,C} \\ \vdots & \vdots & \vdots \\ \mathbf{A}_{C,1} & \cdots & \mathbf{A}_{C,C}. \end{bmatrix}$$

C is the number of class and $A_{i,j}$ is coefficient matrix of bases from class i when projecting data from class j onto the whole dictionary. "$\circ$" is the Hadamard (elementwise) product and $\|\cdot\|_F$ is the Frobenius norm of the matrix.

$\mathbf{Q}$ is a structured penalty matrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{1,1} & \cdots & \mathbf{Q}_{1,C} \\ \vdots & \vdots & \vdots \\ \mathbf{Q}_{C,1} & \cdots & \mathbf{Q}_{C,C}. \end{bmatrix}$$

where $\mathbf{Q}_{i,j}$ is the structured penalty term for $\mathbf{A}_{i,j}$. Specifically, $\mathbf{Q}_{i,j}$ is a zeros matrix when i equals j. Otherwise, it is a matrix with all its elements equal to λ_1. Intuitively, a data point can activate the bases with the same class label without any penalty. But it needs to pay a constant penalty λ_1 if it activates bases from other classes. This regularization explicitly enforced the coefficient matrix A to be block-diagonal and implicitly enforced the bases of different classes to be independent and far apart.

The Common Dictionary: Shared Bases

Considering that different classes may share some common patterns, for example, actions of drawing tick and drawing circle are almost the same except the motion of hand. It would be impossible to prevent such a pose from activating bases from other classes. So it is reasonable to add a common dictionary part whose bases are shared by all classes. Ideally, it will only represent the commonality between different classes which is useless for telling the poses apart. Introducing the common dictionary set can enhance the discriminability of each class-specific dictionary and make the representation more compact. The new cost function differs only slightly from Eq(2). Specifically, $\mathbf{Z} = [\mathbf{Z}^{(0)}, \mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(C)}]$, where $\mathbf{Z}^{(0)}$ denotes the common dictionary. $\mathbf{A}$ and $\mathbf{Q}$ can be updated accordingly.

Noting that the common dictionary should not dominate the whole dictionary, which means we do not want data to be represented mainly by the common bases (in extreme cases, all data points only activate the shared dictionary). So the penalty coefficients of the shared dictionary ($\mathbf{Q}_{0,j}$) are set to be λ_2, which should be less than λ_1. In practice,λ_1 is set to be around 0.1, and λ_2 is set to be around 0.01 .

Optimization

Although the above optimization problem (2) is non-convex, it is convex with respect to each of the variables ($\mathbf{Z}$ and $\mathbf{A}$) when the other one is fixed. In this paper, we alternatively optimize for the two variables in an online way. Specifically, we first fix the dictionary $\mathbf{Z}$ and update the coefficient matrix $\mathbf{A}$. The optimization can be reformulated as

$$\begin{aligned} &\min_{\mathbf{A}} \left\| \begin{bmatrix} \mathbf{Y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\mathbf{A} \\ \mathbf{Q} \circ \mathbf{A} \end{bmatrix} \right\|_F^2 \\ &\text{s.t.} \|\mathbf{a}_i\|_1 = 1, \mathbf{a}_i \succeq 0, \forall i \\ &\quad \|\mathbf{z}_i\|_2 \leq 1, \forall i. \end{aligned} \tag{3}$$

Thus we can use the active-set based algorithm proposed in (Chen, Mairal, and Harchaoui 2014) to solve for the coefficient matrix $\mathbf{A}$.

Then fixing the coefficient matrix $\mathbf{A}$, we obtain the following problem:

$$\min_{\mathbf{Z}} \|\mathbf{Y} - \mathbf{Z}\mathbf{A}\|_F^2 \\ \|\mathbf{z}_i\|_2 \leq 1, \quad \forall i, \tag{4}$$

which turns out to be the classical dictionary learning problem and can be solved efficiently using online dictionary learning algorithm(Mairal et al. 2009).

Extracting Discriminative Activated Simplices

As in (Wang et al. 2016), after learning the dictionary for all the classes, we first represent each datum using the dictionary of its class as well as the common dictionary. They we group the coefficient vectors of each class into several clusters and obtain discriminative activated simplices using the same way as (Wang et al. 2016). So our final representation for each class is a set of discriminative activated simplice $\mathbb{S} = \{\Delta_1, \cdots, \Delta_m\}$. The model represents each pose y by its closest simplex $\Delta_{a(y)}$: $\Delta_{a(y)} = \arg\min_{\Delta_a \in \mathbb{S}_\partial} Dist(\Delta_a, y)$.

Action Recognition Experiments

In this section, we evaluate the action recognition performance on three popular benchmark datasets (Li, Zhang, and Liu 2010) (Seidenari et al. 2013) (Xia, Chen, and Aggarwal 2012) and a self-composed large dataset.

Since the DAS are purposely (discriminatively) learned to be far apart from those of other classes, we take advantage of this property and propose a simple nearest-neighbor based action classifier. Specifically, we first learn a mixture of DAS for each action class by the proposed method. Then given a pose sequence, we compute the distance between its poses and the DAS of each class, the class that achieves the smallest distance is the predicted class.

Data Preprocessing

An action instance is a 3D pose sequence $(y_1, \cdots, y_N)$ where a pose $y_\mu \in \mathbb{R}^{3k}$ is a high dimensional vector of the 3D coordinates of k body joints. The number of joints is about 15 although it may slightly vary for different datasets. Considering that some actions, e.g., standing up and sitting down, are difficult to separate by only static poses (Wang, Wang, and Yuille 2013), following (Wang et al. 2016), we adopt the action-snippet representation which encodes the temporal order of poses by combining ten consecutive poses together as an element: $\mathbf{y}_\mu = [y_\mu \cdots y_{\mu+9}]$ where y_j is a 3D pose at time j. The neighboring action-snippets overlap so the first snippet starts at pose one, the second at pose two, and so on. Then we represent an action by a set of action-snippets: $A = \{\mathbf{y}_1, \cdots, \mathbf{y}_{N-9}\}$.

We learn DAS models for action-snippets rather than static poses. Similarly when classifying a sequence we project its action-snippets onto the DAS of each action class and output the class with smallest projection error.

Table 1: Average action recognition accuracy of all 252 5-5 splits on MSR-Action3D.

Methods	Accuracy (%)	Year
(Oreifej and Liu 2013)	82.15	2013
(Rahmani et al. 2014)	82.70	2014
(Tran and Ly 2013)	84.54	2013
(Du, Wang, and Wang 2015)	89.00	2015
(Wang et al. 2016)	91.40	2016
Our method	**95.62**	

Results on The MSR-Action3D Dataset

The MSR-Action3D dataset (Li, Zhang, and Liu 2010) provides 557 3D human pose sequences of ten subjects performing 20 actions. Each sequence has about 50 frames. This is a challenging dataset first because some actions in the dataset are similar (e.g. the drawing tick and drawing circle actions) which makes them difficult to distinguish. Second, the estimated 3D poses are sometimes inaccurate especially when occlusion happens which increases the intra-class variations.

Following (Wang et al. 2016), we choose the cross-subject evaluation criterion. Most existing works choose five subjects for training and the remaining five subjects for testing, e.g. in (Li, Zhang, and Liu 2010), and report the result based on a single split. However, it is shown in (Padilla-López, Chaaraoui, and Flórez-Revuelta 2014) that choosing which five subjects for training has large influence on the results. To make the results more comparable, in this work, we experiment with all 252 possible subject splits and report the average accuracy.

Table 1 compares our method with the state-of-the-art methods using the protocol of "average over all splits". We learned 20 bases for each class and 5 shared bases.Our method outperforms all of the four state-of-the-art methods (Oreifej and Liu 2013; Rahmani et al. 2014; Tran and Ly 2013; Wang et al. 2016) under the same protocol. In particular, it outperforms (Wang et al. 2016) by about 4.3% which shows the enhanced discriminative power gained by joint learning. It is worth noting that our method and (Wang et al. 2016) are the simplest in terms of both features and classifiers which makes them less prone to over-fitting and more extendable to address other related tasks.

Diagnostic Analysis To understand the reasons behind the recognition performance, we propose to inspect the models from various aspects. Both the AS and DAS models learn an individual basis dictionary for each class. In the current nearest neighborhood classification framework, misclassifying a pose sequence means that using the bases from other classes gives a smaller reconstruction error. To put it in another way, if we combine the dictionaries of all classes together, then the optimal set of activated bases will include those from other classes. In this case, the coefficient matrix maynotbeblock-diagonal.So we plot the coefficient matrices obtained by the AS model and our DAS model respectively. Figure 1 shows the results. We can see from the left penal that for the AS model,data from a certain class tend to ac-

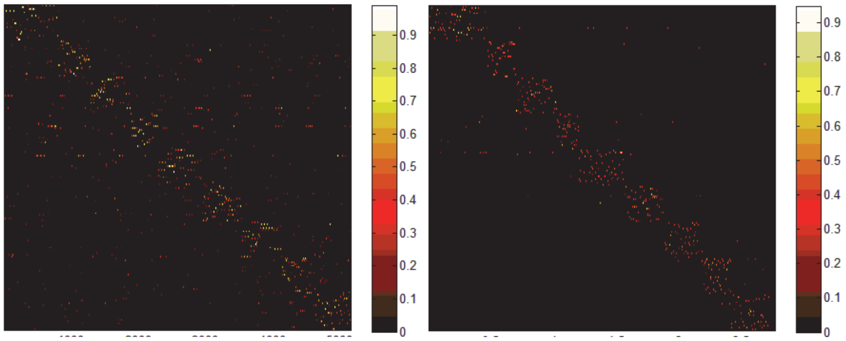

Figure 1: This figure shows the coefficient matrix when encoding the first 9 classes of poses by the dictionaries of (Wang et al. 2016) (left) and our method (right), respectively. We do not enforce the penalty term after learning the dictionary. We can see that the coefficient matrix of our method is more block-diagonal which shows that the poses from a particular class will activate bases learned for that class. In other words, the bases of other classes are less likely to be activated. Best see by zooming in.

tivate bases from other classes. On the contrary, our DAS model activates only bases from the same class. This shows that the bases learned by DAS are far away from data of other classes. Hence the DAS model is more discriminative.

We also explore the model in terms of projection errors directly. When the pose sequences of classes A and B are similar (e.g., having similar poses), then it is highly probable that the AS models of A and B will obtain very similar projection errors on the pose sequences of the two classes. The problem makes classification trivial and vulnerable to noisy or inaccurate poses. This is shown in the top sub-figures in Figure 2. We can see that, for some classes, the reconstruction error difference between the "correct" model and other models is small. That is the situation where the nearest neighbor based classifier fails. However, using our discriminatively learned activated simplices can increase that difference and thus improve the final classification accuracy. Please compare the top and the bottom panels in Figure 2 to tell the difference.

We now evaluate the influence of the shared bases. In particular, we report the results for the common basis dictionary size of $0, 5, 10, 15$ and 20, respectively. See Figure 3. We can see that the method without the shared dictionary (size=0) achieves the lowest accuracy of 95.00%. In contrast, using shared bases can improve the performance by about 1%. Besides, the performance reaches maximum when the shared dictionary size is 10 and begins to decrease when the dictionary size continues to grow. This is reasonable because increasing shared dictionary size reduces the model complexity, hence may decrease overfitting. But when the number of shared bases keeps enlarging, it can hurt the model's discriminative power.

Results on The Florence Dataset

The Florence dataset (Seidenari et al. 2013) was captured using a Kinect camera at the University of Florence. It includes nine activities: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. Ten subjects perform the each of the nine actions for two or three times. In total, there are 215 activity samples. Following the data suggestion, we adopt a leave-one-actor-out protocol: we train the classifier using all the sequences from nine out of ten actors and test on the remaining one. We repeat this procedure for all actors and compute the average classification accuracy values of the ten actors.

Table 2: Action recognition accuracy on the Florence dataset using leave-one-actor-out setting.

Methods	Accuracy (%)
(Seidenari et al. 2013)	82.15
(Vemulapalli, Arrate, and Chellappa 2014)	90.88
(Devanne et al. 2014)	87.04
(Wang et al. 2016)	94.25
Our method	**95.30**

Table 3: Action recognition accuracy using leave-one-sequence-out setting on the UTKinect dataset.

Methods	Accuracy (%)	Year
(Devanne et al. 2014)	91.50	2014
(Xia, Chen, and Aggarwal 2012)	90.92	2014
(Wang et al. 2016)	96.48	2016
Our method	**97.99**	

We set the number of bases for each class to be 30 and the number of common bases to be $5(275$ in total) by cross-validation. Table 2 compares our method with the state-of-art methods on this dataset. Our approach achieves the highest recognition accuracy, outperforms (Devanne et al. 2014) (Xia, Chen, and Aggarwal 2012) by about 6% and (Wang et al. 2016) by about 1%. The results justify the usefulness of the discriminative learning.

Results on The UTKinect Dataset

The UTKinect dataset (Xia, Chen, and Aggarwal 2012) provides 3D human pose sequences obtained from Kinect. The dataset contains ten daily activities including walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands and clap hands. Ten subjects are instructed to repeat each action two times. So the total number of pose sequences is 200.

We use the common "leave-one-sequence-out" evaluation

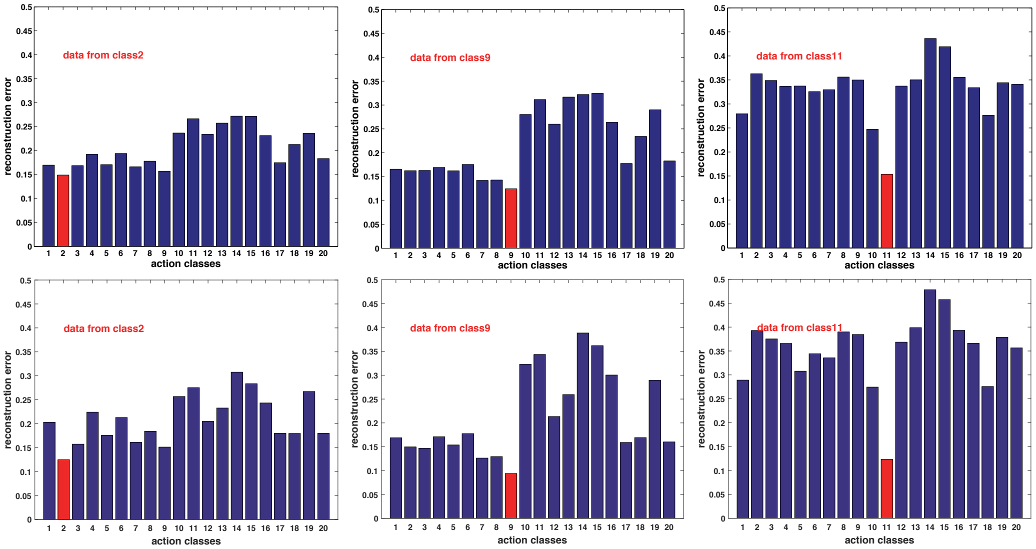

Figure 2: Each figure in the top row shows the reconstruction error when projecting poses of a particular class (e.g. 1,2,9 and 11) to the activated simplices (Wang et al. 2016) of the 20 classes on the MSR-Action3D dataset. The figures in the bottom row show the results of our model. The error is the minimum if the sequences are reconstructed by the DAS model of the right class. More importantly, we can see that the error difference among all classes is increased by using discriminative activated simplices than (Wang et al. 2016) (top figures).

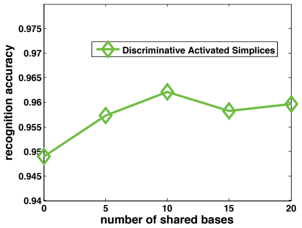

Figure 3: Action recognition accuracy on the MSR-Action3D dataset when the number of shared bases varies.

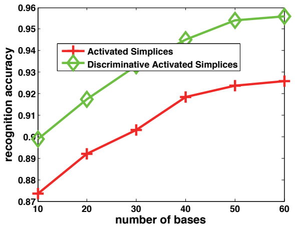

Figure 4: The figure shows the action recognition accuracy of the two methods (i.e. the AS, DAS) as the number of the bases vary on the complied dataset.

criterion to report the performance. More specifically, one sequence is used for testing and the rest of the sequences are used for training. We repeat the process for all sequences and report the average accuracy. We learn 20 bases for each class and 5 bases for the common part, which ends up with about 15 discriminative activated simplices for each action class. The dimension of the simplices is five on average. Table 3 shows the results. We can see that our method outperforms all the state-of-the-arts.

Results on a Composed Dataset

We compose a larger dataset to validate the scalability of the proposed method by combining the above three datasets together. We obtain about one thousand pose sequences (about $40K$ poses) with 39 action classes. We split the dataset into two halves for training and testing respectively, mainly by subjects to help prevent from over-fitting.

We compare out model with the AS model,using the code provided by (Wang et al. 2016). We set the number of the common dictionary to be 5. We keep the number of bases in our method and (Wang et al. 2016) the same for fair comparison. Figure 4 shows the recognition accuracy. We can see that our method consistently outperforms (Wang et al. 2016) when the number of bases varies. Besides, our DAS model can achieve the quite similar recognition accuracy with much less bases needed. This shows the compactness of the DAS model. In addition, the results also show that increasing the size of dataset (both the number of classes and the number of pose sequences) will not harm the performance much, this demonstrate the scalability of our model.

Conclusion

In this paper, we propose a discriminative structure learning method which enhances discriminative capacity. We apply it to the action recognition task, the simple nearest neighbor based classifier outperforms all of the state-of-the-arts on 3 popular datasets. Besides, classifying a typical pose sequence is fast with a careful implementation. It is also worth noting that this is a general discriminative learning method and can be used for other data/tasks as well.

Acknowledgments

This work was supported in part by 973-2015CB351800, NSFC-61231010, 61527804, 61421062 and 61210005.

References

Chen, X., and Yuille, A. L. 2015. Parsing occluded people by flexible compositions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3945–3954.

Chen, Y.; Mairal, J.; and Harchaoui, Z. 2014. Fast and robust archetypal analysis for representation learning. *arXiv preprint arXiv:1405.6472*.

Cherian, A.; Mairal, J.; Alahari, K.; and Schmid, C. 2014. Mixing body-part sequences for human pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; and Del Bimbo, A. 2014. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 1110–1118.

Girshick, R.; Shotton, J.; Kohli, P.; Criminisi, A.; and Fitzgibbon, A. 2011. Efficient regression of general-activity human poses from depth images. In *ICCV, 2011*, 415–422. IEEE.

Guo, H.; Jiang, Z.; and Davis, L. S. 2012. Discriminative dictionary learning with pairwise constraints. In *Computer Vision–ACCV 2012*. Springer. 328–342.

Jiang, Z.; Lin, Z.; and Davis, L. S. 2011. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR, 2011 IEEE Conference on*, 1697–1704. IEEE.

Li, W.; Zhang, Z.; and Liu, Z. 2010. Action recognition based on a bag of 3d points. In *(CVPRW), 2010 IEEE Conference on*, 9–14. IEEE.

Luo, J.; Wang, W.; and Qi, H. 2014. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *CVPR*. IEEE.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In *ICML*, 689–696. ACM.

Oreifej, O., and Liu, Z. 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 716–723. IEEE.

Padilla-López, J. R.; Chaaraoui, A. A.; and Flórez-Revuelta, F. 2014. A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset. *arXiv preprint arXiv:1407.7390*.

Rahmani, H.; Mahmood, A.; Huynh, D. Q.; and Mian, A. 2014. Real time action recognition using histograms of depth gradients and random decision forests. In *WACV*, 626–633. IEEE.

Ramirez, I.; Sprechmann, P.; and Sapiro, G. 2010. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 3501–3508. IEEE.

Seidenari, L.; Varano, V.; Berretti, S.; Del Bimbo, A.; and Pala, P. 2013. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *CVPRW*, 479–485. IEEE.

Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; and Moore, R. 2013. Real-time human pose recognition in parts from single depth images. *Communications of the ACM* 56(1):116–124.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1653–1660.

Tran, Q. D., and Ly, N. Q. 2013. Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences. In *RIVF, 2013 IEEE RIVF International Conference on*, 253–258. IEEE.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 588–595. IEEE.

Wang, J.; Liu, Z.; Chorowski, J.; Chen, Z.; and Wu, Y. 2012a. Robust 3d action recognition with random occupancy patterns. In *ECCV*. Springer. 872–885.

Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012b. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR, IEEE Conference on*, 1290–1297. IEEE.

Wang, C.; Wang, Y.; Lin, Z.; Yuille, A.; and Gao, W. 2014. Robust estimation of 3d human poses from a single image. In *CVPR*, 2361–2368.

Wang, Y.; Wang, B.; Yu, Y.; Dai, Q.; and Tu, Z. 2015. Action-gons: Action recognition with a discriminative dictionary of structured elements with varying granularity. In *ACCV*. Springer. 259–274.

Wang, C.; Flynn, J.; Wang, Y.; and Yuille, A. 2016. Recognizing actions in 3d using action-snippets and activated simplices. In *AAAI*.

Wang, C.; Wang, Y.; and Yuille, A. L. 2013. An Approach to Pose-Based Action Recognition. In *CVPR*, 915–922.

Xia, L.; Chen, C.-C.; and Aggarwal, J. 2012. View invariant human action recognition using histograms of 3d joints. In *CVPRW, IEEE Conference on*, 20–27. IEEE.

Yao, A.; Gall, J.; Fanelli, G.; and Van Gool, L. J. 2011. Does human action recognition benefit from pose estimation?. In *BMVC*, volume 3, 6.