Activity Recognition by Learning Structural and Pairwise Mid-level Features Using Random Forest

Jie Hu
Department of Computer
Science and Engineering
University at Buffalo, SUNY
Buffalo, NY 14260 USA
jhu6@buffalo.edu

Yu Kong
Department of Electrical and
Computer Engineering
Northeastern University
Boston, MA 02115 USA
yu.kong@neu.edu

Yun Fu
Department of Electrical and
Computer Engineering
Northeastern University
Boston, MA 02115 USA
yunfu@ece.neu.edu

Abstract— **This paper presents a novel random forest based method to build mid-level features describing spatial and temporal structure information for activity recognition. Our model consists of two separate parts, spatial part and temporal part, which are employed to capture the distinctive characteristics in spatial and temporal domains of activity analysis. In the spatial part, densely sampled low level features are passed through the first level random forest and concatenated structurally to form spatial mid-level features. In the temporal part, we use results from the first level random forest on sparsely sampled interest points to build pairwise mid-level features. The second level random forests operate on all the mid-level features and compute scores for these two parts. Then final recognition is based on the weighted sum of these two parts. Our method smoothly fuses both spatial and temporal information and builds more descriptive models, which can better represent human activities in large variations. Experimental results show that our method achieves promising performance on three available action and facial expression datasets.**

I. INTRODUCTION

Vision based activity recognition has been widely applied in human-computer interaction, visual surveillance, digital entertainment and some other fields. A reliable recognition of activities mainly depends on features extracted from image sequences, as well as the statistical model applied.

As we can see, activities can be regarded as a spatial-temporal combination of motions presented by different parts of the subject. For example, we can find wide open eyes and mouth in a "fear" face; People move their legs and arms alternatively when walking. Previous methods mainly consider local features or use bag-of-words approaches to cluster local features into a set of words [17][4][15]. Although promising results have been achieved, these methods do not consider spatial and temporal structural information. Recent work [12][18][5] improves recognition accuracy using mid-level features which model the spatial and temporal relationships of local features. However, modeling all kinds of relationships in a homogenous way may reduce performance, since it loses the inhomogeneity of different relationships and is sensitive to types and positions of local features.

To address the aforementioned problems, we propose to describe spatial and temporal relationships in different manners. Local features in spatial neighborhood make up the specific appearance of the action for each temporal interval. We construct spatial mid-level features based on densely sampled features, which give effective description of action and can be easily combined according to their relative positions. From the temporal perspective, for each spatial position, though local features of the same category may vary a lot depending on the subject and speed along the time axis, their pairwise correspondence may comply with some regularity. We construct temporal mid-level features to encode pairwise relationships based on sparsely sampled features, which are informative and robust to noise.

In this paper, we propose a discriminative action recognition model that separately constructs spatial and temporal structures, and combine results from these two parts to make the final recognition. The overview of our method is shown as Fig. 1. Our method consists of two parts, the spatial part (the upper part in Fig. 1) and the temporal part (the bottom part in Fig. 1). We first use the first-level random forest to obtain the histogram features of densely sampled cuboids and space-time interest points. Structural mid-level features for temporal intervals are based on the histogram features of densely sampled cuboids. All intervals obtain their scores using second level random forests and combine them to obtain the spatial part's score. Histogram features of two space-time interest points with specific temporal relationships are multiplied to form the temporal pairwise mid-level features. Pairwise relations contribute to the scores for spatial sub-volumes containing them by passing their mid-level features through the second level random forest, furthermore contributing to the temporal part's score. Then scores of both spatial and temporal parts are combined to obtain the final score to make recognition.

One challenge in the description of structural information is how to choose the forms of features. Instead of using combination of low-level features directly, which are of high dimensionality and diversity, our two-level random forest framework makes full use of the characteristic of random forest that can provide an effective measure for local feature related to category information. When fused structurally or pair-wisely to compose mid-level features, these features give a concise but precise depiction of information for some spatial or temporal relationships.

II. RELATED WORK

Recent work shows excellent performance when fusing local features in multiple ways for activity recognition. Niebles et al. [12] present a constellation of bags-of-features to hierarchically combine spatial-temporal features. Wang et al. [18] regard actions as a set of hidden parts, building their

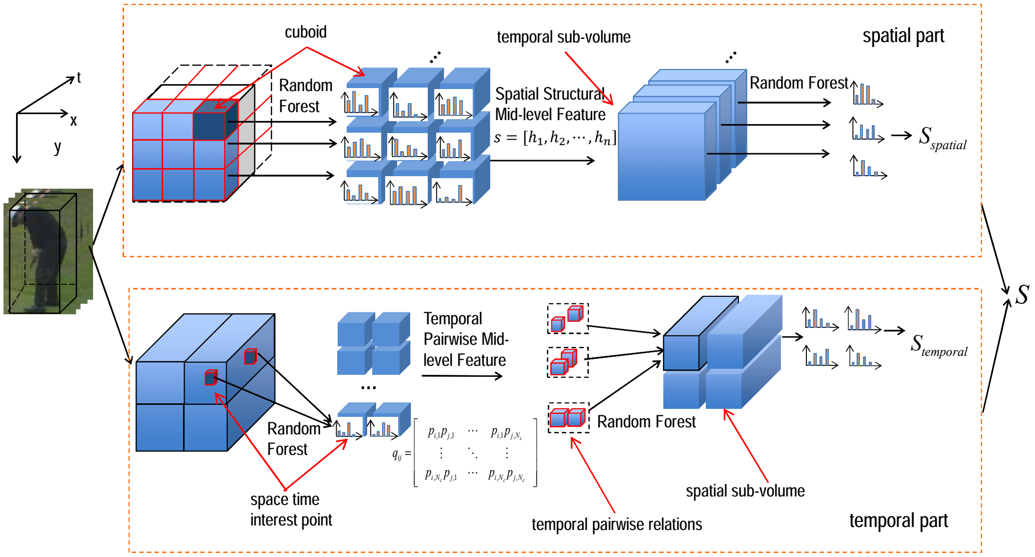

cuboid
temporal sub-volume
spatial part
t
x
y
Random Forest
Spatial Structural Mid-level Feature
$s = [h_1, h_2, \cdots, h_n]$
Random Forest
$S_{spatial}$
S
Temporal Pairwise Mid-level Feature
$S_{temporal}$
Random Forest
$q_{ij} = \begin{bmatrix} p_{i,1}p_{j,1} & \cdots & p_{i,1}p_{j,N_c} \\ \vdots & \ddots & \vdots \\ p_{i,N_c}p_{j,1} & \cdots & p_{i,N_c}p_{j,N_c} \end{bmatrix}$
Random Forest
space time interest point
spatial sub-volume
temporal pairwise relations
temporal part

Fig. 1. Overview of our method. Spatial-temporal volumes cropped from videos go through two parallel parts to get the category label. **Top:** Spatial part. Low level features for densely sampled cuboids go through the first level random forests for histogram representation h and assemble to form the spatial structural mid-level feature s. Spatial part score $S_{spatial}$ is got from the combination results of all temporal sub-volume passing through the second level random forest. **Bottom:** Temporal part. Histogram representations p for space-time interest points are multiplied mutually to make up mid-level feature for temporal pairwise relations. Scores for spatial sub-volumes are got from results that temporal mid-level features go through the second level random forest and further more make up the temporal part score $S_{temporal}$. $S_{spatial}$ and $S_{temporal}$ are combined to get the final score S to make the final recognition.

inter-relationships using hidden conditional random field. There are also some models presented to describe structure information in pairwise ways. Ryoo et al. [14] introduces a set of spatial and temporal pairwise relationships, represented as high dimensional histograms. Kovashka et al. [7] presents a method that constructs a hierarchy of vocabularies using pairwise neighbor information. In our model, densely sampled cuboids are structured spatially and sparse space time interest points are fused in a pairwise way for recognition.

Random forest, first proposed in classification problems [3], has explored its use in computer vision. With the property that all leaf nodes give the distribution of sample points falling some specific interval in feature space, random forest behaves as a mapping function in some applications. Yao et al. [19][20] use the posterior probability that each sample patch of object belongs to different categories as part of the transformation from feature space to Hough space to obtain the vote for label. In the work of [22][21], posterior probability for each space time interest point obtained from random forest are used to calculate the mutual information between the query video and sub-volume in a dataset.

Two most related work to this paper are [5] and [11]. In [5], an AdaBoost based framework is used to compose mid-level features for action patch and a second AdaBoost is used to evaluate all mid-level features to give the final action label. Only two-class recognition can be done under this framework. Similar to the work [11], distributions of samples obtained from leafs of the first level random forest are regarded as vectors to compose the structural or pairwise mid-level features. However, the work in [11] only computes mid-level features for temporally sub-volumes while we separately learns mid-level features for both spatial and temporal structures. We use random forest make an efficient integration of both spatial and temporal, structural and pairwise information for multi-class recognition.

III. Method

Our goal is to recognize human activities, such as actions or facial expressions from videos. The input of our model is the content of human or face centric cropped bounding boxes detected from image sequences. We regard these bounding boxes as 3D spatial-temporal volumes in the whole process. The model we use consists of two parts, each of which considers spatial structural (section III. C) and temporal pairwise relations (section III. D), respectively. Each of them can be represented in 3 layers: 1) Low-level representation (section III. A). 2) Mid-level representation (section III. C and section III. D). 3)Final classifier (section III. E). Two levels of random forests are used to make the connection between layers. The construction of random forest to transfer low-level representation into mid-level features is introduced in section III. B. The second level random forest that works on mid-level features is quite similar with the first level one.

A. Low-level Feature

We extract HOG/HOF [9] feature from videos. The combination of HOG (histograms of oriented gradients) and HOF (histograms of optical flow) integrates both static appearance information and local motion information, and shows promising performance in action recognition [17].

Both dense and sparse sampling are used in this work. For dense sampling, the whole 3D spatial-temporal volume is cut into a set of cuboids with size of $w \times w \times l$. Each cuboid is composed of $n_x \times n_y \times n_t$ cells HOG/HOF. To extract optical flows, we use the algorithm in [16]. Orientations are quantized into 9 bins for both HOG and HOF. For sparse sampling, space-time interest point [8] detector is applied. For each interest point p, similar as dense sampled features, HOG and HOF are calculated in its $n_x \times n_y \times n_t$ cell space-time neighborhood. Histograms of 4-bin for HOG and 5-bin for HOF are computed for each cell.

B. Random forest construction

In both spatial and temporal parts, a 3D space-time volume V is represented by a set of local features D, denoted as $V = \{D_i\}$. We assume that each local feature in a video shares the same category label, $c \in \mathscr{C}$, with the 3D space-time volume containing it. Each local feature D is represented as $D = \{f, c, d\}$, where $f = \{f^{HOG}, f^{HOF}\}$ is the extracted HOG/HOF feature, c is its label and $d = \{x, y, t\}$ is its relative displacement from the 3D space-time volume center.

In the first-level random forest, training samples are densely sampled cuboids (spatial part) or space-time interest points (temporal part). HOG, HOF or displacement is chosen in test functions with probability $\delta_{HOG}, \delta_{HOF}, \delta_d$, s.t. $\delta_{HOG} + \delta_{HOF} + \delta_d = 1$, which decide their weights in the recognition. Assume f is the chosen feature, f^τ is the value of feature f's τ-th dimension and $\theta \in \mathbb{R}$ is a threshold. Suppose C is a training sample, we randomly pick up one of the following four types of test functions investigated in [6]:

$$\begin{aligned} t^{(1)}(C, f, \tau, \theta) &= [f^\tau > \theta] \\ t^{(2)}(C, f, \tau_1, \tau_2, \theta) &= [f^{\tau_1} - f^{\tau_2} > \theta] \\ t^{(3)}(C, f, \tau_1, \tau_2, \theta) &= [f^{\tau_1} + f^{\tau_2} > \theta] \\ t^{(4)}(C, f, \tau_1, \tau_2, \theta) &= [|f^{\tau_1} - f^{\tau_2}| > \theta] \end{aligned} \tag{1}$$

A sample goes to the left branch when the test function equals 1; otherwise, it goes to the right branch.

We create a pool of test functions $\{t^k\}$ for each non-leaf node, where function type, dimension τ_1 and τ_2 (or τ for test function of type 1), and threshold θ are randomly generated. The test function t from the pool that maximizes the information gain is assigned as the test function for this node. Let S be the original training sample set at non-leaf node pa, and $S_l^{t'}, S_r^{t'}$ be two sub-sets of S by splitting method t'. Then the splitting function t should satisfy [10]:

$$\begin{aligned} t &= \arg\max_{t' \in t^k} \{E(S) - E(S; t')\} = \arg\min_{t' \in t^k} E(S; t') \\ &= \arg\min_{t' \in t^k} \{ \frac{|S_l^{t'}|}{|S|} \sum_{cl} p_{c_l} \ln(p_{c_l}) + \frac{|S_r^{t'}|}{|S|} \sum_{cr} p_{c_r} \ln(p_{c_r}) \} \end{aligned} \tag{2}$$

where $E(.)$ is a function for computing entropy. p_{cl}, p_{cr} are the proportions of cuboids for different categories in the left and right child node.

All training samples start from the root, split recursively to construct trees until one of the following three conditions occurs: 1) All samples in the node are with the same label. 2) The number of samples are small. 3) The depth of the node reaches the predefined maximal value. Each leaf node stores the proportions of samples belonging to different categories, which can be written as a vector $[p_1, p_2, ..., p_{N_c}]$, assuming there are N_c categories in total. During testing, we pass a test sample through the forest and average the probability vectors it receives from all trees and obtain a normalized histograms $h = [\bar{p}_1, \bar{p}_2, ..., \bar{p}_{N_c}]$, which have less dimensions than low level features but well describe the confidence of each sample belonging to different categories.

To realize randomization, we train each tree on a random subset of the training data using bagging method. Suppose the whole training set is T with size s and the training set for the ith tree is T_i with size s_i. If s is large enough and $s = s_i$, by sampling examples uniformly with replacement, about $2/3$ training samples for tree i are expected unique [2]. We use out-of-bag estimation to obtain the normalized histograms for each training sample as [11].

Intuitively, features vary a lot for different spatial locations. To make the recognition more accurate, we train separate first level random forests for samples in different spatial locations. Each 3D space-time volume is cut into $N_h \times N_w$ sub-volumes spatially and local feature samples falling in the same spatial sub-volume are trained and tested using the same random forest. We use the center of the sample to decide whether one cuboid is inside or outside of a spatial sub-volume.

C. Spatial Structural Mid-Level Feature

For the spatial part, densely sampled features at certain temporal position are combined structurally to construct mid-level features. After the first level random forest, we get a set of normalized histograms $\{h_i\}$ for densely sampled cuboids which only model the local information, where i is the index for cuboids. To describe the spatial relationship, we need to consider structural information as well. Since densely sampled cuboids are at regular positions and scales in space and time, we simply concatenate histograms for all the cuboids with the same temporal coordinate in a video as one long vector $s = [h_1, h_2, \cdots, h_n]$, regarded as spatial structural a mid-level feature. The dimensions that h_i anchored in the vector s describe the spatial location for cuboid i. Then we can obtain a set of such structural mid-level features which depict overall spatial information along the time axis. Structural mid-level features are the input to the second level random forest in spatial part to make a further recognition.

D. Temporal Pairwise Mid-Level Feature

For temporal part, since activities may have large temporal variation for different subjects and speed, simply concatenation of sample points' information along time axis cannot give a proper description of the temporal relationship. Instead of modeling all sample points in one time sequence, we consider two sample points with certain temporal relationship to construct a temporal pairwise mid-level feature.

We use sparse space-time interest points to build temporal pairwise relationships, which are invariant to the scale change in both spatial and temporal domains. All space-time interest points get their normalized histograms from the spatial location depended first level random forests. We define three types of temporal pairwise relationships in the following for any two space-time interest points i and j located in the same spatial sub-volume. The spatial sub-volumes are the same as those obtained in section III. C.

$$\begin{array}{ll} R1 & equal : |t_i - t_j| < \gamma_0 \\ R2 & near : \gamma_0 < |t_i - t_j| < \gamma_1 \\ R3 & far : |t_i - t_j| > \gamma_1 \end{array} \tag{3}$$

where t_i and t_j are the temporal coordinate of i and j, γ_0 and γ_1 are two thresholds.

Since h_i and h_j give the distributions of space-time interest points i and j, assuming i and j are independent given their low-level features and locations, $h_i^T \cdot h_j$ represents the joint distribution of interest points i and j.

$$q_{ij} = h_i^T \cdot h_j = \begin{bmatrix} p_{i,1}p_{j,1} & \cdots & p_{i,1}p_{j,N_c} \\ \vdots & \ddots & \vdots \\ p_{i,N_c}p_{j,1} & \cdots & p_{i,N_c}p_{j,N_c} \end{bmatrix} \tag{4}$$

$p_{i,c_i}p_{j,c_j}$ gives the confidence that i belongs to c_i and j belongs to c_j simultaneously. q_{ij}, which is a temporal pairwise pairwise mid-level feature, is the input of the second level random forest for the temporal part.

E. Final Classification

For final classification, all structural mid-level features (obtained from section III. C) and pairwise mid-level features (obtained from section III. D) are passed through the second level random forests. Final recognition is made based on the weighted sum of scores from both spatial and temporal parts.

For the spatial part, the second level random forest is built on all the training spatial structural mid-level features. The construction of the random forest is similar to section III. B except that there is no feature selection step.

When test, each temporal sub-volume obtains a vector containing the distribution belonging to different categories from the second level random forest. Suppose there are M sub-volumes in the test video and $P_k = [P_{k,1}, \cdots, P_{k,N_c}]$ is the distribution for sub-volume k. We average all M distributions to give the spatial part scores for different categories.

$$S_1 = \frac{1}{M}\sum_k P_k = \frac{1}{M}[\sum_k P_{k,1}, \cdots, \sum_k P_{k,N_c}] = [S_{1,1}, \cdots, S_{1,N_c}] \tag{5}$$

For the temporal part, we build the second level random forest for each type of pairwise relationship in spatial sub-volumes on pairwise mid-level features q from section III. D. During test, suppose there are N_h temporal pairwise relationships falling in spatial sub-volume h, with the ith relationship having $P_i^h = [P_{i,1}^h, \cdots, P_{i,N_c}^h]$ as its distribution from the corresponding random forest. The hth sub-volume's scores for different categories are the average of all the P_i^h:

$$S_2^h = \frac{1}{N_h}\sum_i P_i^h = \frac{1}{N_h}[\sum_k P_{i,1}^h, \cdots, \sum_k P_{i,N_c}^h] = [S_{2,1}^h, \cdots, S_{2,N_c}^h] \tag{6}$$

Since different spatial sub-volumes have different significance to the final recognition, we assign different weights to the their scores to obtain the temporal part score:

$$S_2 = \sum_h w_h S_2^h = [S_{2,1}, \cdots, S_{2,N_c}] \tag{7}$$

In our experiment, we set w_h according to the proportion of pairwise relationships falling in spatial sub-volume h.

The final decision is based on the combination of the results from both spatial and temporal parts. The category label that has the highest score will be assigned to the video:

$$c = \arg\max_{c'}(\alpha S_{1,c'} + (1-\alpha) S_{2,c'}) \tag{8}$$

where $\alpha < 1$ is a parameter to balance the significance of these two parts.

IV. EXPERIMENTS

We evaluate our methods on three public datasets: Weizmann action [1], UCF sports [13] and facial expression dataset [4]. In this section, we first describe the datasets and experimental setting, then give experimental results.

A. Datasets

The Weizmann action dataset contains 9 human actions presented by 10 different people: bend, jumping-jack (jack), jump-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), run, gallop-sideways (side), skip, walk, wave-one-hands (wave1), wave-two-hands (wave2). We use leave-one-subject-out cross validation to evaluate its performance and report the average recognition accuracy.

The facial expression dataset is composed of 4 sub-sets for 6 different emotions by 2 different individuals and 2 lighting setups. The 6 types of expressions are: anger, disgust, fear, joy, sadness and surprise. We use the same experiment framework as [4]. We train our model on one sub-set, containing expressions from one person under one lighting setup and test on all four subsets. The best achievement is reported for all four settings.

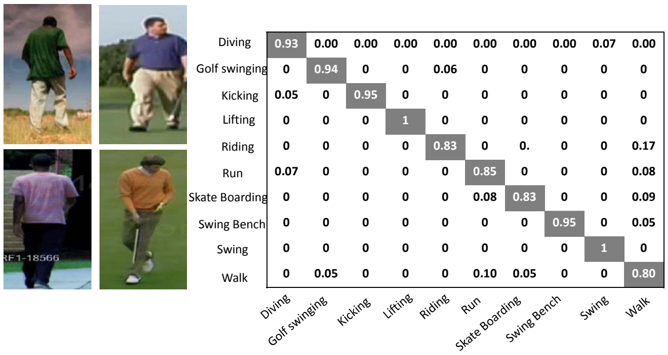
UCF sports dataset contains 150 video sequences of 10 human actions: diving, golf swinging, kicking, lifting, horseback riding, running, skateboarding, swinging on the pommel horse and on the floor, swinging at the high bar and walking. All these actions are various in scenes and view points. We add horizontally flipped version of each sequence to the dataset to extend the amount of samples. We use the evaluation methods in [11] that randomly select 85% of the samples for training and 15% for testing, split the dataset for seven times and report the average results.

B. Experimental settings

For Weizmann dataset, we crop and align the figure centric volumes using the background subtraction masks with the dataset. $15 \times 15 \times 10$ cuboids are densely sampled from 3D spatial temporal volumes. We set $n_x = n_y = 3, n_t = 2$ for each cuboid. Space time interest points are detected using the original setting. Every 3D spatial-temporal volume is cut into 5×2 sub-volumes spatially, containing one dense sampled cuboid in the center of each spacial sub-volume at every sampled temporal coordinate. Such spatial sub-volumes are also used to train separate random forest on space time interest points in temporal part.

For facial expression dataset, we simply extract 90×90 pixels containing the face from all frames and concatenate them to form the 3D space-time volume. We use the same settings as that on Weizmann dataset except the 3D space time volume is cut into 5×5 spatial sub-volumes for densely sampled cuboids training and 2×2 spatial sub-volumes for space time interest points.

For UCF sports dataset, we use the coordinates of figures in all frames together with the dataset to crop the figure volume, then align them into a fixed size of 400×250. Each space-time volume is cut into 5×3 spatial sub-volumes and each cuboid is set to $80 \times 80 \times 10$. We randomly select 85% samples for training and use the rest for testing.



	Diving	Golf swinging	Kicking	Lifting	Riding	Run	Skate Boarding	Swing Bench	Swing	Walk
Diving	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00
Golf swinging	0	0.94	0	0	0.06	0	0	0	0	0
Kicking	0.05	0	0.95	0	0	0	0	0	0	0
Lifting	0	0	0	1	0	0	0	0	0	0
Riding	0	0	0	0	0.83	0	0.	0	0	0.17
Run	0.07	0	0	0	0	0.85	0	0	0	0.08
Skate Boarding	0	0	0	0	0	0.08	0.83	0	0	0.09
Swing Bench	0	0	0	0	0	0	0	0.95	0	0.05
Swing	0	0	0	0	0	0	0	0	1	0
Walk	0	0.05	0	0	0	0.10	0.05	0	0	0.80

Fig. 2. Left: Misclassified examples of walking category. Right: Confusion matrix of our method on UCF sports dataset.

Experiments on all datasets have parameters to tune. In our experiments, we use 100 trees and set the depth of each tree to 20. We run 10 times for each experiment to reduce the influence of the randomness. The weights for different scores are tuned according to the proportion of training samples falling in spatial or temporal sub-volumes. We empirically set α is to 0.5 for the Weizmann and UCF sports datasets and 0.3 for the facial expression dataset.

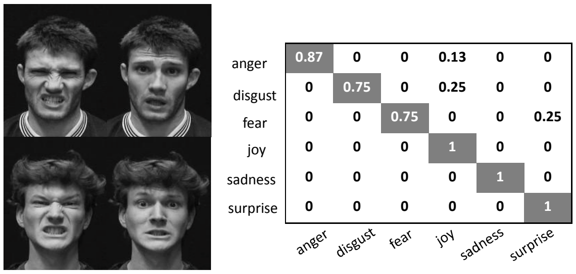
C. *Results*

Human Action Datasets. Our method achieves 100% recognition accuracy on the Weizmann dataset. We do not show the confusion matrix in this paper since it is simply a perfect diagonal matrix. The confusion matrix of our method on the UCF sports dataset is displayed as the right part of Fig. 2. Our method achieves 91.47% accuracy on this dataset. As we observed in the experiment, the "walking" category is easily misclassified. The underlying reason is that they look quite similar to actions in other categories such as run, skateboarding, etc. We show misclassified examples as the left part of Fig. 2.

TABLE I
COMPARISON RESULTS OF ACCURACY ON WEIZMANN AND UCF SPORTS DATASETS.

Method	Weizmann	UCF
Kovashka et al. [7]	-	87.27%
Yao et.al. [19]	95.60%	86.60%
Wang et. al. [17]	-	85.60%
Rodriguez et. al. [13]	-	69.20%
Fathi et. al. [5]	100%	-
Liu et al. [11]	100%	90.10%
Our method	**100%**	**91.47%**

We compare our method with previous methods and show results in Table I on both the Weizmann dataset and UCF sports dataset. Our method achieves 100% accuracy on the Weizmann dataset which is the same with the state-of-the-art methods in [11], [5]. On the UCF sports dataset, our method achieves 91.47% recognition accuracy which is higher than [11]. Compared with [11], our method separately learns spatial and temporal mid-level features, and integrates both low-level appearance and mid-level structural information for recognition. Therefore, our method outperforms other comparison methods.



	anger	disgust	fear	joy	sadness	surprise
anger	0.87	0	0	0.13	0	0
disgust	0	0.75	0	0.25	0	0
fear	0	0	0.75	0	0	0.25
joy	0	0	0	1	0	0
sadness	0	0	0	0	1	0
surprise	0	0	0	0	0	1

Fig. 3. Left: Sample frames of "anger" and "fear" categories. Right: Confusion matrix for training and test on different subjects and the same illumination on facial expression dataset.

Facial Expression Dataset. We test our method on facial expression dataset. Fig. 3 shows some representative frames of categories "anger" and "fear" under one lighting setup and the confusion matrix obtained from the experimental setting that training and testing are completed on different subject and the same illumination.

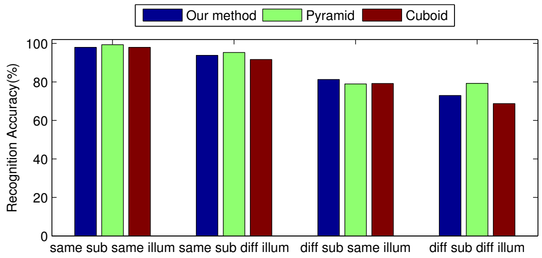


Fig. 4. Comparison results of our method, method in [23] (Pyramid) and method in [4] (Cuboid) on facial expression dataset. We list comparison results in all four experimental settings, where training on one sub-set, test on a sub-set with 1) same subject same illumination (same sub same illum), 2) same subject different illumination (same sub diff illum), 3) different subject same illumination (diff sub same illum), 4)different subject different illumination (diff sub diff illum).

We compare our method with [4] and [23] on facial expression dataset and show comparison results in Fig. 4. The method in [4], simply extracts spatiotemporal interest points and applies bag-of-words model to represent activities. Results show that our method outperforms [4] due to we learn expressive mid-level features to elegantly describe space-time structure. Zhao et al. [23] use a pyramid representation to capture the distribution of point sets in each sub-volume of video. Histograms from all sub-volumes are concatenated to form the feature for the whole video in a homogeneous way in both spatial and temporal perspective. Our method outperforms [23] in the experimental settings that using different subjects for training and testing where modeling spatial and temporal structure information respectively makes our model more robust to variation in local features. For the experimental settings that using the same subject for training and testing, the recognition accuracy of our method is a little lower than [23]. We should note that the method in [23] have different experimental settings from ours , which would significantly affect recognition accuracy. For example, whether some preprocessing (e.g., tracking, background subtraction) is needed, etc. We achieve overall recognition accuracy of 85.15% on facial expression dataset.

Evaluation of different components. To evaluate the effectiveness of the spatial part and temporal part, we compare the recognition accuracy of each part with the complete method. The comparison experiment is conducted on the Weizmann dataset. Comparison results in Fig. 5 show that the complete method outperforms method using one part. We also observe that spatial information and temporal information contribute differently to action classes.

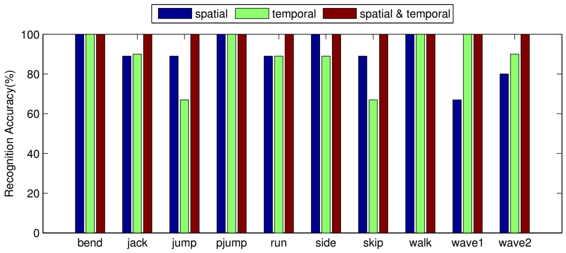


Fig. 5. Comparison results using spatial, temporal part and both.

We also evaluate the performance of mid-level features in recognition. We remove the mid-level features from the full model and obtain the 2-layer mode , which recognizes human actions only by dense low-level features. Another model for comparison experiment is the one with only sparse mid-level features. Comparison results on the Weizmann dataset between the two models and the full model are shown in Fig. 6. It is clear that the full model which uses both low-level feature and mid-level feature outperforms the model only uses low-level features. The mid-level feature learns spatial and temporal structures of action videos and better represents structural information in the videos. Therefore, the recognition accuracy of the full model can be boosted using both appearance information and structural information.

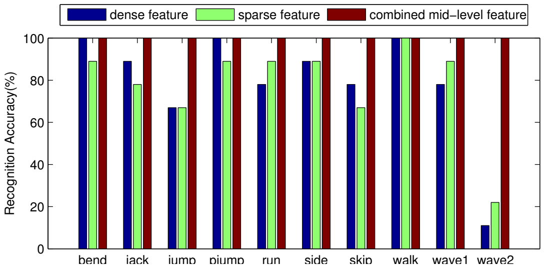


Fig. 6. Comparison results using only low-level and mid-level features.

V. Conclusion

In this paper, we presented a random forest based method for human behavior recognition. Structural and pairwise mid-level features were constructed from densely or sparsely sampled low level HOG/HOF features to represent spatial and temporal sub-volumes of action video. Recognition results of spatial and temporal parts are calculated separately and combined to make the final recognition. We tested our method on Weizmann, UCF and facial expression datasets and showed promising performance of our method.

VI. Acknowledgments

This research is supported in part by the NSF CNS 1135660 and 1314484, Office of Naval Research award N00014-12-1-0125, and Air Force Office of Scientific Research award FA9550-12-1-0201.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *ICCV*, 2, 2005.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *IEEE International Workshop on VS-PETS*, pages 65–72, 2005.

[5] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. *CVPR*, 2008.

[6] P. Kontschieder, S. R. Bulo, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. *ICCV*, pages 2190–2197, 2011.

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

[8] I. Laptev and T. Lindeberg. Space-time interest points. *ICCV*, 1, 2003.

[9] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, pages 1–8, 2008.

[10] V. Lepetit, P. Lagger, and P. Fua. Randomised trees for real-time keypoint recognition. *CVPR*, 2:775–781, 2005.

[11] C. Liu, Y. Kong, X. Wu, and Y. Jia. Action recognition with discriminative mid-level features. *ICPR*, 2012.

[12] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. *CVPR*, 2007.

[13] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. *CVPR*, 2008.

[14] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *ICCV*, pages 1593–1600, 2009.

[15] C. Schapire, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. *ICPR*, 3:32–36, 2004.

[16] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. *CVPR*, pages 2432–2439, 2010.

[17] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatiotemporal features for action recognition. *BMVC*, 2009.

[18] Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic versus max-margin. *PAMI*, 33(7):1310–1323, 2011.

[19] A. Yao, J. Gall, and L. V. Gool. A hough transform-based voting framework for action recognition. *CVPR*, pages 2061–2068, 2010.

[20] A. Yao, J. Gall, and V. Lempitsky. Class-specific hough forests for object detection. *CVPR*, pages 1022–1029, 2009.

[21] G. Yu, A. Norberto, J. Yuan, and Z. Liu. Fast action detection via discriminative random forest voting and top-k subvolume search. *IEEE Transactions on Multimedia*, 13(3):507–517, 2011.

[22] G. Yu, J. Yuan, and Z. Liu. Unsupervised random forest indexing for fast action search. *CVPR*, pages 865–872, 2011.

[23] Z. Zhao and A. Elgammal. Spatiotemporal pyramid representation for recognition of facial expressions and hand gestures. *FG*, 2008.