2014 22nd International Conference on Pattern Recognition

Modeling the Relationship of Action, Object and Scene

Jing Liu, Xinxiao Wu, and Yang Feng
Beijing Laboratory of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing 100081, P.R. China
Email:{liu_jing, wuxinxiao, fengyangbit}@bit.edu.cn

***Abstract*—In the task of action recognition, object and scene can provide rich source of contextual information for analyzing human actions, as human actions often occur under particular scene settings with certain related objects. Therefore, we try to utilize the contextual object and scene for improving the performance of action recognition. Specifically, a latent structural SVM is introduced to build the co-occurrence relationship among action, object and scene, in which the object class label and scene class label are treated as latent variables. Using this framework, we can simultaneously predict action class labels, object class labels as well as scene class labels. Moreover, we use a mid-level discriminative feature to separately describe the information of action, object and scene. The feature is actually a set of decision values from the pre-learned classifiers of each class, measuring the likelihood that the input video belongs to the corresponding class. In this paper, we use SVM as action and scene pre-learned classifiers, and use deformable part-based object detector as the object pre-learned classifier, so that object location can be obtained as a by-product. Experimental results on UCF Sports, YouTube and UCF50 datasets demonstrate the effectiveness of the proposed approach.**

***Keywords*-action recognition; context modeling; object detection; scene recognition; LSSVM.**

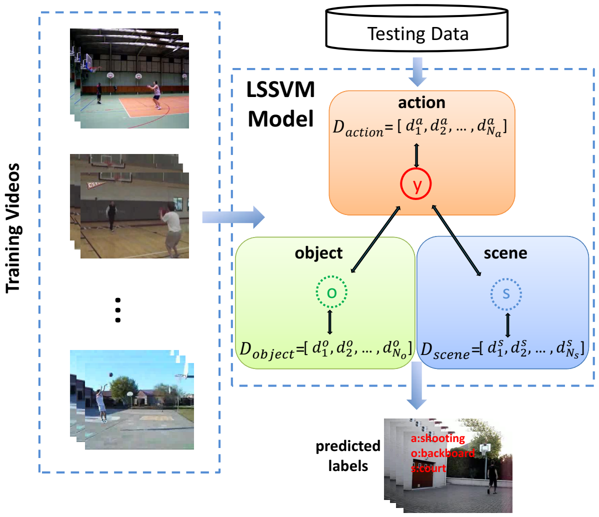

Training Videos
Testing Data
LSSVM Model
action
D_{action}=[$d_1^a, d_2^a, \dots, d_{N_a}^a$]
y
object
o
D_{object}=[$d_1^o, d_2^o, \dots, d_{N_o}^o$]
scene
s
D_{scene}=[$d_1^s, d_2^s, \dots, d_{N_s}^s$]
predicted labels
a:shooting
o:backboard
s:court

Figure 1. Overview of our method. We jointly model the relationship among action, object and scene. The action label, object label and scene label are predicted simultaneously. D_{action}, D_{object} and D_{scene} are the class correlation features. y is the action label while o and s are the latent object label and latent scene label respectively.

I. Introduction

Recognizing human actions is receiving increasing attention due to its wide range of applications. Although impressive results have been reported on datasets collected from controlled environments, recognizing actions in realistic videos from unconstrained environments still remains a challenging problem due to the large appearance variations of human bodies, background clutter and camera movement. In realistic environment, human actions are usually associated with some certain objects and settings, then the co-occurrence relationship can be observed among action, object and scene. For example, the action of "playing basketball" usually happens at the court with the presence of a basketball while the "swing" action often happens outdoors with a child sitting at the swing. Therefore, it is reasonable to utilize the object and scene information as a contextual clue for action recognition. In this paper, we focus on capturing the co-occurrence relationship to discover the mutual contextual constraints among action, object and scene.

We formulate our problem in an LSSVM framework modeling the contextual relationship among action, object and scene. Different from most of previous work [1], [2], we use LSSVM to jointly model the compatibility between features and classes, and the contextual relationship among action classes, object classes and scene classes. Specifically, our model contains five parts: action class model, object class model, scene class model, action-object interaction model and action-scene interaction model. Moreover, the object label and scene label are treated as latent variables which can be inferred implicitly during both learning and inference processes.

To describe the motion, object and scene information in the videos, we utilize a set of mid-level discriminative features. Due to the biological and kinematic constraints, some different action classes may share certain similar motion patterns of human parts. Furthermore, some different objects may have similar appearance or shape patterns and some different scenes may also share similar texture data distributions. So it is beneficial to develop a descriptor to capture the class correlations. Here we utilize a set of decision values from the pre-learned classifiers of each class as class correlation feature [3]. Due to the high level relationship among different classes, this mid-level feature exhibits more discriminative power and robustness than the low-level feature. An overview of our method is illustrated in Fig. 1.

We test our approach over the UCF Sports dataset [4], the YouTube dataset [5] and the UCF50 dataset [6], and the experimental results demonstrate that the proposed framework works effectively in human action recognition.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.350

II. Related Work

Modeling the action and contextual information interaction has been explored in several recent works for human action recognition. Moore et al. [7] employed belief networks for modeling the relationship between human actions and objects context to perform action recognition. Gupta et al. [8] tried to segment and recognize actions by applying context on human movements from the objects. Wu et al. [9] proposed a object-use based method for activity recognition in the kitchen. In this work, a dynamic Bayesian network model is applied in to jointly infer the most likely activity and object labels. Tran et al. [10] tried to explicit common sense rules for event recognition in the parking lot using the Markov logic networks (MLN).

Different from [7], [8], [9], [10], some methods [11], [12], [13], [14] focus on the video data collected from the uncontrolled environment. Sun et al. [11] modeled action context in a feature representation that is designed to capture the proximity of local keypoint tracks. Marszalek et al. [12] used a text-mining approach to discover the relations between action and scene from movie script. Han et al. [13] used contextual scene descriptors and Bayesian multiple kernel learning methods for realistic action recognition. Yao et al. [14] proposed a random field model to encode the objects and human poses in interaction activities. Wu et al. [3] modeled the scene and action interaction by integrating the scene and action co-occurrence relationship into a latent SVM. Motivated by this method, we consider the action-object-scene interaction and try to model the action and contextual information for action recognition.

Unlike all of these approaches, we model both object and scene contexts for action recognition in realistic world with a large categories of actions. Our work is similar to [1] and [2]. In [1], multiple features of people, objects and scene are combined for classification using linear SVM and multiple kernel learning methods in a multiple instance learning framework. Different from their work, we explicitly exploit the action-object-scene interaction by integrating the contextual co-occurrence relationship among them. In [2], the object classifier and scene classifier are separately learned and these priori information are treated as cues for the relationship of action, object and scene. In our work, not only the action classifier, the scene classifier and object classifier, but also the interaction between action and scene and interaction between action and object are jointly modeled in a unified framework. Therefore, they can be learned simultaneously during the optimization process. Moreover, we treat the object label and scene label as latent variables and do not require the ground truth of the object label and scene label in the training data.

III. Modeling Action, Object and Scene Interaction using LSSVM

A. Model Formulation

Let $x \in \mathcal{X}$ be the input action video and $y \in \mathcal{Y}$ be the output action label. Our model aims to learn a discriminative function $F(x, y)$ which measures how compatible the action label y is suited to an input video x. In our work, the input-output relationship also depends on the unobserved object label $o \in \mathcal{O}$ and latent scene label $s \in \mathcal{S}$. Accordingly, the discriminative function takes the following form:

$$\begin{aligned} F(x,y) &= \max_{y\in\mathcal{Y}} \max_{o\in\mathcal{O}} \max_{s\in\mathcal{S}} f_w(x,y,o,s), \\ f_w(x,y,o,s) &= w^T\Phi(x,y,o,s), \end{aligned}$$

where w is the learned parameter of the model including five parts $w = \{w_a; w_o; w_s; w_{ao}; w_{as}\}$ and $\Phi(x,y,o,s)$ is a joint feature vector which describes the relationship among the input action video x, the output action label y, the latent object label o and the latent scene label s. $w^T\Phi(x,y,o,s)$ is defined as

$$\begin{aligned} w^T\Phi(x,y,o,s) =& w_a^T\Phi_a(x,y) + w_o^T\Phi_o(x,o) \\ &+ w_s^T\Phi_s(x,s) + w_{ao}\Phi_{ao}(y,o) \\ &+ w_{as}\Phi_{as}(y,s). \end{aligned}$$

1) Action Class Model $w_a^T\Phi_a(x,y)$: This potential function measures the compatibility between an action video x and an action label y, defined by

$$w_a^T\Phi_a(x,y) = \sum_{i=1}^{N_a}\sum_{j=1}^{T_a} {w_{ij}^{a}}^T \phi_j^a(x) I_i(y). \tag{1}$$

N_a and T_a respectively denote the number of action classes and that of action feature types. $\phi_j^a(x) \in R^{N_a}$ represents the j-th type of action feature. The parameters $w_{ij}^a \in R^{N_a}$ is the weight vector for the feature $\phi_j^a(x)$. $I_i(y)$ is an indicator function, namely, $I_i(y) = 1$ if $y = i$, and $I_i(y) = 0$ otherwise.

2) Object Class Model $w_o^T\Phi_o(x,o)$: This potential function measures the compatibility between an action video x and an object label o, defined by

$$w_o^T\Phi_o(x,o) = \sum_{i=1}^{N_o} {w_i^o}^T \phi^o(x) I_i(o). \tag{2}$$

N_o denotes the number of object classes. $\phi^o(x) \in R^{N_o}$ represents the object feature. The parameters $w_i^o \in R^{N_o}$ is the weight vector for the feature $\phi^o(x)$ and $I_i(o)$ is an indicator function.

3) Scene Class Model $w_s^T\Phi_s(x,s)$: This potential function measures the compatibility between an action video x and a scene label s, defined by

$$w_s^T\Phi_s(x,s) = \sum_{i=1}^{N_s} {w_i^s}^T \phi^s(x) I_i(s). \tag{3}$$

N_s denotes the number of scene classes. $\phi^s(x) \in R^{N_s}$ represents the scene feature. The parameters $w_i^s \in R^{N_s}$ is the weight vector for the feature $\phi^s(x)$ and $I_i(s)$ is an indicator function.

4) Action-Object Interaction Model $w_{ao}\Phi_{ao}(y,o)$: This potential function represents the contextual co-occurrence relation between an action label y and an object label o. It is parameterized as

$$w_{ao}\Phi_{ao}(y,o) = \sum_{i=1}^{N_a}\sum_{j=1}^{N_o} w_{ij}^{ao} I_i(y) \cdot I_j(o), \tag{4}$$

where the parameter w_{ao} is a $N_a \times N_o$ matrix with each entry w_{ij}^{ao} indicating how likely the action class is i and the object class is j.

5) Action-Scene Interaction Model $w_{as}\Phi_{as}(y,s)$: This potential function represents the contextual co-occurrence relation between an action label y and a scene label s. It is parameterized as

$$w_{as}\Phi_{as}(y,s) = \sum_{i=1}^{N_a}\sum_{j=1}^{N_s} w_{ij}^{as} I_i(y) \cdot I_j(s), \tag{5}$$

where the parameter w_{as} is a $N_a \times N_s$ matrix with each entry w_{ij}^{as} indicating how likely the action class is i and the scene class is j.

B. Learning

Given a set of training examples $\{(x_i, y_i)_{i=1,2,...,N}\}$, our goal is to learn the model parameter w. Since the object labels $\{o_i\}_{i=1:N}$ and the scene labels $\{s_i\}_{i=1:N}$ are unobserved and treated as latent variables, our optimization problem is formulated in a latent structural SVM framework for learning:

$$\min_{w,\xi_i} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i, \tag{6}$$

$$\begin{aligned} s.t. \quad & \max_{o,s} w^T\Phi(x_i,y_i,o,s) - \max_{o,s} w^T\Phi(x_i,y,o,s) \\ & \geq \triangle(y_i,y) - \xi_i, \ \xi_i \geq 0 \ \forall i \ \forall y, \end{aligned} \tag{7}$$

where $\Delta(y_i,y)$ is a loss function measuring the cost incurred by predicting the ground-truth label y_i as y. Here we use a simple 0-1 loss as $\Delta(y_i,y) = 1$ if $y_i \neq y$, and $\Delta(y_i,y) = 0$ otherwise. The constraint in Eq. (7) can be explained as follows: for the i-th training sample, the score $\max_s w^T\Phi(x_i,y_i,o,s)$ associated with the ground-truth class label y_i should be no less than the score $\max_s w^T\Phi(x_i,y,o,s)$ associated with any hypothesized class label y. The constrained optimization problem in Eq. (6) and Eq. (7) can be rewritten as an unconstrained problem:

$$\min_{w,\xi_i} \frac{1}{2}w^T w + C\sum_{i=1}^{N}(M_i - G_i), \tag{8}$$

where

$$M_i = \max_{y,o,s}(\Delta(y_i,y) + w^T\Phi(x_i,y,o,s)), \tag{9}$$

$$G_i = \max_{o,s} w^T\Phi(x_i,y_i,o,s). \tag{10}$$

We employ the convex bundle optimization technique in [15] to solve Eq. (8). Specifically, this optimization algorithm iteratively builds an increasingly accurate piecewise quadratic approximation of Eq. (8) and converges to an optical solution of w, which requires the calculation of the subgradient of $M_i - G_i$. Suppose

$$(\tilde{y},\tilde{o},\tilde{s}) = \arg\max_{y,o,s}(\Delta(y_i,y) + w^T\Phi(x_i,y,o,s))$$

and

$$(o',s') = \arg\max_{o,s} w^T\Phi(x_i,y_i,o,s),$$

then $\partial_w(M_i - G_i)$ can be calculated as

$$\partial_w(M_i - G_i) = \Phi(x_i,\tilde{y},\tilde{o},\tilde{s}) - \Phi(x_i,y_i,o',s').$$

Note that the inferences on Eq. (9) and Eq. (10) can be respectively solved via the enumeration of all the possible label pairs (y,o,s) and all the possible label pairs (o,s).

C. Inference

With the learned parameter w, the inference problem is to simultaneously find the optimal action label y^*, object label o^* and scene label s^* for a test video x. The inference is equal to the following optimization problem:

$$(y^*,o^*,s^*) = \arg\max_{y,o,s} w^T\Phi(x,y,o,s). \tag{11}$$

For simplicity, we can solve Eq. (11) by directly enumerating all the possible action-object-scene label pairs (y,o,s) for a test video x. The values of y, o and s are respectively set from 1 to N_a, 1 to N_o and 1 to N_s.

IV. Representations of Action, Object and Scene

As the limited discriminative capability of the extracted low-level visual features, we utilize a mid-level feature, called class correlation feature proposed by [3], which captures the correlations between different classes, to abstract the visual content of video. This class correlation feature is represented by a set of decision values from all the pre-learned classifiers, which can represent the semantic meaning to some extent.

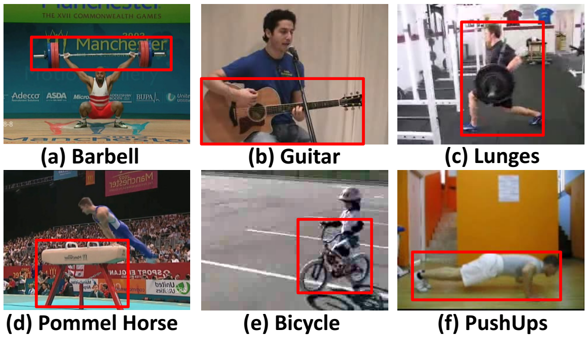
A. Action Class Correlation Feature

To train the pre-learned action classifiers, we extract five types of low-level visual features from the action videos. We first extract the spatio-temporal context distribution feature proposed by [16] considering the spatio-temporal contextual information of interest points. Complementally, we extract the appearance information from the cuboids around the interest points. Moreover, three dense trajectory features (i.e. trajectory, HOG, MBH) proposed by Wang et al. [17] are also extracted for further improvement of recognition performance.

Based on the five types of low-level features, we trained five independent SVMs for each action class to produce the decision values. We denote $f_c^k(x)$ as the c-th action class pre-learned classifier from the k-th type of low-level feature and x as the action video. Using the k-th type of low-level feature, the likelihood that the video x belongs to the c-th action class is described by the classification score $d^k = f_c^k(x)$. Then the action class correlation feature of x for the k-th type of low-level feature is represented by

$$D_{action}^k = [d_1^k, d_2^k, ..., d_{N_a}^k]^T \in R^{N_a},$$

where N_a is the number of action classes. Finally, we simply concatenate D_{action}^k from all types of low-level features to construct the action class correlation feature of x.



(a) Barbell (b) Guitar (c) Lunges (d) Pommel Horse (e) Bicycle (f) PushUps

Figure 2. Some examples of the defined object classes. (c) and (f) are the specific human gestures which are the important objects of interest.

B. Object Class Correlation Feature

We employ the deformable part model (DPM) object detector [18] as our object pre-learned classifier. Firstly, we define a set of distinctive objects. Note that, in some cases like "Lunges" and "PushUps", the person with specific gesture is obviously the most important object of interest. Fig. 2 shows some examples of our defined object classes.

To train these object models, we randomly select 20-50 frames as positive examples from each class which contain the corresponding object. To detect objects in an input video x, we randomly select K ($K \leq 5$) frames and detect objects using all the pre-learned object detectors. As the limited test examples from each video, we use max pooling in our experiments to be robust to noise. Finally, we associate the video with the frame with the highest score, and use the highest score as the score of the whole video. Let us denote $f_c(x^l_{img})$ as the pre-learned detector of the c-th object class and x^l_{img} as the l-th (l=1, 2, ..., K) test frame randomly selected from the input video x. The likelihood that the frame x^l_{img} contains the c-th object class is described by the detection score $d^l = f_c(x^l_{img})$. Then the object class correlation feature of x is represented by

$$D_{object} = [\max_l d^l_1, \max_l d^l_2, ..., \max_l d^l_{N_o}]^T \in R^{N_o},$$

where N_o is the number of object classes.

C. Scene Class Correlation Feature

We extract the local SIFT features [19] to describe the visual information of scene from randomly selected frames in the video and use the bag-of-words model. Similar to the progress of obtaining action class correlation feature, we train the binary linear SVMs for each scene class to get the decision values of an input video x. Then the scene class correlation feature of x can be represented by

$$D_{scene} = [d_1, d_2, ..., d_{N_s}]^T \in R^{N_s},$$

where N_s is the number of scene classes. Note that in the training process, we simply assume that the scene class labels are the same with the action class labels.

V. Experiments

A. Datasets and Settings

The UCF Sports dataset [4] contains 149 real videos of ten different types of sports actions. Videos in this dataset are extracted from sports broadcasts with large intra-class variabilities. We extend the dataset by adding a horizontally flipped version of each video sequence as suggested in [20] to increase the amount of training samples. In the leave-one-sample-out cross validation setting, one original video sequence is used as the test data while the rest original video sequences together with their flipped versions are employed as the training data. Following [20], the flipped version of the test video sequence is not included in the training set.

The YouTube dataset [5] contains 1168 videos divided into 11 action classes. The videos for each class are divided into 25 related groups with each group consisting of 4 or more than 4 videos. Following [5], we use the leave-one-group-out cross validation setting for training and testing.

The UCF50 dataset [21] is one of the most difficult datasets for action recognition. It contains 50 action classes and each action class consists of 25 to 30 related action groups. In our experiment, we just use the first 25 groups with at least 100 videos for each class. We adopt the leave one-group-out cross validation over these groups similar to the YouTube dataset, as suggested in [5].

For the spatio-temporal context distribution and appearance features of action, the spatial and temporal scale parameters σ and τ are respectively set by $\sigma = 2$ and $\tau = 2.5$. The size of cuboid is empirically fixed as $7 \times 7 \times 5$ and 1000 interest points are extracted from each video. For the three dense trajectory descriptors of action [17] (i.e., trajectory, HOG, MBH), we use the bag-of-words approach and the number of visual words per descriptor is fixed to 4000. For the class correlation feature of object, the number of test frames randomly selected from each video is fixed to 4 (i.e., $K = 4$) to limit the complexity. That is to say we detect objects on the four frames to obtain the score and use the max pooling as the score of the whole video. For the SIFT feature of scene, we extracted the SIFT descriptors from the 20% frames randomly selected from each video and the number of SIFT words is fixed to 2000. The parameter C in LSSVM is fixed as the default value (i.e., C = 1) as in SVM.

B. Experimental Results

Recognition Results Using Different Information: Table I lists the recognition accuracies using different information on the datasets of UCF Sports, YouTube and UCF50. The first three rows indicate the performance of the individual correlation feature of action, object and scene, respectively. The results show that, even using the single object feature or scene feature can produce a recognition accuracy more than 40%, from which we can infer that object feature and scene feature can provide much useful information for the possible action.

The last row in Table I shows the results using multiple features and their contextual co-occurrence relationship. "Ac-

Table I
ACCURACIES ON THREE DATASETS

	UCF Sports	YouTube	UCF50
Action	91.35%	84.37%	83.63%
Object	82.18%	56.06%	54.84%
Scene	73.50%	54.71%	40.34%
Action+Object+Scene	93.14%	86.62%	87.14%

Table II
ACCURACIES OF DIFFERENT METHODS ON THE UCF SPORTS DATASET

Methods	Recognition Accuracy
Wang et al. [20].	85.6%
Wang et al. [17]	88.2%
Le et al. [4].	86.5%
Shabani et al. [22]	91.5%
Sadanand and Corso [23]	95.0%
Wu et al. [16]	91.3%
Kovashka and Grauman [24]	87.27%
Wu et al. [3]	92.48%
our method	93.14%

tion+Object+Scene" refers to the combination of action, object and scene information. These results demonstrate that in most of the cases, although the object feature and the scene feature is not as effective as the action feature, the combination of them outperforms anyone of them and can improve the performance significantly.

Comparison with Other Methods: We compare our method with other methods on three datasets, and the results are shown in Tables II, III and IV. In Table II, we report the recognition accuracies of the state-of-the-art methods on UCF Sports dataset. These recent methods [24], [20], [17], [4], [22], [23], [16] employ the same leave-one-sample-out cross validation setting as ours. The results clearly show that our method outperforms most of the state-of-the art methods and is comparable to [23], though they use additional datasets for training. In Table III, we report the recognition accuracies of the state-of-the-art methods which adopt the same leave-one-group-out cross validation strategy on the YouTube dataset and our method is comparable to the best results in [3] without using the multiple low-level visual features. In Table IV, we show the recognition results on the UCF50 dataset. It is obvious that on this complex and challenging action dataset with 50 action classes and large intra-class variations, our method outperforms the state-of-the-art methods.

Table III
ACCURACIES OF DIFFERENT METHODS ON THE YOUTUBE DATASET

Methods	Recognition Accuracy
Liu et al. [5]	71.2%
Wang et al. [17]	84.2%
Le et al. [4]	75.8%
Bregonzio et al. [25]	64.0%
Ikizler-Cinbis et al. [1]	75.21%
Wu et al. [3]	87.01%
our method	86.62%

Table IV
ACCURACIES OF DIFFERENT METHODS ON THE UCF50 DATASET

Methods	Recognition Accuracy
Sadanand et al. [23]	57.9%
Kliper-Gross et al. [21]	72.6%
Reddy et al. [17]	76.9%
Wang et al. [26]	78.4%
Wang et al. [27]	85.6%
Wang et al. [28]	85.7%
Wu et al. [3]	85.87%
our method	87.14%

Visualization: Fig. 3 shows some examples of the predicting results on the three public datasets. During testing, the action label, the object label and the scene label can be predicted simultaneously using our LSSVM framework. Furthermore, the object pre-learned classifiers can detect the object location, so that using the predicted object label, object location can be obtained as a by-product.

VI. CONCLUSION

In this paper, we have presented an action recognition approach which models the co-occurrence relationship of action, object and scene. We utilize a latent structural SVM to build the interaction treating object label and scene label as latent variables. By using contextual object and scene information, the performance of action recognition is significantly improved in complex realistic videos. Experimental results on UCF Sports, YouTube, and UCF50 datasets have demonstrated the effectiveness of our approach.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundations of China (NSFC) under Grant No.61203274, the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission, and the Excellent Young Scholars Research Fund of BIT (2013).

REFERENCES

[1] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *European Conference on Computer Vision*. Springer, 2010, pp. 494–507. 1, 2, 5

[2] Y.-G. Jiang, Z. Li, and S.-F. Chang, "Modeling scene and object contexts for human action retrieval with few examples," *Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 674–681, 2011. 1, 2

[3] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm," *Circuits and Systems for Video Technology*, vol. 23, no. 8, pp. 1422–1431, 2013. 1, 2, 3, 5

[4] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3361–3368. 1, 4, 5

[5] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1996–2003. 1, 4, 5

[6] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013. 1

[7] D. J. Moore, I. A. Essa, and M. H. Hayes III, "Exploiting human actions and object context for recognition tasks," in *International Conference on Computer Vision*, vol. 1. IEEE, 1999, pp. 80–86. 2

[8] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8. 2

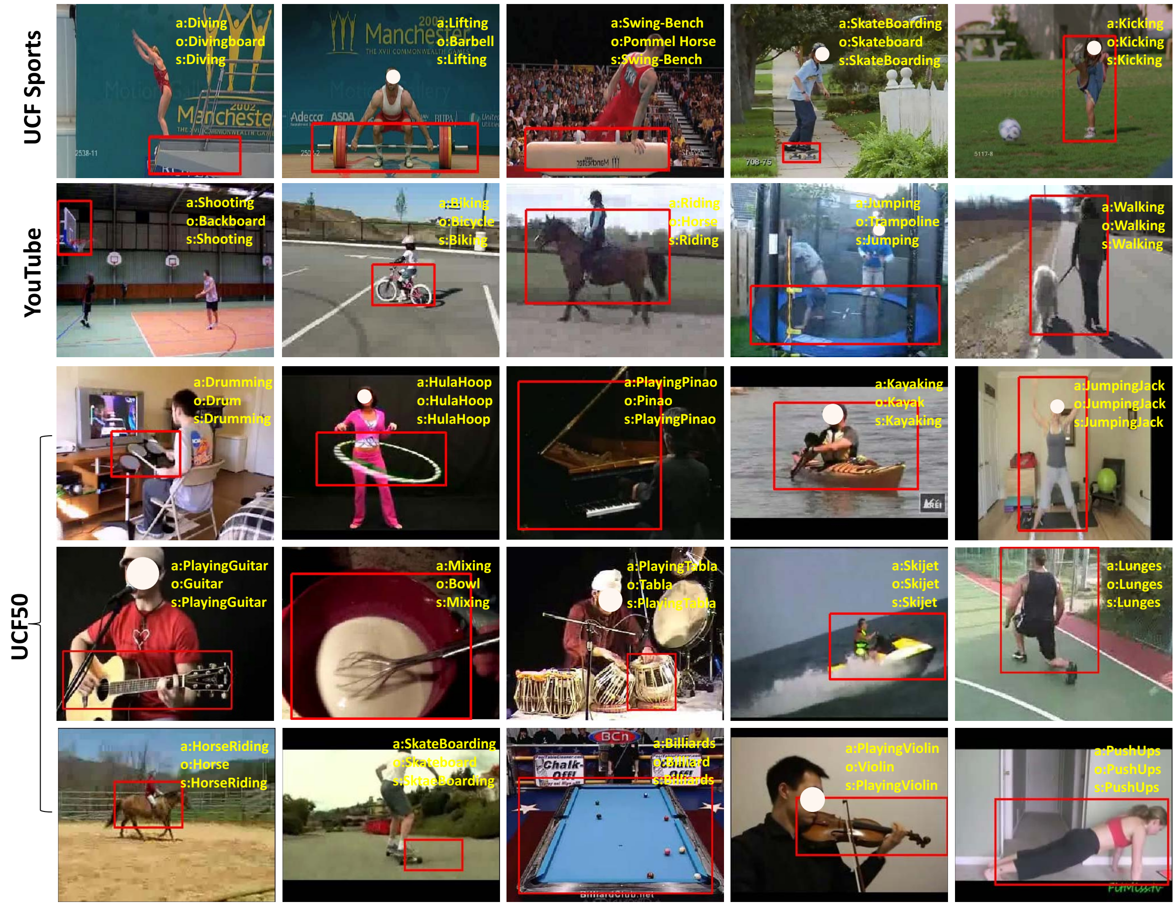


Figure 3. Some examples of our recognition results. The text shows the predicting action labels, object labels and scene labels respectively and the bounding boxes show the localization results of object detection. The last column shows the results with specific human gestures detected as described in section IV-B.

[9] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8. 2

[10] S. D. Tran and L. S. Davis, "Event modeling and recognition using markov logic networks," in *European Conference on Computer Vision*. Springer, 2008, pp. 610–623. 2

[11] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2004–2011. 2

[12] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2929–2936. 2

[13] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *International Conference on Computer Vision*. IEEE, 2009, pp. 1933–1940. 2

[14] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 17–24. 2

[15] T.-M.-T. Do and T. Artières, "Large margin training for hidden markov models with partially observed states," in *Annual International Conference on Machine Learning*. ACM, 2009, pp. 265–272. 3

[16] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 489–496. 3, 5

[17] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3169–3176. 3, 4, 5

[18] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8. 4

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 4

[20] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid *et al.*, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009. 4, 5

[21] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, pp. 1–11, 2012. 4, 5

[22] A. H. Shabani, D. A. Clausi, and J. S. Zelek, "Improved spatio-temporal salient feature detection for action recognition." in *British Machine Vision Conference*, 2011, pp. 1–12. 5

[23] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1234–1241. 5

[24] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2046–2053. 5

[25] M. Bregonzio, J. Li, S. Gong, and T. Xiang, "Discriminative topics modelling for action feature selection and recognition." in *British Machine Vision Conference*, 2010, pp. 1–11. 5

[26] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3d parts for human motion recognition." 5

[27] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, pp. 1–20, 2013. 5

[28] L. Wang, Y. Qiao, and X. Tang, "Mining motion atoms and phrases for complex action recognition," in *International Conference on Computer Vision*. IEEE, 2013. 5