Human-Object Interaction Recognition by Modeling Context

Qun Zhang(✉), Wei Liang, Xiabing Liu, and Yumeng Wang

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, People's Republic of China
{zhangqun,liangwei,liuxiabing,wangyumeng}@bit.edu.cn

Abstract. In this paper, we present a new method to recognize human-object interactions by modeling the context between human actions and manipulated objects. It is a challenging task due to severe occlusion between human and objects during the interacting process. While human actions and objects can provide strong context information, such as some action happening is usually related to a certain object, by which we can improve the accuracy of recognition for both of them. In this paper, we use global and local temporal features from skeleton sequences to model actions, and kernel features are applied to describe objects. We optimize all possible solutions from actions and objects by modeling the context between them. The results of experiments show the effectiveness of our method.

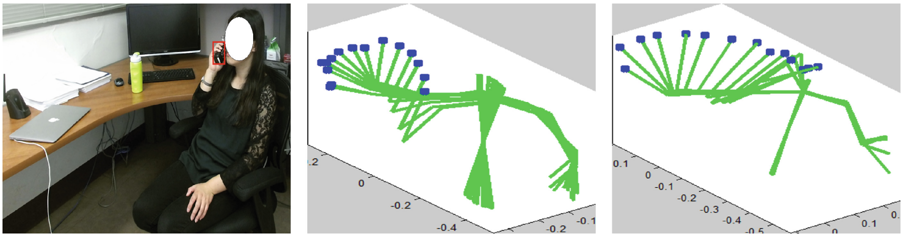
Keywords: Human-object interaction · Action recognition · Object classification · Context

1 Introduction

Human-object interaction recognition has been studied in the field of computer vision for years and it has a broad prospective application. Although human can distinguish interactions easily, it is still a difficult task for computers. The reasons are: (1) The appearance of objects and human vary a lot due to occlusion between objects and human during interactions, which leads to the failure of recognition. As shown in Fig. 1(a), objects are occluded by hand. In this situation, it is challenging for object recognition only by appearance. (2) There are ambiguities in human actions if we only consider the pose from a single frame. Even for pose sequences, it is not easy. Sometimes different actions have similar pose sequences. In Fig. 1(b) and (c), *"calling"* and *"drinking"* can not be separated well from skeleton sequences.

Inspired by the work of Yao and Fei-Fei [22], we propose a method to recognize human-object interactions by modeling the context between human actions and manipulated objects with RGBD videos which are captured by Kinect sensor. Firstly, we train a classifier by SVM algorithm for actions with pose sequence features, which can score each action, and then we search all possible manipulated objects by a sliding window near human hand region. We keep all possible

Y.-J. Zhang (Ed.): ICIG 2015, Part II, LNCS 9218, pp. 385–395, 2015.
DOI: 10.1007/978-3-319-21963-9_35



(a) Occluded object (b) Calling (c) Drinking

Fig. 1. Human-object interaction examples. (a) The manipulated objects are small and occluded in human-object interactions, (b) and (c) are ambiguous with similar pose sequences.

interpretations of action and object labels. By modeling the context between them, we get the most reasonable results of human actions and manipulated objects with optimization. Our framework is shown in Fig. 2.

The rest of this paper is organized as follows. In Sect. 2, we introduce the related works. In Sect. 3, more details of our method are presented. In Sect. 4, our dataset and experimental results are described. Finally, the paper is concluded in Sect. 5 with discussion.

2 Related Work

Action Recognition. Action recognition is very important in the field of computer vision. Researchers have proposed many different approaches to solve this problem. Most of traditional methods focused on 2D data. They used features like silhouettes and shapes to extract spatio-temporal feature descriptors to recognize actions [4,6,17,19]. Raptis and Soatto [17] proposed a hierarchical structure which included SIFT average descriptor, trajectory transition descriptor, and trajectory proximity descriptor. Some works [13,14] modeled actions with the coordinates of skeletons or the relative position of body parts. The approach in [13] extracted local joints structure as local skeleton features and histograms of 3D joints as global skeleton features for action recognition. In most of the above work, actions are performed without occlusion. When occlusions happen, the features usually tend to fail. It is one of the most difficult issues in many kinds of vision work.

Object Features. The "good" features are very important for object classification. Researchers have presented various kinds of features. For low-level image features, SIFT [8,15] and HOG [7,25] are the most popular features in vision tasks. Many researchers adopted multiple kinds of features to represent various aspects of objects for classification in [1,2,11]. Ito and Kubota [11] introduced three different co-occurrence features named color-CoHOG (color-Co-occurrence Histograms of Oriented Gradients), CoHED (Co-occurrence Histograms of pairs of Edge orientations and color Difference), and CoHD (Co-occurrence Histograms of color Difference) to classify objects. Benefitted from the performance of 3D sensors, such as the Kinect sensor, some researchers [1,2] extracted features from RGBD data. Bo et al. [2] presented five depth kernel descriptors that captured different cues including size, shape and edges. Some other approaches used a set of semantic attributes to classify objects [12,18]. Su et al. [18] used five groups of semantic attributes including scene, color, part, shape, and material, they demonstrated that the semantic attributes can be helpful to improve the performance of object classification.

Context. Psychology experiments have shown that context plays an important role in recognition in the system of human vision. It has been used for some tasks in computer vision, such as object detection, object classification, action recognition, scene recognition, and semantic segmentation [10,16,19,20,23]. Marszalek et al. [16] claimed human actions have relations to the purpose and scene. Hence, they modeled the context between human actions and scenes to recognize actions. Sun et al. [19] adopted a hierarchical structure to represent the spatio-temporal context information for action recognition. In [20], they proposed a 4D human-object interaction model to recognize the events and objects in the video sequences. Gupta et al. [10] combined the spatial with function constraint between human and objects to recognize actions and objects. Yao et al. [23] modeled the mutual context of objects and human poses in human-object interaction activities.

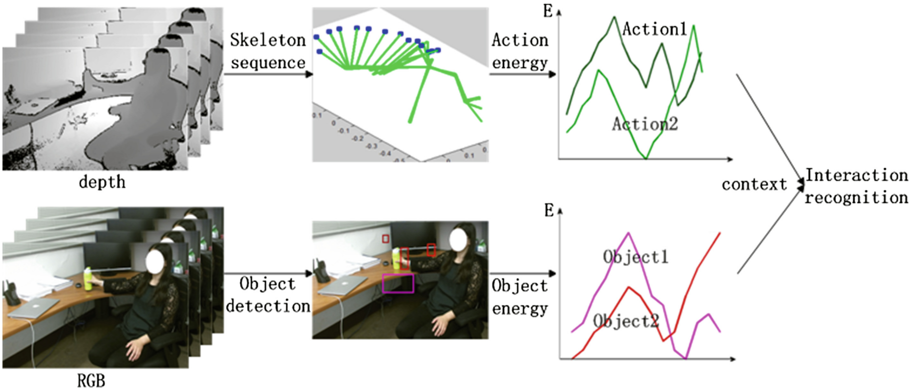


Fig. 2. An overview of our framework

3 Modeling Context of Actions and Objects

We define human-object interactions as $I =< A, O >$. Given a RGBD video V in time interval $T = [1, T]$, our goal is to recognize manipulated objects O and actions A respectively. We cast recognition into an optimization problem:

$$E(I,V) = \sum_{t=1}^{T}(\lambda_1 E(A,V_t) + \lambda_2 E(O,V_t) + E(A,O)). \tag{1}$$

Where $E(A,V_t)$ is the energy of human actions in temporal space. $E(O,V_t)$ is the energy of manipulated objects in spatial space. $E(A,O)$ is the energy of context between human actions and manipulated objects. λ_1 and λ_2 are weights to balance the contribution of each energy term.

3.1 Action Energy

$E(A,V_t)$ models the energy of human actions. We train a classifier by multi-class SVM [5] to score each skeleton sequence as class a. The energy is:

$$E(A,V_t) = \omega_a \theta_a. \tag{2}$$

Where ω_a is the template parameter of action class a. θ_a is the skeleton features of human action.

As shown in Fig. 3, we improve the method of global features in [21], histograms of 3D joints (HOJ3D), by combining local features. This feature is computed with aligned spherical coordinate system and it is independent of views. We denote features as $\theta_a = \{F_L, F_G\}$. The features include two parts: *local features* $F_L = (l_1, l_2, ..., l_{N_l})$ and *global features* $F_G = (g_1, g_2, ..., g_{N_g})$. For global feature, 3D space is firstly divided into many bins and we count how many skeletons are in these bins. It describes the overall statistical information about human skeletons distribution without local features. Besides the global features, we also extract another feature to describe the local structure of skeletons, which uses the triangle areas of every three skeletons. In order to obtain more robust and dense features, we apply linear discriminant analysis (LDA) to reduce the dimension of features space. LDA can extract dominant features and produce the best discrimination between classes by searching in the subspace.

3.2 Object Energy

$E(O,V_t)$ models the energy of object label. Extracting discriminative features is critical for object classification. Various features are explored by researchers [3,9,24]. In this paper, we apply kernel descriptors [1] to model objects. We use gradient, color, and local binary pattern match kernel descriptors to turn images into patch-level features. We sample the image patches with different sizes, such as 4 × 4 rectangle or 8 × 8 rectangle. Then, the gradient match kernel is used to measure gradient orientations similarity between patches from two different images. In the same way, the color match kernel and local binary pattern match kernel can represent image appearance and local shape respectively. We visualize these three types of kernel features in Fig. 4. As the number of image patches is large and evaluating kernels is time consuming, we utilize kernel principal component analysis to extract compact basis vectors. In human-object interactions, object is often small and occluded. Fortunately, benefitted from the Kinect, we

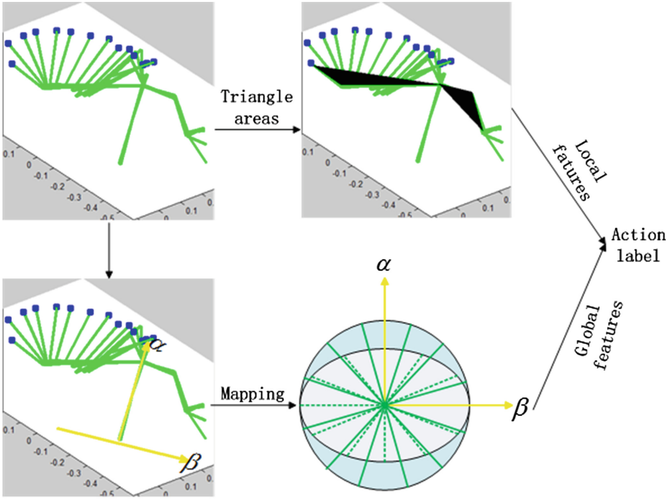


Fig. 3. The features for action recognition. (1) Extracting local features by computing triangle areas. (2) Extracting global features by mapping the skeleton coordinates into spherical bins.

can get more stable and reliable human pose. So we use sliding window to detect object near human hand region with several sizes. For each window, we extract above features and then compute the cost that assigning the label o to the object by a linear SVM classifier. The energy is defined as:

$$E(O, V_t) = \omega_o \theta_o. \tag{3}$$

Where ω_o is the template parameter of object class o. θ_o is the kernel features of manipulated object.

3.3 Context Between Actions and Objects

Human can recognize human-object interactions easily even some information is missing. There is the context between action and related object in human mind. The knowledge of context is helpful to get the most reasonable interpretation for action and the manipulated object together. For example, when someone is drinking, severe occlusion usually happened because of human's hand holding the cup. But, for human, it is easy to fill in the information of the object. It is supposed that there exists a cup in human hand when the human is doing the action of *"drink"*. Some kinds of action has a higher possibility to manipulate certain objects. That is to say, we can infer the human action with the related object and vice versa. The context between human actions and manipulated objects is defined as:

$$E(A, O) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_O} N_{(A=i)} N_{(O=j)}. \tag{4}$$

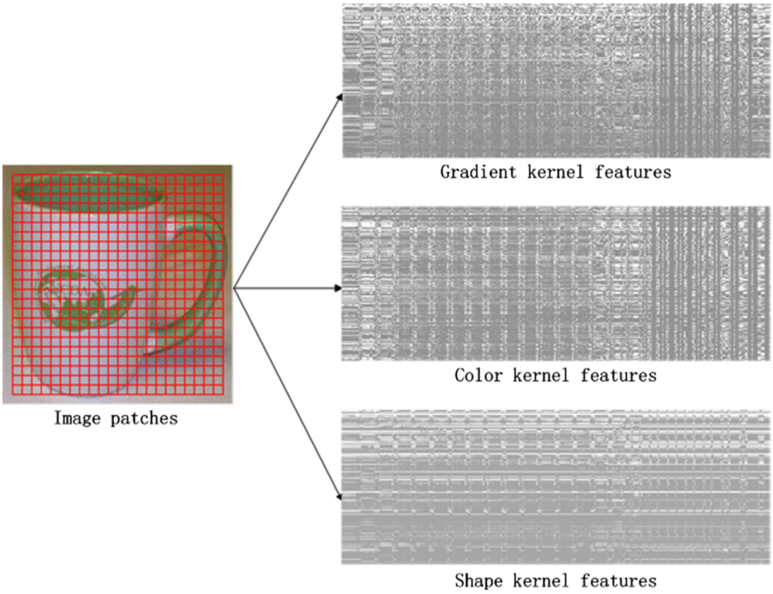


Fig. 4. The gradient, color, and shape kernel features of the cup (Color figure online)

Where N_A and N_O represent the number of action classes and object classes respectively. $N_{(A=i)}$ and $N_{(O=j)}$ represent the number of the ith category action and the jth category object respectively.

3.4 Learning and Inference

We adopt SVM algorithm to learn the action model parameter ω_a, defined by:

$$\underset{\omega,\xi_i}{min} \quad \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{N}\xi_i. \tag{5}$$

$$\text{s.t.} \quad y_i(\omega\theta + b_i) \geq 1 - \xi_i.$$

Where y_i is the label for data x_i. θ is the feature for data x_i. Similarly, we can learn the object model parameter ω_o by SVM.

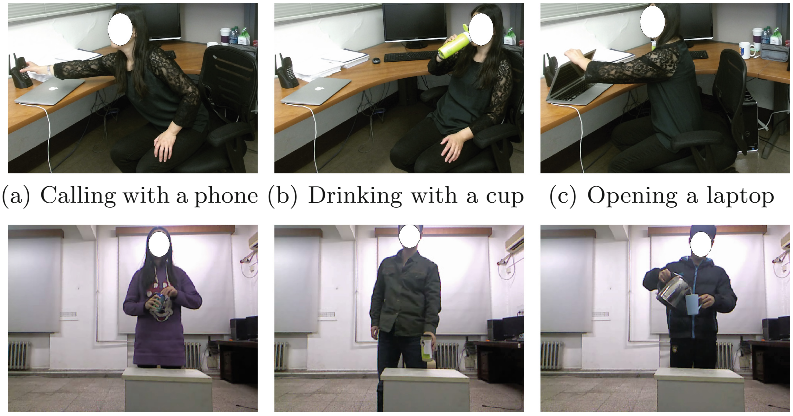
Our final optimization function is to solve the minimum energy in Eq. (1), defined by:

$$I^\star = \underset{I_1 I_2 \cdots I_t}{\arg\min} \sum_{t=1}^{T} E(I_t, V_t). \tag{6}$$

We optimize the Eq. (6) via greedy algorithm framework. We always choose the locally optimal choice at each stage with the hope of finding a global optimum.

4 Experiments

We collect a human-object interaction dataset by Kinect sensor with RGB, depth and skeletons. There are ten subjects, six males and four females. Each person performs each interaction four or five times. In our dataset, there are eight different daily interactions including *calling with a phone, drinking with a cup, picking up a mouse, putting down a cup, opening a laptop, turning off a laptop, opening a soda can, and pouring into a cup.* Some examples from our dataset are shown in Fig. 5.



(a) Calling with a phone (b) Drinking with a cup (c) Opening a laptop

(d) Opening a soda can (e) Putting down a cup (f) Pouring into a cup

Fig. 5. Some examples of the human-object interactions in our dataset

Action Recognition. The human actions are ambiguous without manipulated objects, for example, making a call has similar skeleton sequences with drinking. We use global features and local features from skeleton data described in Sect. 3.1. The confusion matrix of the action recognition without and with the manipulated objects context are shown in Fig. 7(a) and (b). The results in Fig. 7(a) demonstrate that making a call, drinking, putting down and pouring water have a large confusion probability. In Fig. 6, the *"drink"* action energy is minimum, but considering the manipulated objects, it has a large chance to be the phone, we finally infer the action label is *"make a call"*. The results of our method show that the action recognition accuracy can be improved with the related objects as context.

Object Classification. In human-object interactions, related objects are always small or occluded by human hand. Occlusion is one of the most difficult issues in the field of computer vision. Bo et al. [1] designed a family of kernel features that described gradient, color and shape. Our dataset is randomly split into ten parts, and we adopt leave-one-sample-out cross validation strategy, namely, one original video sequence is used as the test data while the rest original video

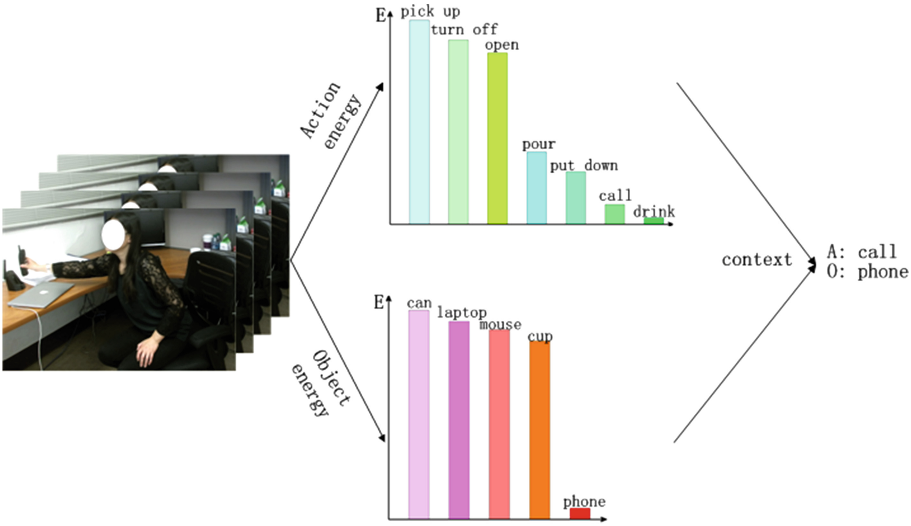


Fig. 6. Optimize action recognition and object recognition by the context between them.

sequences as the training data. The confusion matrix of the object recognition without and with the human actions information are shown in Fig. 8(a) and (b). We can see that the context of the human actions and manipulated objects is effective in improving the accuracy of object classification.

Human-Object Interaction Recognition. In addition to the results of human actions and objects from the RGBD video, we can also recognize human-object interaction. Figure 9 indicates the performance of the human-object interaction recognition of our method. We can see that our results demonstrate the effectiveness of human action recognition and object classification for better human-object interaction recognition.

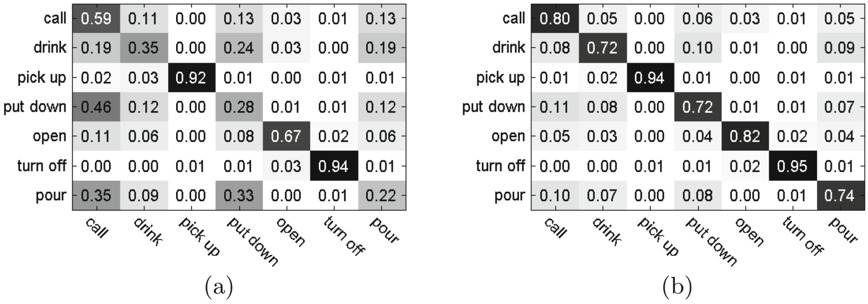
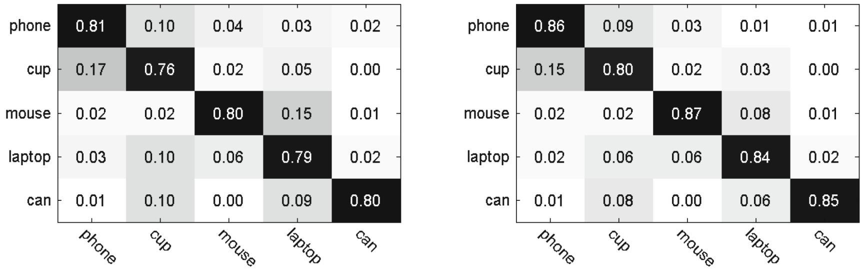


Fig. 7. (a) The confusion matrix of action recognition in [21]. (b) The confusion matrix of our method



(a) (b)

Fig. 8. (a) The confusion matrix of object recognition in [1]. (b) The confusion matrix of our method

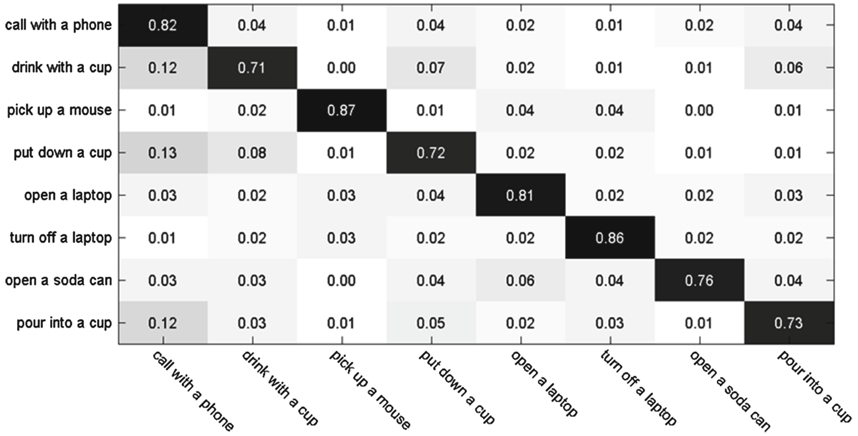


Fig. 9. The confusion matrix of human-object interaction recognition

5 Conclusions

Human-object interaction recognition is one of the most important topics in the field of computer vision. In this paper, we model human actions and manipulated objects in a unified framework for recognition. For human actions, we use local features together with global features to improve the accuracy of recognition. For object recognition, we apply kernel features. Experiments show that the features are more discriminative. Then we model the context between actions and manipulated objects. It is helpful to improve human-object interaction recognition. In the future, we will extend our dataset and consider more kinds of actions and objects.

References

1. Bo, L., Ren, X., Fox, D.: Kernel descriptors for visual recognition. In: Advances in Neural Information Processing Systems, pp. 244–252 (2010)

2. Bo, L., Ren, X., Fox, D.: Depth kernel descriptors for object recognition. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 821–826. IEEE (2011)
3. Bo, L., Sminchisescu, C.: Efficient match kernel between sets of features for visual recognition. In: Advances in Neural Information Processing Systems, pp. 135–143 (2009)
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
5. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
6. Chen, D.Y., Shih, S.W., Liao, H.Y.: Human action recognition using 2-d spatio-temporal templates. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 667–670. IEEE (2007)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
8. Flitton, G.T., Breckon, T.P., Bouallagu, N.M.: Object recognition using 3d sift in complex CT volumes. In: BMVC, pp. 1–12 (2010)
9. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: 2005 Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1458–1465. IEEE (2005)
10. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: using spatial and functional compatibility for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1775–1789 (2009)
11. Ito, S., Kubota, S.: Object classification using heterogeneous co-occurrence features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 701–714. Springer, Heidelberg (2010)
12. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 951–958. IEEE (2009)
13. Liang, Y., Lu, W., Liang, W., Wang, Y.: Action recognition using local joints structure and histograms of 3d joints. In: 2014 Tenth International Conference on Computational Intelligence and Security (CIS), pp. 185–188. IEEE (2014)
14. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 444–451. IEEE (2009)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
16. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2929–2936. IEEE (2009)
17. Raptis, M., Soatto, S.: Tracklet descriptors for action modeling and video analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 577–590. Springer, Heidelberg (2010)
18. Su, Y., Allan, M., Jurie, F.: Improving object classification using semantic attributes. In: BMVC, pp. 1–10 (2010)
19. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2004–2011. IEEE (2009)

20. Wei, P., Zhao, Y., Zheng, N., Zhu, S.C.: Modeling 4d human-object interactions for event and object recognition. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3272–3279. IEEE (2013)
21. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27. IEEE (2012)
22. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17–24. IEEE (2010)
23. Yao, B., Fei-Fei, L.: Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1691–1703 (2012)
24. Yu, K., Xu, W., Gong, Y.: Deep learning with kernel regularization for visual recognition. In: Advances in Neural Information Processing Systems, pp. 1889–1896 (2009)
25. Zhang, J., Huang, K., Yu, Y., Tan, T.: Boosted local structured HOG-LBP for object localization. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1393–1400. IEEE (2011)