A Novel Group-sparsity-optimization-based Feature Selection Model for Complex Interaction Recognition

Luyu Yang[1], Chenqiang Gao[1], Deyu Meng[2], Lu Jiang[3]

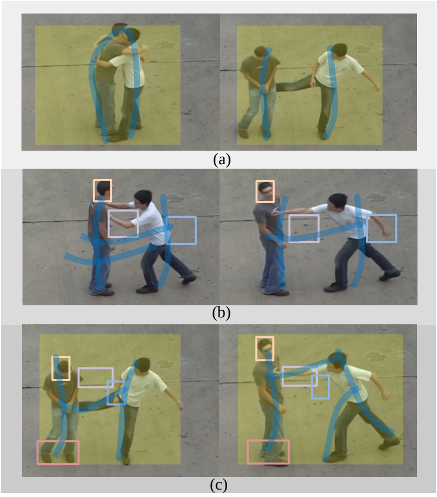
[1]Chongqing Key laboratory of Signal and Information Processing, Chongqing University of Posts and Telecommunications
[2]School of Mathematics and Statistics, Xi'an Jiaotong University
[3]School of Computer Science, Carnegie Mellon University

Abstract. Interaction recognition is an important part of action recognition and has various applications such as surveillance systems, human computer interface, and machine intelligence. In this paper, we propose a novel group-sparsity-optimization-based feature selection model for complex interaction recognition. Firstly multiple local and global features are concatenated into a feature pool, and then based on the group sparsity optimization, different feature types are automatically selected to fit specific interaction categorization. We test our method on the benchmark dataset: the UT-interaction dataset. Experimental results substantiate the effectiveness of the proposed method on complex interaction recognition tasks as compared with current state-of-the-art methods.

1 Introduction

Action recognition aims to recognize the ongoing action from an unknown video. This technique has a variety of potential applications, such as intelligent surveillance systems, human computer interface, machine intelligence et al. In the past decades, the research focus was mainly on the task of single-person action recognition [1, 2] and good performance was achieved. In some typical datasets [3, 4], the recognition accuracy has reached over 90% [5, 6]. Good progress for single-person action recognition makes many researchers devote efforts to the interaction recognition which is a more complex recognition task. Besides the challenges of background clutter, partial occlusion and the perspective effect, compared with single-person action recognition, the interactive action recognition task additionally suffers from: (1) variations within an interaction among different performers; (2) similar patterns among different interactions or with background interference.

Various methods [7–9] have been proposed to address this challenge in recent years. Among them, methods under the framework of machine learning have received more attention. Most of such methods try to recognize various interactions using only one feature or one concatenation of features. Although good performance can be achieved among classes with more prominent discrimination, classes with obscure discrimination were always recognized poorly. It is a


(a)

(b)

(c)

Fig. 1. Three image pairs demonstrate one-vs-one interaction classification of (a) “hug” and “kick”; (b) “punch” and “push”; (c) “kick” and “push”. The bigger shadowed-areas and smaller bounding boxes respectively indicate global and local regions containing discriminative information for classification. Each pair of bounding boxes of comparison is colored differently. The strokes around human torsos indicate the interaction pose of two performers.

challenging task to improve accuracy on interactions with obscure discrimination while still keep high accuracy on interactions with prominent discrimination. In [10], the recognition accuracy of interactions “hug” and “kick” were above 95%, but only around 70% for “punch” and “push”. One reason is that the former two interactions are different from other interactions in a more prominent way, while the latter two interactions have more obscure discriminative information. As shown in Fig. 1, between “hug” and “kick”, the region which provides discriminative information is prominent and large, so a global feature such as the popular dense trajectory would be sufficient. However, between “punch” and “push”, discriminative information is obscure and exists only in small local areas. In this case, local features around those areas have to be rationally utilized to provide effective description. Differently, in “kick” and “push”, both global and local areas can provide some discriminative information, but not typical enough when being considered separately. Therefore, it is more reasonable to combinationally consider multiple feature types here. Faced with various classes of interactions, the questions are, **which** types of features should be chosen and **how** can this choice be made automatically.

To answer these questions, in this paper we propose a novel feature selection model for interaction recognition. The proposed model automatically learns to select feature types which optimize the recognition. We choose multiple local and global features and concatenate them into a feature pool, and our model selectively learns the best feature types from the pool. To the best of our knowledge, it is new to utilize the group sparsity for human interaction recognition. We test our method on the interaction benchmark, the UT-interaction dataset, and experimental results demonstrate the effectiveness of the proposed method.

The rest of paper is organized as follows: Section 2 reviewed related methods of interaction recognition. The features we utilized are introduced in Section 3. In Section 4, how our model selects feature types and why it has feature selection capability are explained. Finally, Section 5 exhibits experimental results.

2 Related work

As compared with the action recognition problem, which has been investigated for more than a decade, the interaction recognition has not attracted much attention until the first attempt by Oliver et. al [11]. This method handled the interaction recognition problem by employing motion trajectories obtained from blob-tracking of human. Another remarkable milestone is the successful use of a new spatio-temporal feature detector [12] in action recognition which received much attention in the field [13–16]. Its invariability under illumination change and noisy background has largely benefited the action recognition task under real-world scenes. Recognition based on key frames [17, 18] is also a widely-used method in interaction recognition. These methods analyzed descriptors extracted from key frames of a video, trying to model the relations between interactions and poses of key frames, while somewhat underused the contextual information of motion trajectory. To utilize contextual information within an interaction, a respectable amount of works have been presented to model the context of interactions [19, 20]. By presenting interactions by action context descriptors, in [20], action context was encoded by interactive phrases which were composed of atomic actions of elementary movements, namely attributes. Their method obtained improvement compared with previous methods. However, in this method, the attributes need to be manually labeled and specified to certain data sets, which makes the method less automatic in recognition and less scalability onto other data sets. In many works, fusion of multiple features was used to balance the contribution of each feature type. In [21], training scores of each feature type were employed as an input to learn the fusion weight vector. However, these methods tend to lose the discriminative description of combined features at the early-fusion stage. Our method capitalizes on selecting feature types using the group sparsity technique and feature types are selected before training. To realize group sparsity, a weight vector indicating the importance of all sample features is involved. The $L_{2,1}$-norm is imposed on these weights to enforce its group sparsity on simultaneously selecting certain scales of features. Results show that our feature selection model makes feature-fusion more effective. Although there exist some feature selection models [22][23] , it is new to use group sparsity to achieve the feature selection goal in interaction recognition.

3 Interaction representation

We utilize five scales of features as a feature pool, including three features for local context and two for global context. The local features are extracted based on images. We use the pedestrian detector [24] to detect each interacting person and his body parts in one image and thus 1 full-body bounding box and a group of 8 body-part bounding boxes for one person can be obtained. The first and second local features are the HOG of full-body bounding box and body-part bounding boxes, respectively. The third local feature is the configuration of body-part bounding boxes. The global features are extracted along the complete video based on dense trajectory. It should be noted that our method represents a general implementation scheme, and any other local or global features can be readily integrated to further extend its capability.

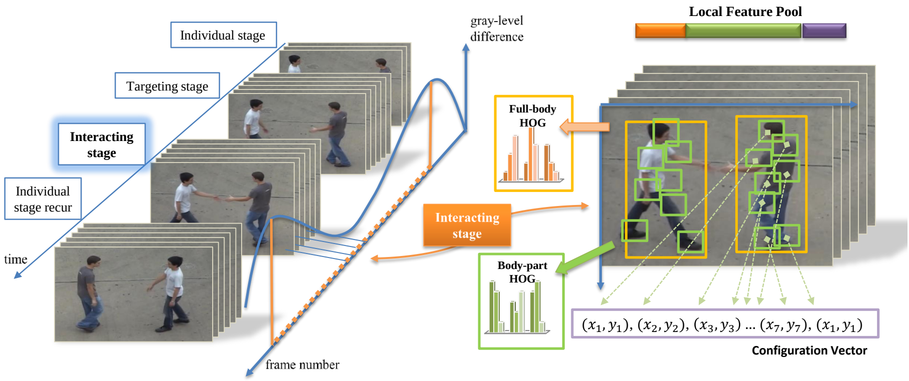


Fig. 2. The left demonstrates Interacting Stage Detection process: Frame-sequences of the four stages are represented and the right-side curve corresponds to the changes of gray-level differences during the stages. The right shows how local features: full-body HOG, body-part HOG and body-part configuration vector are extracted from the interacting stage. On top of the right is the concatenated local features.

3.1 Local feature representation

Local feature representations are important in interaction recognition, since in some interactions the discriminative information is obscure and exists only in small local regions with few frames. In order to properly recognize these interactions, we need to make full use of discriminative information of local regions. In our work, we utilized three types of features as the representation of local regions.

Instead of using the whole video for local feature extraction, the core part of video is used in this paper. According to our observation, most interactions can be regarded as a four-stage sequential transition including: individual stage, targeting stage, interacting stage, and individual stage recurrence. Only the interacting stage, which provides more prominent in this state is to our interest. This is mainly due to two reasons. Firstly, when a video is regarded as frames, the other three stages, including individual stage, targeting stage and individual stage recurrence, might appear very similar among different interactions, which inclines to reduce the distinguishability among interaction classes. Secondly, the other three stages are of much randomness, and may not be contributory to the classification since how performers would like to act before or after they interact is much up to their willingness.

Interacting stage detection Fig. 2 demonstrates the process of the interacting stage detection. In order to automatically detect the third interacting stage, for each video we compute the gray-value difference between each two consecutive frames and the difference value of each pixel is added up, thus obtaining a gray-value difference curve. According to our experience, the curve is saddle-shaped with two peaks which respectively indicates the starting and ending frames of the interacting stage. To get the starting and ending frame numbers of the interacting stage, we employ an n-degree curve fitting

$$y = ax^n + bx^{(n-1)} + cx^{(n-2)} + ... + dx + e \tag{1}$$

to smooth the curve with an initial $n = 10$. If more than 2 local maxima is found in the fitted curve, we continue the curve fitting with increased n till only 2 local maxima are left. The frame numbers which correspond to the maxima are used as the starting and ending frames of the interacting stage. Frames between the two frames are those which we later extract local feature from.

Local HOG We try to detect the local regions where discriminative information is more likely to exist. So features are extracted only in regions where full-body and body-part bounding boxes are detected.

Histogram of gradients (HOG) [25] is a powerful description of texture in action recognition [24, 26, 27], so we use it as a feature of the detected local regions. We resize each bounding box to 64×128 using nearest neighbor interpolation, and then an 8×8 grid is superimposed upon each full-body bounding box and body-part bounding box corresponding to each interacting person. Finally, a full-body local descriptor with a size of $S_f = 8 \times 16 \times 31$ and a body-part local descriptor with a size of $S_p = 64 \times 16 \times 31$ are obtained for each interacting person.

Body-part configuration Besides texture description using HOG, spatial information of tracked body-part is another type of description for local regions. Fig. 2 demonstrates the extraction process of this local feature. According to our observations, the configuration of body-part bounding boxes is discriminative among different interactions. In some obscurely discriminative interactions, such spatial information of body parts can be important clues when texture of local regions appears similar. We employ the relative location of detected body-part as a representation of configuration. Concatenation of coordinate centers (x_b, y_b) of 8 body-part bounding boxes, where $b = (1, 2, ..., 8)$ is used in our work and thus a configuration vector of length 2×8 is obtained.

In order to cover the discriminative regions as much as possible, three types of local features are used, including full-body HOG, body-part HOG and body-part configuration covered both texture and spatial description. Moreover, the texture description is of both larger and smaller local regions which is more comprehensive compared with [18, 21], in which texture descriptions were only refined to the full-body scale.

3.2 Global feature representation

In videos, motion is a most informative cue for action recognition, and the motion trajectory is one of ways to describe motion. The dense trajectory extraction method described in [28] is popular in action recognition in recent years. We employ this method in interaction recognition to obtain a good representation of interaction trajectory. In our work, interactive motion is tracked, which forms a trajectory of interest points, and descriptors are extracted along the trajectory. The more detailed process includes three steps. Firstly, densely sampled points at multiple scales are tracked using the optical flow method used in [29]. Secondly, we track the sampled points to form trajectories. Finally, descriptors are computed by space-time volume around the trajectory. We utilize state-of-art descriptors including HOGHOF which shows prominent performance on various datasets [15, 30], and the motion boundary histogram (MBH) [31] which can capture the relative motion between pixels both vertically and horizontally. The two descriptors are computed in the same parameter setup as in [28]. As a result, two types of global features based on trajectory are obtained with the size of 204 for HOGHOF and 192 for MBH.

All together we used 5 types of features of both local and global representation. Each type of feature has different representation capability, and by utilizing them, we try to capture discriminative information which might exist in texture, brightness and spatial locations. Before we concatenate them into a feature pool, a powerful fisher vector tool [32] is employed to encode each type of feature. Therefore, the final feature pool consists of 5 types of encoded features.

4 Intrinsic feature selection model

With multiple types of features obtained, the easiest way is to concatenate all of them into one feature and directly use it as an input of a machine learning model. However, as we analyze earlier in section 2, the recognition complexity is not always on the same level among different interactions. For interactions with prominent discriminative information such as “hug” and “kick”, features at global scale would be sufficient, but such features are not sufficient for those with obscure discriminative information such as "punch" and "push" which share similar patterns that can even be confused by human eyes. So in order to achieve good performance among interactions with obscurely discriminative information, we have to utilize local features to make up for or even replace the insufficient description of global features. However in practice, interactions are not just divided into two poles of complexity. There exist different mixtures of discriminative information which correspond to the concatenation of different feature types, and it is difficult to decide which feature types should be selected. To address this problem, we formulate the feature selection into a learning process, in which feature types are selected according to how well they perform. Aiming at effective feature selection, our method uses the group sparsity technique [33] and feature types are selected to optimize the recognition performance. The samples as well as features can also be simultaneously selected during training based on the intrinsic mechanism of SVM. More details are presented as follows.

4.1 Formulation of our model

Given an input video, we extract K types of features from it and formulate a concatenation of K types of features with total dimension $d = d_1 + d_2 + ... + d_k$. In order to enable the model with feature selection capability, we define a weight vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_d)^T$, where the elements of $\boldsymbol{\alpha}$ can be grouped into $(\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, ..., \boldsymbol{\alpha_k})$ according to the lengths of K feature types. Hence the weighted feature group is presented as $\boldsymbol{\alpha} \odot \boldsymbol{x}$. Given the i^{th} sample $\boldsymbol{x_i}$ with label y_i, the interaction recognition problem can be formulated as an optimization problem:

$$min_{\boldsymbol{\omega},b,\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_i (1 - y_i(b + \boldsymbol{\omega}^T(\boldsymbol{\alpha} \odot \boldsymbol{x_i})))_+^2 \qquad s.t. \qquad \|\boldsymbol{\alpha}\|_{2,1} \leq s, \quad (2)$$

where $\boldsymbol{\omega}$ is the model parameter, b is the offset value, C is the cost coefficient and s is the constraint of $\|\boldsymbol{\alpha}\|_{2,1}$. In Eq. 2, $L_{2,1}$-norm of $\boldsymbol{\alpha}$ can be written as

$$\|\boldsymbol{\alpha}\|_{2,1} = \sum_{k=1}^{K} \|\boldsymbol{\alpha}_k\|_2.$$

By optimizing Eq. 2, we can simultaneously make the calculated $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$ sparse. $\boldsymbol{\alpha}$ is sparse among K feature groups while dense within each type of features, which indicates that this model can have both sample selection and feature selection capability.

4.2 Learning and inference

Inference: Given the model parameters $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$ and b, the inference problem is to find the right interaction class label y for a test video $\boldsymbol{x}$. We define the following function to score $\boldsymbol{x}$.

$$y = b + \boldsymbol{\omega}^T(\boldsymbol{\alpha} \odot \boldsymbol{x}). \quad (3)$$

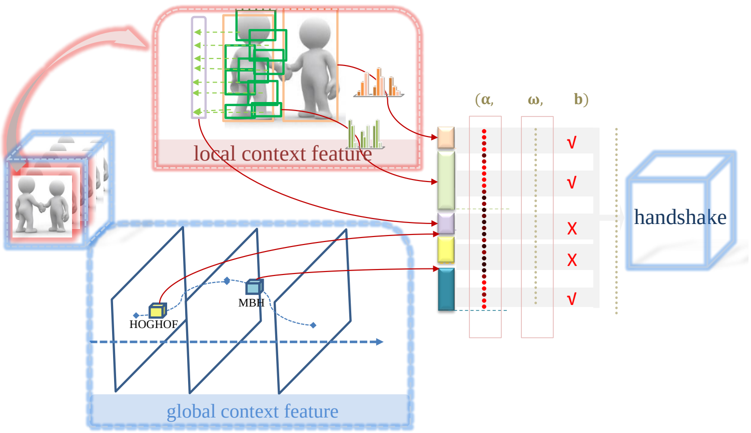


Fig. 3. The testing procedure: two global features (shown at the bottom) and three local features (shown on the top) are extracted from the test sample, and a concatenation of five features is shown as the rectangular patches in five colors in the middle. A set of trained $\boldsymbol{\alpha}$ which corresponds to feature selection, $\boldsymbol{\omega}$ and b, correspond to a test score calculated with the score function. The class which has the highest score is the test result of this video. The recognition result is "handshake" as shown, which rightly matches the interaction class.

The complete inference procedure is demonstrated in Fig. 3. Five features are extracted from the test video to form a feature pool. We employ the one-vs-one classification method in the learning phase, so between each two interaction classes there is a set of trained $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$ and b, corresponding to a test score calculated with Eq. 3. The class which has the highest score is the test result of this video.

Learning: Given N training samples $(x_n, y_n)(n = 1, 2, ..., N)$, the training task is to learn the model parameters $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$ and b. Our optimization algorithm includes mainly two steps to iteratively learn these three parameters.

(1) Holding $\boldsymbol{\alpha}$ fixed, the optimization problem is:

$$min_{\boldsymbol{\omega},b}\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_i(1 - y_i(b + boldsymbolomega^T(\boldsymbol{\alpha}\odot\boldsymbol{x_i})))_+^2, \tag{4}$$

which can be written as

$$min_{\boldsymbol{\omega},b}\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_i(1 - y_i(b + \boldsymbol{\omega}^T Q^T\boldsymbol{x_i}))_+^2,$$

where Q =

$$\begin{pmatrix}\alpha_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \alpha_d\end{pmatrix}$$

is a diagonal matrix. The optimization problem above is a standard SVM [34] model and can be directly solved by virtue of off-the-shelf tools, among which LIBSVM described in [35] is adopted.

(2) Holding $\boldsymbol{\omega}$, b fixed, the optimization problem is,

$$min_{\boldsymbol{\alpha}} \sum_i (1 - y_i(b + \boldsymbol{\omega}^T(\boldsymbol{\alpha} \odot \boldsymbol{x_i})))_+^2 \qquad s.t. \qquad \|\boldsymbol{\alpha}\|_{2,1} \leq s, \tag{5}$$

which can be written as:

$$min_{\boldsymbol{\alpha}} \sum_i (1 - y_i(b + \boldsymbol{\alpha}^T P^T \boldsymbol{x_i}))_+^2 \qquad s.t. \qquad \|\boldsymbol{\alpha}\|_{2,1} \leq s, \tag{6}$$

where P =

$$\begin{pmatrix} \omega_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \omega_d \end{pmatrix}$$

is a diagonal matrix and s is constraint parameter that controls sparsity level.

For being better solvable, the constrained optimization problem of the Eq. 6 is transformed into a unconstrained optimization problem with Lagrangian expression as follows:

$$L(\boldsymbol{x_i}, y_i, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_i (1 - y_i(b + \boldsymbol{\omega}^T(\boldsymbol{\alpha}^T P^T \boldsymbol{x_i})))_+^2 + \boldsymbol{\lambda}\|\boldsymbol{\alpha}\|_{2,1}, \tag{7}$$

Under certain $\boldsymbol{\lambda}$, this is a convex optimization model with respect to $\boldsymbol{\alpha}$, and can be readily solved by gradient descent method [36]. We can then derive $\boldsymbol{\alpha}$ based on the obtained result. The appropriate $\boldsymbol{\lambda}$ can be properly specified by cross-validation.

5 Experiments

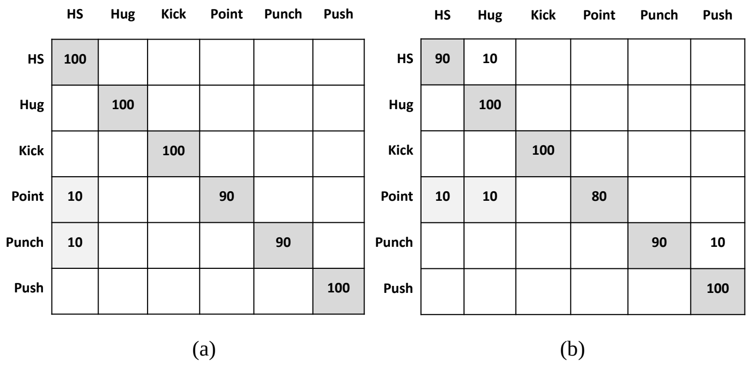
We test our method on the UT-Interaction dataset. This dataset consists of 20 videos in total, containing 6 classes of human-human interactions: “handshake”, “hug”, “kick”, “point”, “punch” and “push”. On average, there 8 instances of interactions per video and each video contains at least one instance. According to the filming condition, the dataset is divided into two sets. Set 1 is recorded at a parking plot with a stationary background, and Set 2 is recorded on a lawn with slight background movement and camera jitter. In accordance with experimental settings of the recognition task described in High-level Human Interaction Recognition Challenge [37], bounding boxes are used and the performance of our method is evaluated using leave-one-out cross validation on each set. The information of main actors (standing on the left or right side) is provided in the dataset as ground-truth. However, we did not use this information since it is hard to be obtained in realistic situations.

5.1 Implementation details

By using the interacting stage detection method mentioned in Section 3, we obtain frames of the interacting stage upon which we detect a full human body and 8 body parts of each interacting person using the deformable part-based model [24]. To ensure at least two interacting person detection results, we set the detection score threshold at a lower $threshold = -1.5$ compared with the default $threshold = -0.5$. And the Top two detection results of the rank list are chosen as the interacting person detection results.

We compute HOG of each full-body bounding box and body-part bounding boxes as described in [24]. Next we compute HOGHOF descriptor and MBH descriptor as described in [28]. Fisher vector is utilized to generate a codebook for each feature type.

When generating codebooks, the number of Gaussians G in Gaussian mixture model is an important parameter, so we evaluate a variety of G on UT-interaction dataset with HOGHOF, MBH and their combinations. We process parameter search, and it turned out the best performance is obtained when $G = 65$. Model parameter C and $\boldsymbol{\lambda}$ are optimized using cross validation.



	HS	Hug	Kick	Point	Punch	Push
HS	100					
Hug		100				
Kick			100			
Point	10			90		
Punch	10				90	
Push						100

(a)

	HS	Hug	Kick	Point	Punch	Push
HS	90	10				
Hug		100				
Kick			100			
Point	10	10		80		
Punch					90	10
Push						100

(b)

Fig. 4. (a) Confusion matrix of our method on Set 1 of UT-interaction dataset. (b) Confusion matrix of our method on Set 2. Note that "HS" stands for "Handshake".

5.2 Results

Fig. 4 shows the confusion matrix of the Set 1 and Set 2 in the UT-interaction dataset. It can be seen from Fig. 4 that the interactions "hug", "kick" and "push" are recognized better than other interactions, and achieve 100% recognition precision. In addition, the interactions "handshake" and "punch" which are usually regarded as interactions with obscure discriminative information with lower average precision [37, 10], also achieve precision above 90% for our method.

Relatively, “point” is the difficulty-recognized class, with a precision of 85%, since it is a special class in the dataset, with only one performer performing the action. Performance on Set 2 is not as good as on Set 1 in “handshake” and “point”, since Set 2 is filmed with camera jitter and partial occlusions of the background. Among the interactions with obscure discriminative information, “handshake” is slightly confused with “hug” in Set 2, and “punch” is misclassified as “push” in Set 2, which indicates that clutter increases the difficulty of recognizing interactions.

Table 1. Per-class precision (%) on UT-interaction dataset. 7 previous methods are compared with ours according to their average degree of precision. Average precision is listed in the last column.

Methods	Handshake	Hug	Kick	Point	Punch	Push	Average
Ryoo et al.[37]	75	87.5	75	62.5	50	75	70.8
Waltisberg et al.[10]	60	95	**100**	**100**	75	60	81.5
Yu et al.[14]	**100**	65	75	**100**	85	75	83.3
Ryoo et al.[26]	80	90	90	80	**90**	80	85
Patron-Perez et al.[21]	95	95	85	-	65	85	85
Kong et al.[20]	**100**	90	**100**	80	**90**	90	91.67
Vahda et al.[18]	85	**100**	95	95	80	95	92
Our Method	95	**100**	**100**	85	**90**	**100**	**95**

We also compare our classification accuracy in each interaction class with the methods proposed in [37, 10, 14, 26, 21, 20, 18] and the results are listed in Table 1. From Table 1 we can see that our method keeps high accuracy among interactions with prominent discrimination such as “hug” and “kick”, meanwhile improves accuracy among interactions with obscure discrimination such as “punch” and “push”. Best average precision is achieved using our method compared to the other 7 competing methods as well as best accuracy among four interactions out of six. Among interactions with obscure discriminative information, our method prominently outperforms state-of-art methods, especially in “punch” and “push” which demonstrates bad performance in most previous methods. However, “point” does not show strong performance in our method. This is since “point” is an exceptional class in the UT-interaction dataset, containing only one person performing the activity with no interaction information. Since our model is specifically designed for capturing interaction information between humans, it might not be so appropriate for this specific class. Yet our method still gets a reasonable result on this class (85%), comparable to most current methods along this line.

The average precision of all competing methods on each set are listed in Table 2. Our method achieves 96.7% and 93.3% precision on Set 1 and Set 2, respectively, which shows that our method outperforms the state-of-art methods on both sets.

Table 2. Average precision (%) on UT Interaction Dataset Set 1 and Set 2. Three state-of-art methods are compared.

Methods	Set1	Set2	Average
Waltisberg et al.[10]	83	80	81.5
Patron-perez et al.[21]	84	86	85
Vahdat et al.[18]	93	90	92
Our Method	**96.7**	**93.3**	**95**

5.3 Conclusion

In this paper, we have proposed a novel group-sparsity-optimization- based feature selection model for complex interaction recognition. We have used various combines of feature types with different representation capacity to recognize interactions with different prominent/obscure discrimination. Aiming at this goal, we have proposed a model which automatically selects feature types for specific interactions. We test our method on interaction benchmark UT-interaction dataset and extensive experimental results show the effectiveness of the proposed method for complex interaction recognition tasks compared to the state-of-the-art methods. Specifically, our method improves accuracy on interactions with obscure discrimination, while still keeps high accuracy on interactions with prominent discrimination.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 61102131, 61373114, 61275099), the Natural Science Foundation of Chongqing Science and Technology Commission (No. c-stc2014jcyjA40048), the Project of Key Laboratory of Signal and Information Processing of Chongqing (No. CSTC2009CA2003).

References

1. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action recognition in the premotor cortex. Brain **119** (1996) 593–609
2. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding **104** (2006) 249–257
3. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition, 2004 IEEE International Conference on. Volume 3., IEEE (2004) 32–36
4. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Computer Vision, 2005 IEEE International Conference on. Volume 2., IEEE (2005) 1395–1402
5. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: Computer Vision and Pattern Recognition, 2009 IEEE Conference on, IEEE (2009) 1948–1955

6. Hoai, M., Lan, Z.Z., De la Torre, F.: Joint segmentation and classification of human actions in video. In: Computer Vision and Pattern Recognition, 2011 IEEE Conference on, IEEE (2011) 3265–3272
7. Joo, S.W., Chellappa, R.: Attribute grammar-based event recognition and anomaly detection. In: Computer Vision and Pattern Recognition Workshop, 2006 Conference on, IEEE (2006) 107–107
8. Khan, S.M., Shah, M.: Detecting group activities using rigidity of formation. In: Proceedings of the 13th ACM international conference on Multimedia, ACM (2005) 403–406
9. Kitani, K.M., Sato, Y., Sugimoto, A.: Deleted interpolation using a hierarchical bayesian grammar network for recognizing human activity. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005 Joint IEEE International Workshop on, IEEE (2005) 239–246
10. Waltisberg, D., Yao, A., Gall, J., Van Gool, L.: Variations of a hough-voting action recognition system. In: Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos. Springer (2010) 306–312
11. Oliver, N., Rosario, B., Pentland, A.: Graphical models for recognizing human interactions. In: Proceedings of International Conference on Neural Information and Processing Systems, Citeseer (1998) 924–930
12. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. (2005) 65–72
13. Wu, X., Ngo, C.W., Li, J., Zhang, Y.: Localizing volumetric motion for action recognition in realistic videos. In: Proceedings of the 17th ACM international conference on Multimedia. (2009) 505–508
14. Yu, T.H., Kim, T.K., Cipolla, R.: Real-time action recognition by spatiotemporal semantic and structural forest. In: Proceedings of the British Machine Vision Conference, BMVA Press (2010) 52.1–52.12
15. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008 IEEE Conference on, IEEE (2008) 1–8
16. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Proceedings of the 12th European Conference on Computer Vision. Volume 4. Springer (2012) 215–230
17. Laptev, I., Pérez, P.: Retrieving actions in movies. In: Computer Vision. IEEE 11th International Conference on, IEEE (2007) 1–8
18. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: Computer Vision Workshops, 2011 IEEE International Conference on, IEEE (2011) 1729–1736
19. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. Pattern Analysis and Machine Intelligence, 2012 IEEE Transactions on **34** (2012) 1549–1562
20. Kong, Y., Jia, Y., Fu, Y.: Interactive phrases: Semantic descriptions for human interaction recognition. In: Pattern Analysis and Machine Intelligence, 2014 IEEE Transactions on. Volume 36. (2014) 1775–188
21. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in tv shows. Pattern Analysis and Machine Intelligence, 2012 IEEE Transactions on **34** (2012) 2441–2453

22. Tan, M., Wang, L., Tsang, I.W.: Learning sparse svm for feature selection on very high dimensional datasets. In: Proceedings of the 27th International Conference on Machine Learning. (2010) 1047–1054
23. Qian, Y., Zhou, J., Ye, M., Wang, Q.: Structured sparse model based feature selection and classification for hyperspectral imagery. In: Geoscience and Remote Sensing Symposium, 2011 IEEE International Conference on, IEEE (2011) 1771–1774
24. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, 2008 IEEE Conference on, IEEE (2008) 1–8
25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005 IEEE Conference on. Volume 1., IEEE (2005) 886–893
26. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: Computer Vision, 2011 IEEE International Conference on, IEEE (2011) 1036–1043
27. Dong, Z., Kong, Y., Liu, C., Li, H., Jia, Y.: Recognizing human interaction by multiple features. In: Pattern Recognition, 2011 First Asian Conference on, IEEE (2011) 77–81
28. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition, 2011 IEEE Conference on, IEEE (2011) 3169–3176
29. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis. Springer (2003) 363–370
30. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., et al.: Evaluation of local spatio-temporal features for action recognition. In: 2009 British Machine Vision Conference. (2009) 127
31. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Computer Vision, 2006 European Conference on. Volume 3952. Springer (2006) 428–441
32. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: Computer Vision, 2013 IEEE International Conference on, IEEE (2013) 1817–1824
33. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint 2, 1-norms minimization. In: Advances in Neural Information Processing Systems. (2010) 1813–1821
34. Xu, Z., Dai, M., Meng, D.: Fast and efficient strategies for model selection of gaussian support vector machine. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **39** (2009) 1292–1307
35. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27
36. Baird, L., Moore, A.W.: Gradient descent for general reinforcement learning. Advances in Neural Information Processing Systems (1999) 968–974
37. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 1593–1600