Human behaviour recognition with mid-level representations for crowd understanding and analysis

Bangyong Sun¹,² | Nianzeng Yuan¹ | Shuying Li⁴ | Siyuan Wu² | Nan Wang²,³

¹ College of Printing, Packaging Engineering and Digital Media, Xi’an University of Technology, Xi’an, Shaanxi 710048, China
² The Key Laboratory of Spectral Imaging Technology CAS, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, Shaanxi 710119, China
³ University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, P. R. China
⁴ School of Automation, Xi’an University of Posts & Telecommunications, Xi’an, Shaanxi 710121, P. R. China

Correspondence
Shuying Li, School of Automation, Xi’an University of Posts & Telecommunications, Xi’an, 710121, Shaanxi, P. R. China.
Email: angle_li@163.com
Siyuan Wu, The Key Laboratory of Spectral Imaging Technology CAS, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an 710119, China.
Email: wsuyuan@opt.ac.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 62076199; China Postdoctoral Science Foundation, Grant/Award Number: 2019M653784; National Key R&D Program of China, Grant/Award Number: 2017YFB0502900; CAS Light of West China Program, Grant/Award Number: XAB2017B15; Key Laboratory of Spectral Imaging Technology of Chinese Academy of Sciences, Grant/Award Number: LSIT201801D

Abstract
Crowd understanding and analysis have received increasing attention for couples of decades, and development of human behaviour recognition strongly supports the application of crowd understanding and analysis. Human behaviour recognition usually seeks to automatically analyse ongoing movements and actions in different camera views by using various machine learning methodologies in unknown video clips or image sequences. Compared to other data modalities such as documents and images, processing video data demands much higher computational and storage resources. The idea of using middle level semantic concepts to represent human actions from videos is explored and it is argued that these semantic attributes enable the construction of more descriptive methods for human action recognition. The mid-level attributes, initialized by a cluster processing, are built upon low level features and fully utilize the discrepancies in different action classes, which can capture the importance of each attribute for each action class. In this way, the representation is constructed to be semantically rich and capable of highly discriminative performance even paired with simple linear classifiers. The method is verified on three challenging datasets (KTH, UCF50 and HMDB51), and the experimental results demonstrate that our method achieves better results than the baseline methods on human action recognition.

1 | INTRODUCTION

Visual recognition and understanding [1–5] of human behaviours have been widely used in crowd analysis [6–8] and security surveillance. Moreover, the development of human action recognition can support the application of crowd understanding and analysis. In simple terms, the objective of human behaviour recognitions is to correctly classify the video into its action category, and the video is fragmented to contain only one implementation of human movement. In more general cases, the purpose of human action recognition is focused on performing the continuous recognition on every appeared human action of the input video from start to end. Human action recognition task is of significant meaning in several applications. For instance, a train station surveillance system can automatically recognize suspicious activities like “people suddenly run panicky ” or "a person waving his/her harms with swords in hands.” Recognition of human actions is also helpful in the real-time monitoring of patients, children and elderly persons. What is more important, a high performance action
recognition system can make the construction of vision-based intelligent environments and gesture-based human computer interfaces become possible.

Though a lot of enthusiasm has been generated in the vision community by recent advances in machine recognition of human actions [9–12], there are still several important issues left to be addressed. One of the most crucial issues is to find algorithms that can robustly overcome the variability of features with the same action class label. This is because there are large variations in performance. For example, hands waving motions can differ in speed and radian length. In addition, there are anthropometric differences between individuals. Moreover, similar observations can be made for different actions, especially for non-period actions or actions that are adapted to the environment (e.g. walking with a dog or handing somebody something). A thorough human action recognition framework should be able to tolerate variations within one class and distinguish actions of different classes. For increasing numbers of action classes, this will even be more challenging as the overlap between classes will be higher.

Many previous action recognition frameworks directly match low-level features with action class labels [13–16], in which abundant visual spatiotemporal information can hardly be generalized by the raw low-level features. To overcome the aforementioned drawback, recent works show that attributes built upon the raw low-level features can act as higher semantic concepts and bridge the gap between low-level features and action class labels. Some works [17–21] even treated attributes as middle-level features for classification or learning problems. Lampert et al. [22] proposed the direct attribute prediction model which was learned to predict the presence of each attribute, and then used these predictions to train object models. Liu et al. [23] treated the action attributes as latent variables, whose classifiers are pre-trained by linear SVMs with outputs being the inputs of latent SVM. The aforementioned methods have proved that the attributes are functional in many ways and effective for recognizing actions.

In this work, we simultaneously consider the relationship between low-level features and attributes to bridge the gap between low-level features and action class labels. Attributes can put up a semantically expressive bridge between raw data and higher level representations as they focus on depicting characters of all instances within the same class rather than naming one directly. But the intra-class variability can still make the attributes to be inaccurate for specific members of the training classes. As a result, some action samples may be associated with slightly different sets of attributes even if they share the same action class label. To address this problem, we construct a mid-level representation by combining low level features and attributes together. Figure 1 shows the procedure to generate the mid-level representation of our method. Each action attribute corresponds to a set of low-level features which share the same action class label. Meanwhile, every mid-level representation contains both intra-class and inter-class information, which make it very discriminative. The contributions of this paper would be summarized as follows:

1. The mid-level representations can conquer the appearance variation across cameras to some extent and represent the discriminative information of motion patterns in the video sequences
2. The proposed model is discriminative by making the same actions more similar after the middle-level representation while different motion patterns are more distinct. In this case, it can reduce the negative impact of the similar actions
3. Extensive experiments have been conducted on three action datasets to verify the performance of the proposed method. The experimental results show that our mid-level representation of human action patterns is capable of constructing a powerful action recognition framework. In addition, our method can give prominent performance on several standard datasets (KTH, UCF50, HMDB51).

The rest of this paper is organized as follows. In Section II, we review recent works on action recognition. In Section III, we introduce the proposed method in detail, including the construction of mid-level representation and the whole algorithm framework for classifying human actions. Then we present experimental results on several data bases for video based action recognition in Section IV. Finally, we conclude this paper in Section V.

2 RELATED WORKS

The recognition of human activities from video sequences is one of the most encouraging applications of computer vision which effectively support the crowd understanding and analysis. In recent years, this task has attracted the attention of researchers from academia, security agencies, industry and the general populace as well. According to the features used for recognizing human actions, the methods in related works can be mainly classified into two categories: global feature-based methods and local feature-based methods. And method based on neural network also obtain attention because of their outstanding performance in many computer vision tasks.
Moreover, methods of crowd understanding and analysis are essential to be stated.

2.1 | Global feature based method

Global feature based methods [24–26] encode the visual observation as a whole and the region of interest (ROI) is usually detected through background subtraction or tracking. Common global features are obtained from edges, silhouettes or optical flow [27]. The representations are powerful since they encode much of the information. The work of Bobick and Davis [27,28] is one of the earliest methods utilizing silhouettes. They extracted silhouettes from a single view and accumulated the differences between subsequent frames of a motion clip with the generated binary motion energy image (MEI) and motion history image. Instead of silhouette, the observation within the ROI can also be described with motion information—optical flow [28]. Flow methods are usually executed when background subtraction cannot be performed. Efros et al. [28] calculated optical flow in person-centered video frames on sports footage, where persons in the frame were very small. There are also a lot of work combining flow and shape descriptors together, which can overcome the limitations of single representation. Optical flow was employed in ref. [29], and it researched what makes a flow algorithm good for action recognition. In addition, the spatio-temporal volume approach is also a global representation, despite the fact that this approach shares many similarities with local approaches.

2.2 | Local feature based method

Local representations describe the object as a collection of individual patches. To estimate the local features, spatiotemporal interest points should be detected first. After that, local cubes will be calculated around these points. Then, the cubes are combined into a final feature. Finally, bag-of-feature approaches are always followed. Local features are less sensitive to noise and partial occlusion, and do not strictly require background subtraction or tracking. However, as they depend on the extraction of a sufficient amount of relevant interest points, pre-processing is sometimes indispensable.

To detect the interest points or salient regions, a large variety of methods [30–34] have been put forward. Laptev et al. [28] modified the 2D spatial interest points to detect 3D spatiotemporal interest points for action recognition. The detected local features preserve some rotation invariance, which is robust to the succeeding action recognition. Wang et al. [34] presented a detector to obtain spatio-temporal interest points by employing a Geometric Algebra for action recognition in multi-channel videos. Yi et al. [32] developed the salient point detector by taking advantage of trajectory space information based multiple kernel learning.

After the detection of interesting points or salient regions, many methods have been proposed to describe them. Scovanner et al. [35] extended the SIFT descriptor to 3D-SIFT, by using the spatiotemporal information which is denoted by a sub-histogram. Laptev et al. [36] applied histograms of gradient orientations (HoG) and histograms of optic flows to describe the information of local motion and cubes of the space-time neighbourhoods of detected interest points, and Klaser et al. [37] ameliorated HoG to the 3D case.

2.3 | Deep learning based method

C2-LSTM (Correlational Convolutional) [38] can effectively perceive the spatial features, temporal dependencies and motion structure. Zufan et al. [39] address the human action recognition by applying Conv-LSTM and FC-LSTM (full connection) which are combined with different attention mechanisms. AGC-LSTM [40] (attention enhanced graph convolutional) can explore the relationship and capture discriminative features in time-space domain. Meng et al. [41] propose a novel sample fusion network to achieve data augmentation for skeleton-based human action recognition (SHAR). Graph convolutions technology are considered to address SHAR, G3D [42] applies cross-time-space edges as skip connections which can propagate direct information for disentangling the importance of nodes.

2.4 | Crowd understanding and analysis

Besides human behaviour recognition, crowd understanding and analysis have many subproblems such as crowd counting and person re-identification. Wang et al. [6] proposed a large-scale benchmark for crowd localization and counting named NWPU-Crowd. Synthetic data [7] are also adopted for pixel-wise crowd understanding. In ref. [8], pixel shuffle decoder and density-aware curriculum learning are designed to obtain the potential of crowd counting models. Moreover, Zhao et al. [43] present PISNet to address the interference of other pedestrians for person re-identification. And Li et al. [44] focus on different and similar components for person re-identification.

3 | PROPOSED METHOD

In this section, details about the proposed method are introduced. Framework of our method is shown in Figure 2. In III-A, we illustrate the ‘low-level’ features used in this work. In III-B, we present our method for human action recognition in detail, which explicitly employs the attribute-based mid-level representation. In III-C, we give the optimization procedures for the recognition method.

3.1 | Low-level features

To extract low-level features, we follow the procedure described in ref. [45] and use the features named ‘action bank.’ In this implementation, action bank is obtained using the action
spotting detector [46] following the previous work [45]. As described in [45], action bank is comprised of many individual action detectors sampled broadly in semantic space as well as viewpoint space and is constructed to be semantically rich. What's more, it is capable of highly discriminative performance. So it is capable of being the basis of a powerful action recognition method.

3.2 | Attribute-based mid-level representations

Let \( X \in \mathbb{R}^{d \times N} \) represent the raw low-level features extracted from action video clips. Each column of \( X \) corresponds to the feature representation of a human action and \( N \) is the number of training samples. The low-level action bank feature \( x_i \in X \) represents the \( i \)-th sample from training set. \( y_i \) is the corresponding action class label of \( x_i \). Most previous works just represent actions with raw extracted features and simply connect the feature vectors to action class labels. In this case, the raw features cannot convey all the information the recognition needs, so that there exists a gap between the low-level features and class labels. In this paper, we consider that human actions can be better described by attribute-based mid-level representations, since these mid-level representations can act as higher semantic concepts and bridge the gap between low-level features and action class labels.

Attributes are mainly used to describing an instance (e.g. rather than naming one. In one human action recognition context, for example, a ‘leg motion’ may be shared between ‘jogging’ action and ‘running’ action. This kind of descriptions of shared components or parts between similar or different classes can be called as attributes. And attribute learning has been popularized recently as an effective method for image and video understanding [22]. In this paper, we design an action attribute space \( A^{d \times r} \) as a semantic metric space in which each column denotes one kind of information of an action class. The attribute space is composed of \( r \) attributes

\[
C = \{c_i\}_{1}^{r},
\]

and each attribute is mined from a group of low-level features that share the same action class label by using clustering algorithm. The clustering centers are defined as attributes because they are more discriminative than low-level features and able to capture the inherent intra-class variability of each action class. Thus, they are capable of representing the complete information of action classes. Given a training set

\[
D = \{(x_i, y_i)\}_{i=1}^{N}, \quad y_i \in \{-1, +1\},
\]

where \(+1\) denotes the positive samples and \(-1\) denotes the negative ones, we try to build a classification model to recognize an unknown action \( x \). To this end, both attributes and low-level features are used to generate mid-level representations \( \Phi^C(X) \), which represent the action video clips with higher-level concepts. The mid-level representation \( \Phi^C(x) \) is defined as

\[
\Phi^C(x) = [s(x, c_1), s(x, c_2), \ldots, s(x, c_r)],
\]

where \( c_i \in A^{d \times r} \) is the attribute generated from \( X \) by K-means clustering, and \( r \) is the total number of cluster centers. \( s(x, c) \) is a similarity measure function, and we measure the similarity by Gaussian Hausdorff function, that is,

\[
s(x, c) = \exp\left(-\|x - \hat{c}\|^2 / \delta^2 \right),
\]

where \( \delta \) is set as the average distance between all pairs of instances. For the training set \( D \), we define

\[
\Phi^C(X) = [\Phi^C(x_1); \Phi^C(x_2); \ldots; \Phi^C(x_N)]^T,
\]

where \( N \) is the number of training samples. For each action class \( i \), there is a label matrix

\[
Y_i = [y_{i1}, y_{i2}, \ldots, y_{iN}]^T,
\]

and

\[
y_{ij} \in \{-1, +1\},
\]

where \(+1\) denotes the positive samples and \(-1\) denotes the negative ones. To recognize all different kinds of action classes, we learn a classifier for each action class. And the classifier is denoted as

\[
f_i \left( \Phi^C(X) \right) = w_i^T \Phi^C(X),
\]

where \( w_i \in \mathbb{R}^{1 \times 1} \) is the parameter vector, \( w_i^T \) denotes the transpose of \( w_i \), \( f_i \) is a linear model. And given the training set \( D \), the maximum likelihood estimate for the weight vector \( \hat{w} \), is that which minimizes the sum squared error:

\[
J_{\text{ll}}(w) = \sum_{i=1}^{N} (f_i - y_i)^2.
\]
Let $W$ denote the matrix of parameter vectors, $[w_1, w_2, ..., w_m]$. To find the best parameter matrix $W$, we optimize all loss functions and regularize them with $\ell_{2,1}$-norm:

$$\min_{W} \left( \| W^T \Phi^C (X) - Y \|_{2,1} \right),$$

where

$$W = [w_1, w_2, ..., w_m],$$

$$Y = [Y^1, Y^2, ..., Y^m]^T,$$

and $m$ is the total number of action classes. If $Z \in R^{D \times T}$, then $\| Z \|_{2,1}$ is the $\ell_{2,1}$-norm of matrix $Z$, that is,

$$\| Z \|_{2,1} = \sum_d \sqrt{\sum_t Z_{dt}^2}.$$

This norm first computes the $\ell_2$ norm of the loss function matrix in each dimension, then computes the $\ell_1$ norm which is well known for sparse solutions. The regularization achieves the minimum if and only if

$$W^T \Phi^C (X) = Y,$$

for all corresponding elements. Thus, the $\ell_{2,1}$-norm has an important property so that it can provide the score of the mid-level representation, and find the best parameter matrix $W$. However, there is still a latent defect in this formulation. The K-means clustering technique is a non-convex method, so the process to generate action attributes may result in a suboptimal performance. To address this problem, the proposed method ameliorates the formulation in the following subsection.

### 3.3 Optimization method

So far, we assume that the attributes $C$ are available. However, it may lead to a suboptimal performance because of leaving the learning model out of accounting this situation. One alternative way is to learn $W, C$ simultaneously. So our model is extended as follows:

$$\min_{W,C} \left( \| W^T \Phi^C (X) - Y \|_{2,1} + \beta \Theta (\{X\}^m, C) \right),$$

where $\beta$ is the parameter to measure the influence of the clustering process, and $\Theta (\{X\}^m, C)$ is the optimization function to solve the sub-optimization problem of clustering process. The rationale behind this model is that by minimizing the integrated objective function, we can find a set of perfect attributes $C$ and the classification model $W$. The model $W$ can predict the data correctly with a large margin and minimize the loss of mutual information caused by feature merging. In this paper, we exploit the positive instances to upgrade the attributes $C$ as

$$\Theta \left( \{X\}^m, C \right) = \sum_{i=1}^{p} \min_{j=1,...,k} \left\{ H \left( x_i, c_j \right) \right\},$$

where $\{X\}^m$ represents the positive instances for the $l$-th action class label, and

$$H \left( x, c \right) = \| x - d \|^2 / \beta,$$

is the square of average Euclidean distance.

Although the objective function of Equation (5) is non-convex for a given $W$ or $C$ only, it is convex with respect to both $W$ and $C$. Therefore, it is difficult to get a global optimal solution directly. In this case, it is necessary to solve the objective function of Equation (5) by iterating over one variable, while fixing the other variable. It is not guaranteed that the global solution will be obtained, while the local optimal solution is satisfactory likewise. To simplify this process, we adopt an efficient block coordinate descend algorithm. Specifically, we first optimize the objective function with respect to $W$ when attributes $C$ are fixed, then optimize it with respect to $C$ when $W$ are fixed. These two procedures are repeated until convergence. Algorithm 1 summarizes the pseudo-code of block coordinate descend algorithm.

When $W$ is fixed, we need to solve a optimization which is non-convex and non-differentiable. Let $g(C)$ denote the objective function of Equation (5) and $C^{(0)}$ denote the current solution. Here, we employ the sub-gradient method to find a refined solution $\tilde{C}$, that is, $g(\tilde{C}) \leq g(C)$. Let $\nabla C$ denote the subgradient of $C$, we update $C^{(t+1)}$ as

$$C^{(t+1)} = C^{(t)} - \alpha \nabla g \left( C^{(t)} \right), \forall t = 0, 1, ..., \nu,$$

where $\alpha$ is the step size and $\nu$ is the number of iterations. Because the subgradient method is not a descent method, it is common to keep tracking $C$ until we find the best candidate, that is, the one with the smallest function value,

$$\tilde{C} = \arg \min_{C^{(t)}, t=0,...,\nu} g \left( C^{(t)} \right).$$
SUN ET AL.

FIGURE 3 Diagram of the one-versus-rest vote evaluation method. Each action mid-level representation is treated as the input of all action class classifiers, then the outputs of the classifiers are used as inputs of the response evaluation function to find the maximum output, whose corresponding classifier action label is the mid-level representation’s action label.

FIGURE 4 Illustration of the ‘Precision/Recall’ evaluation method.

It is evident that we will have \( g(\bar{C}) \leq g(C) \). When \( g(\bar{C}) = g(C) \), the algorithm stops and outputs \( \bar{C} \) is as the final result.

4 EXPERIMENTS

Before learning the optimal process, an important step is feature extraction. In this paper, as mentioned above, we use the action bank features as the low-level features. After extracting low-level features, block coordinate descend algorithm is employed to train the model. To verify the effectiveness and robustness of the proposed method, we conduct experiments on three representative datasets, that is, KTH, UCF50 and HMDB51 using different experimental settings.

For KTH, UCF50, and HMDB51 datasets, we employ two methods to predict the label of the test samples in order to better evaluate the algorithm. The first method is called ‘one-versus-rest vote’ method. As can be seen from Figure 3, the mid-level representation \( \Phi^C(x) \) is treated as the input of all classifiers, action label of the representation is chosen to hold the maximum output of all classifiers by using the ‘response evaluation function.’

Another evaluation method is called ‘Precision/Recall’ method. Figure 4 shows the illustration of the ‘Precision/Recall’ method. In this method, for the response of every action classifier, we will calculate the recall \( (R) \) and precision \( (P) \), then the average precision \( (AP) \) is calculated as

\[
AP = \int_0^1 P(R) dR.
\]  

(20)

For the aforementioned evaluation methods, average classification accuracy is served as the final measurement of the classification performance in this paper. \( y_i \) denotes the true label of the unlabeled feature \( x_i \), \( \hat{y}_i \) denotes its predicted label, \( n \) denotes the number of the testing samples of the action class. The average classification accuracy of an action class is defined as

\[
\text{accuracy} = \frac{1}{n} \sum_{i=1}^{n} \delta (y_i, \hat{y}_i),
\]  

(21)

if

\[
y_i = \hat{y}_i \]

(22)

\[
\delta (y_i, \hat{y}_i) = 1
\]  

(23)

otherwise

\[
\delta (y_i, \hat{y}_i) = 0
\]  

(24)

The larger the accuracy is, the better the classification performance is.

4.1 Performance and discussion

1) Performance on different datasets: In this subsection, experimental results on KTH, UCF50 and HMDB51 datasets with different training size and action classes are displayed. All evaluations were done with 3-fold-cross-validation. To verify the performance of the proposed framework, we compare our method with other human action recognition methods.

For KTH datasets, Figure 5 lists the confusion matrix for six actions with varied performance when using ‘one-versus-rest vote’ evaluation method. The column of the matrix represents the instances to be classified, meanwhile each row represents the corresponding classification results. In the confusion table some ‘hand-waving’ actions are misclassified to ‘clapping’ since both classes share ‘alternate hands motion’. While ‘hand waving’ representation does not have strong enough information to differentiate itself from ‘clapping,’ but the leg-related actions are perfectly classified even when they have similar appearance exhibitions (jogging and running). Table 1 compares the average accuracies of our method with other excellent methods. We can see from Table 1 that our method achieves the highest recognition accuracy of 96.1%, and is over 9% better than SA baseline, which proves the effectiveness of our method commendably. When using ‘Precision/Recall’ evaluation method, the final experimental results are shown in Table 2.
To further verify the performance of the proposed method on the UCF50 dataset, one-versus-rest vote evaluation method is exploited in the experiments. As can be seen from Figure 6, the darker the colour of the block is, the higher rate the accuracy of the action is. The errors of this confusion matrix are randomly distributed across action labels with no apparent trends, and the average accuracy of the UCF50 dataset is 68.9%. Since there are many considerable variations in action performance, camera movement, human appearance, viewpoint, illumination and background in this database, the classification result is acceptable. Table 3 shows the comparison results of our method with other methods. In order to further demonstrate the effectiveness of the method on this database, we also use the “Precision/Recall” evaluation method and list the average accuracy rate of each action class in Table 4. It can be seen that the majority of action classes still remain high classification accuracy, while actions like ‘PizzaTossing’, ‘SoccerJuggling’, ‘WalkingWithDog’, ‘TennisSwing’ and ‘Basketball’ have awful results lower than 50%.

We also conduct the experiment on the HMDB51 dataset. Figure 7 shows the confusion table of our proposed method when using ‘one-versus-rest vote’ evaluation method, and the average accuracy of our action classes is 45.1%. As mentioned before, HMDB51 database is a large and most challenging video database for human motion recognition. And it can capture a substantial degree of the richness and the full complexity of

---

**TABLE 1**  Comparison on the KTH dataset

| Methods       | Accuracy (%) |
|---------------|--------------|
| Klaser et al. [37] | 84.5         |
| Savarese et al. [47] | 86.8         |
| Ryoo et al. [48] | 91.1         |
| Liu et al. [23]  | 91.6         |
| Laptev et al. [36] | 91.8         |
| Liu et al. [49]  | 93.8         |
| Le et al. [50]   | 93.9         |
| Gilbert et al. [51] | 94.5         |
| Our method      | 96.1         |

**TABLE 2**  Accuracy(%) for each action class on the KTH dataset

| Action class     | Accuracy (%) |
|------------------|--------------|
| Boxing           | 100.0        |
| Clapping         | 97.15        |
| Handswaving      | 96.61        |
| Jogging          | 100.0        |
| Running          | 100.0        |
| Walking          | 100.0        |
| Average          | 98.96        |
Table 3 shows the comparison of our method with other methods. It can be seen from Table 5 that the performance of our method is better than the baseline method whose classification accuracy is 23.2%. Furthermore, when using 'Precision/Recall' evaluation method, the experimental results are shown in Table 6. It can be seen that only actions like 'golf,' 'pullup,' 'pushup' and 'situp' get high classification accuracy, meanwhile, other actions only have very poor classification results.

2) Parameter selection: In this part, we conduct a series of experiments on two datasets, that is, UCF50 action dataset and HMDB51 action dataset, in order to ulteriorly verify the effect of mid-level representations, and to evaluate the impact of the number of attributes, training sizes and action classes. Moreover, we also make experiments to study the influences of different parameters.

### Table 3: Comparison on the UCF50 dataset

| Methods            | Accuracy (%) |
|--------------------|--------------|
| Gist [52]          | 38.8         |
| Scene context [53] | 47.6         |
| Laptev et al. [36] | 47.9         |
| Sadanand et al. [45]| 57.9         |
| Gist3D [54]        | 65.3         |
| Our method         | 68.9         |

### Table 4: Accuracy rate(%) for each action class of the 'Precision/Recall' evaluation method on the UCF50 dataset

| BaseballPitch 87.60 | GolfSwing 80.41 | Lunges 71.76 | PommelHorse 90.51 | Skijet 58.35 |
|---------------------|-----------------|-------------|-------------------|--------------|
| Basketball 39.85     | High jump 68.91 | Military parade 77.29 | Pull ups 98.17 | Soccer juggling 35.64 |
| Bench press 93.59    | Horse race 78.52| Mixing 85.56 | Punch 96.58      | Swing 49.72  |
| Biking 66.96         | Horse riding 70.14| Nunchucks 59.44 | Push ups 90.01 | TaiChi 69.69   |
| Billiards 99.76      | Hula hoop 69.09 | Pizza tossing 22.93 | Rock climbing indoor 67.01 | Tennis swing 42.01 |
| Breast stroke 85.06  | Javelin throw 79.12| Playing guitar 100.0 | Rope climbing 60.44 | Throw discus 82.67 |
| Clean and jerk 92.45 | Juggling balls 85.15| Playing piano 80.03 | Rowing 61.67 | Trampoline jumping 63.03 |
| Diving 82.74         | Jump rope 95.63 | Playing tabla 89.28 | Salsa spin 81.32 | Volleyball spiking 68.43 |
| Drumming 76.39       | Jumping jack 99.76| Playing violin 88.76 | Skate boarding 69.77 | Walking with dog 35.03 |
| Fencing 91.71        | Kayaking 57.64 | PoleVault 68.93 | Skiing 64.40 | YoYo 78.75 |

**Figure 7** Confusion table of our method on the HMDB51 database
TABLE 6  Accuracy rate(%) for each action class of the 'Precision/Recall' evaluation method on the HMDB51 dataset

| Action Class   | Accuracy Rate (%) |
|----------------|-------------------|
| Brush hair     | 68.14             |
| Eat            | 6.74              |
| Kiss           | 37.92             |
| Shake hands    | 58.86             |
| Sword          | 20.97             |
| Cartwheel      | 36.14             |
| Fall floor     | 21.36             |
| Laugh          | 47.71             |
| Shoot ball     | 60.78             |
| Sword exercise | 23.82             |
| Catch          | 81.14             |
| Fencing        | 61.50             |
| Pick           | 20.97             |
| Shoot BoW      | 80.55             |
| Talk           | 47.09             |
| Chew           | 52.32             |
| Flic_flac      | 49.83             |
| Pour           | 80.48             |
| Shoot gun      | 32.30             |
| Throw          | 47.84             |
| Clap           | 35.42             |
| Golf           | 83.42             |
| Pullup         | 94.40             |
| Sit            | 34.04             |
| Turn           | 21.91             |
| Climb          | 60.18             |
| Handstand      | 36.26             |
| Punch          | 37.70             |
| Situp          | 91.99             |
| Walk           | 11.64             |
| Climb_stairs   | 42.44             |
| Hit            | 42.86             |
| Push           | 23.90             |
| Smile          | 38.11             |
| Wave           | 24.99             |
| Dive           | 45.74             |
| Hug            | 50.78             |
| Pushup         | 96.37             |
| Smoke          | 36.70             |
| Draw_sword     | 39.33             |
| Hug            | 50.78             |
| Ride_bike      | 57.52             |
| Somersault     | 49.50             |
| Dribble        | 46.0              |
| Kick           | 44.06             |
| Ride_horse     | 37.82             |
| Stand          | 22.85             |
| Drink          | 42.77             |
| Kick ball      | 49.81             |
| Run            | 6.86              |
| Swing_baseball | 57.16             |

FIGURE 8  Performance of our method under different numbers of training sizes and attribute scales on the UCF50 dataset

values of the parameter. In order to simplify our expression, the training size \( p \) is defined to represent the number of training samples in each action class, and the scale of cluster centers \( r \) is represented as the percentage of training set for initializing the attributes.

We firstly study the impact of the training sizes and the number of attributes. When the size of training samples is changing from 20 to 50 with fixed cluster sizes, the experimental results on UCF50 dataset and HMDB51 dataset can be seen in Figures 8 and 9. On both datasets, the number of testing samples is fixed to 20 and the number of action classes is also 20. Figure 8 shows the average recognition accuracy with different values of cluster scales \( r \) when the training size is equal to 20, 30, 40 and 50 on 20 action classes, which are selected randomly from 50 actions in the UCF50 dataset. It can be observed from Figure 8 that the experimental results is getting better when the value of \( r \) is increasing. It can be inferred that the proposed method is a little sensitive to parameter \( r \) in a reasonable range on UCF50 dataset.

As for the HMDB51 dataset, the performance of the experiment is significantly affected over a range of factors such as camera position and motion as well as occlusions. Figure 9 shows that as the cluster scale \( r \) increases, the experimental performance also becomes better for all different training sizes.

Then we test the performance under different values of parameter \( \beta \) and different number of action classes. Experimental results are shown in Figures 10 and 11. We employ 3-fold cross-validation method in these experiments, and the training size is 60 in each fold, meanwhile the number of testing samples is fixed to 20. Figure 10 shows that when the number of action class increases, the average classification accuracy decreases significantly, which means our method is sensitive to the number of action class. When the action class number increases, the margins of different classes become blurred, which can make a great difference to the accuracy of the classification results. We can also see from Figure 10 that the performance of the proposed method remains stable when \( \beta \) is smaller than 10 no matter what the number of action class is, which shows the robustness of our action recognition framework. As shown in Figure 10, a larger value of parameter \( \beta \) will cause a lower overall matching rate, and it seems that the proposed method is insensitive to the small parameter value of \( \beta \). In Figure 11, the experimental results indicate that the result is the best for all different numbers of action class when \( \beta = 0.001 \), and the performance becomes unstable when the value of \( \beta \) is larger than 10.
CONCLUSION

In this paper, for support the application of crowd understanding and analysis, we explore the idea of using middle level semantic concepts to represent human behaviours from videos and argue that these semantic attributes enable the construction of more descriptive methods for human action recognition. Compared with previous works in action recognition domain, the proposed mid-level representation exhibits unique performance in fully utilizing the discriminative information among diverse action features. Each mid-level attribute is focused on the dis-similarity between the action itself and other actions, and is built from the low-level features. This method can make the representation more discriminative by driving the same action’s features more similar while different action’s features more distinct.

As reported in section IV, the experimental results conducted on the KTH dataset demonstrate that the proposed method can obtain better action recognition results compared with the state-of-the-art. We also demonstrate that our method is comparable with other traditional methods on the challenging UCF50 and HMDB51 datasets, even with their limited number of training examples and the diversity of the actions. Especially, our method can reduce the negative effect of the similar actions even in the difficult cases, and overcome the illumination and view angle change to large degree, which verifies the effectiveness and robustness of the proposed method. For the future work, methods based on deep learning which integrate middle-level feature for better performance will be presented.

ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China under Grant 62076199, in part by the China Postdoctoral Science Foundation under Grant 2019M653784, in part by the National Natural Science Foundation of China under Grant 61702498, in part by the National Key R&D Program of China under Grant 2017YFB050290, in part by the CAS Light of West China Program under Grant XAB2017B15, in part by the National Natural Science Foundation of China under Grant 61806193, in part by the National Natural Science Foundation of China under Grant 61701387, and in part by Key Laboratory of Spectral Imaging Technology of Chinese Academy of Sciences under Grant LSIT201801D.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Siyuan Wu https://orcid.org/0000-0002-9041-8587
Nan Wang https://orcid.org/0000-0001-5601-7838

REFERENCES

1. Zhang, P., et al.: Semantics-guided neural networks for efficient skeleton-based human action recognition. IEEE Conf Computer Vision and Pattern Recognition, (2020), pp. 1112–1121
2. Ludl, D., et al.: Enhancing data-driven algorithms for human pose estimation and action recognition through simulation. IEEE Trans Intell Transp Syst 21(9), 3990–3999 (2020)
3. Kamel, A., et al.: Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Trans Syst Man Cybern Syst 49(9), 1806–1819 (2019)
4. Zhao, R., et al.: Bayesian hierarchical dynamic model for human action recognition. IEEE Conf Computer Vision and Pattern Recognition, (2019), pp. 7733–7742
5. Zheng, Y., et al.: Unsupervised learning of human action categories in still images with deep representations. ACM Trans Multim Comput Commun Appl 15(4), 112:1–112:20 (2020)
6. Wang, Q., et al.: NWPU-crowd: A large-scale benchmark for crowd counting. arXiv preprint arXiv:200103360 (2020)
7. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vision 1–21 (2020)
8. Wang, Q., et al.: Density-aware curriculum learning for crowd counting. IEEE Trans. Cybern. 1–13, (2020)
9. Zhang, P., et al.: View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans Pattern Anal Mach Intell 41(8), 1963–1978 (2019)
10. Zhang, J., Hu, H.: Domain learning joint with semantic adaptation for human action recognition. Pattern Recognit 90, 196–209 (2019)
11. Shao, Z., et al.: A hierarchical model for human action recognition from body-parts. IEEE Trans Circuits Syst Video Technol 29(10), 2986–3000 (2019)
12. Meng, F., et al.: Sample fusion network: An end-to-end data augmentation network for skeleton-based human action recognition. IEEE Trans Image Process 28(11), 5281–5295 (2019)
13. López-Nava, L.H., et al.: Human action recognition based on low- and high-level data from wearable inertial sensors. Int J Distributed Sens Networks 15(12), (2019)
14. Patel, C.I., et al.: Human action recognition using fusion of features for unconstrained video sequences. Comput Electr Eng 70, 284–301 (2018)
15. Zhang, Z., Tao, D.: Slow feature analysis for human action recognition. IEEE Trans Pattern Anal Mach Intell 34(3), 436–450 (2012)
16. Minhas, R., et al.: Incremental learning in human action recognition based on snippets. IEEE Trans Circuits Syst Video Techn 22(11), 1529–1541 (2012)
17. Lampert, C.H., et al.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans Pattern Anal Mach Intell 36(3), 453–465 (2014)
18. Wang, Y., et al.: Visual attribute-augmented three-dimensional convolutional neural network for enhanced human action recognition. arXiv preprint arXiv:180502860 (2018)
19. Wu, T., et al.: Clustering-aware structure-constrained low-rank representation model for learning human action attributes. In: IEEE Image, Video, and Multidimensional Signal Processing Workshop, (2016), pp. 1–5
20. Wu, T., et al.: Human action attribute learning from video data using low-rank representations. arXiv preprint arXiv:161207857 (2016)
21. Zhang, Z., et al.: Robust relative attributes for human action recognition. Pattern Anal Appl 18(1), 157–171 (2015)
22. Lampert, C.H., et al.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conf. Computer Vision and Pattern Recognition, (2009), pp. 951–958
23. Liu, J., et al.: Recognizing human actions by attributes. In: IEEE Conf. Computer Vision and Pattern Recognition, (2011), pp. 3337–3344
24. Aly, S., Sayed, A.: Human action recognition using bag of global and local zernike moment features. Multim Tools Appl 78(17), 24923–24953 (2019)
25. Sun, B., et al.: Effective human action recognition using global and local offsets of skeleton joints. Multim Tools Appl 78(5), 6329–6353 (2019)
26. Liu, J., et al.: Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Trans Image Process 27(4), 1586–1599 (2018)
27. Bobick, A.F., Davis, J.W.: The recognition of human movement using templeal templates. IEEE Trans Pattern Anal Mach Intell 23(3), 257–267 (2001)
28. Efros, A.A., et al.: Recognizing action at a distance. In: IEEE Int'l Conf. Computer Vision, (2003), pp. 726–733
29. Sevilla-Lara, L., et al.: On the integration of optical flow and action recognition. In: Pattern Recognition, (2018), pp. 281–297
30. Vemulapalli, C., et al.: Unsupervised human action retrieval using salient points in 3d mesh sequences. Multim Tools Appl 78(3), 2789–2814 (2019)
31. Korbar, B., et al.: Sesampler: Sampling salient clips from video for efficient action recognition. In: IEEE Int'l Conf. Computer Vision, (2019), pp. 6231–6241
32. Yi, Y., et al.: Human action recognition with salient trajectories and multiple kernel learning. Multim Tools Appl 77(14), 17709–17730 (2018)
33. Li, Y., et al.: Actions as moving points. European Conference on Computer Vision (2020), pp. 68–84
34. Wang, R., et al.: GA-STIP: Action recognition in multi-channel videos with geometric algebra based spatio-temporal interest points. IEEE Access 6, 56575–56586 (2018)
35. Wang, H., et al.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conf, (2009), pp. 1–11
36. Laptev, I., et al.: Learning realistic human actions from movies. In: IEEE Conf. Computer Vision and Pattern Recognition, (2008), pp. 1–8
37. Kläser, A., et al.: A spatio-temporal descriptor based on 3d-gradients. In: Machine Vision Conf, (2008), pp. 1–10
38. Majid, M., Safabakhsh, R.: Correlational convolutional lstm for human action recognition. Neurocomputing 396, 224–229 (2020)
39. Zhang, Z., et al.: Human action recognition using convolutional lstm and fully-connected lstm with different attentions. Neurocomputing 410, 304–316 (2020)
40. Si, C., et al.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2019), pp. 1227–1236
41. Meng, F., et al.: Sample fusion network: An end-to-end data augmentation network for skeleton-based human action recognition. IEEE Trans. Image Process. 28(11), 5281–5295 (2019)
42. Liu, Z., et al.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2020), pp. 143–152
43. Zhao, S., et al.: Do not disturb me: Person re-identification under the interference of other pedestrians. In: European Conference on Computer Vision, Springer, (2020), pp. 647–663
44. Li, H., et al.: Jointly learning commonality and specificity dictio naries for person re-identification. IEEE Trans. Image Process. 29, 7345–7358 (2020)
45. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: IEEE Conf. Computer Vision and Pattern Recognition, (2012), pp. 1234–1241
46. Derpanis, K.G., et al.: Efficient action spotting based on a spacetime oriented structure representation. In: IEEE Conf. Computer Vision and Pattern Recognition, (2010), pp. 1990–1997
47. Savarese, S., et al.: Spatial-temporal correlatons for unsupervised action classification. In: IEEE Motion and video Computing, (2008), pp. 1–8
48. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: IEEE Conf. Computer Vision and Pattern Recognition, (2009), pp. 1593–1600
49. Liu, J., et al.: Recognizing realistic actions from videos in the wild. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, (2009), pp. 1996–2003
50. Le, Q.V., et al.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE Conf. Computer Vision and Pattern Recognition, (2011), pp. 3361–3368
51. Gilbert, A., et al.: Fast realistic multi-action recognition using mined dense spatio-temporal features. In: IEEE Int'l Conf. Computer Vision, (2009), pp. 925–931
52. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42(3), 145–175 (2001)
53. Kuehne, H., et al.: HMDB: A large video database for human motion recognition. In: IEEE Int'l Conf. Computer Vision, (2011), pp. 2556–2563
54. Solmaz, B., et al.: Classifying web videos using a global video descriptor. Mach Vis Appl 24(7), 1473–1485 (2013)
55. Cao, Y., et al.: Trajectory-based modeling of human actions with motion reference points. In: Euro. Conf. Computer Vision, (2012), pp. 425–438
56. Wang, L., et al.: Motionlets: Mid-level 3d parts for human motion recognition. In: IEEE Conf. Computer Vision and Pattern Recognition, (2013), pp. 2674–2681

How to cite this article: Sun B, Wu S, Li S, Yuan N, Wang N. Human Behavior Recognition with Mid-Level Representations for Crowd Understanding and Analysis. IET Image Processing. 2021;1–11. https://doi.org/10.1049/ipt.12147