Available online at www.sciencedirect.com

ScienceDirect

CAAI Transactions on Intelligence Technology 1 (2016) 125–136
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original Article

Sequential Bag-of-Words model for human action classification

Hong Liu [a,b,c,*], Hao Tang [a,b], Wei Xiao [a,b], ZiYi Guo [d], Lu Tian [a,b], Yuan Gao [e]

[a] School of Electronic and Computer Engineering, Peking University, Shenzhen Graduate Shcool, 518055, China
[b] Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Peking University, Shenzhen Graduate School, 518055, China
[c] Key Laboratory of Machine Perception, Peking University, 100871, China
[d] School of Software and Microelectronics, Peking University, 100871, China
[e] Institute of Computer Science, Christian-Albrechts-University, 24118, Germany

Available online 21 October 2016

Abstract

Recently, approaches utilizing spatial-temporal features to form Bag-of-Words (BoWs) models have achieved great success due to their simplicity and effectiveness. But they still have difficulties when distinguishing between actions with high inter-ambiguity. The main reason is that they describe actions by orderless bag of features, and ignore the spatial and temporal structure information of visual words. In order to improve classification performance, we present a novel approach called sequential Bag-of-Words. It captures temporal sequential structure by segmenting the entire action into sub-actions. Meanwhile, we pay more attention to the distinguishing parts of an action by classifying sub-actions separately, which is then employed to vote for the final result. Extensive experiments are conducted on challenging datasets and real scenes to evaluate our method. Concretely, we compare our results to some state-of-the-art classification approaches and confirm the advantages of our approach to distinguish similar actions. Results show that our approach is robust and outperforms most existing BoWs based classification approaches, especially on complex datasets with interactive activities, cluttered backgrounds and inter-class action ambiguities.

Keywords: Action classification; Sequential Bag-of-Words; STIP; Probalibity

1. Introduction

Human action classification is an important research topic in the fields of computer vision and pattern recognition. It is a growing challenging problem and well worth studying in the task of automatically analyzing and interpreting video data in computer vision and image analysis. It is of significant use in many applications, such as intelligent surveillance, content-based video retrieval and human–computer interaction. As a hot but difficult problem in the field of computer vision, action classification has been researched for years, but remains a very challenging task. It still has many challenges such as intra-class and inter-class variations, body occlusion, camera movement, environment changing, etc. In action recognition, the performance is frequently susceptible to interference under background changes, variable illumination, camera motion and zooming, and intra-class inhomogeneity or variability. To alleviate these difficult circumstance, discriminative descriptors are of vital importance. These video descriptors are invariant to the variability disturbed by the multiple noise such as distortions, occlusions and so on. In addition, they need yet to encode the spatio-temporal information effectively and robustly. One of the most difficult problems is to distinguish between actions with high inter-ambiguities. Regarding an action as a connection of sub-actions, some action classes consist of similar sub-actions, which greatly increase the difficulty of classification (Fig. 3).

Recent work in action recognition are generally divided into four categories:

* Corresponding author. School of Electronic and Computer Engineering, Peking University, Shenzhen Graduate Shcool, 518055, China.
E-mail addresses: hongliu@pku.edu.cn (H. Liu), haotang@sz.pku.edu.cn (H. Tang), xiaoweipkusz@pkusz.edu.cn (W. Xiao), 1401210562@pku.edu.cn (Z. Guo), lutian@sz.pku.edu.cn (L. Tian), yuan.gao@stu.uni-kiel.de (Y. Gao).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.10.001

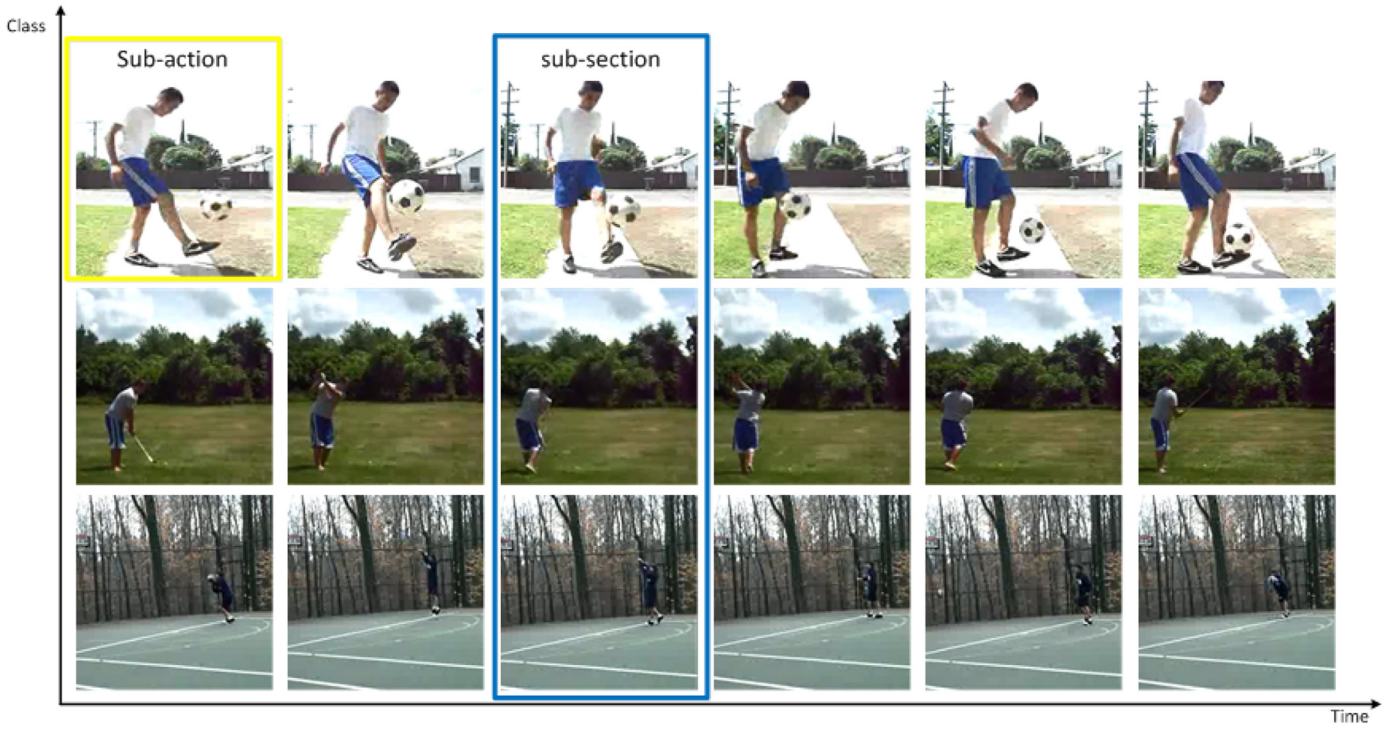


Fig. 1. Sub-sections and sub-actions. Each section is called a sub-action. The *i*th sub-actions of all actions compose the *i*th sub-section.

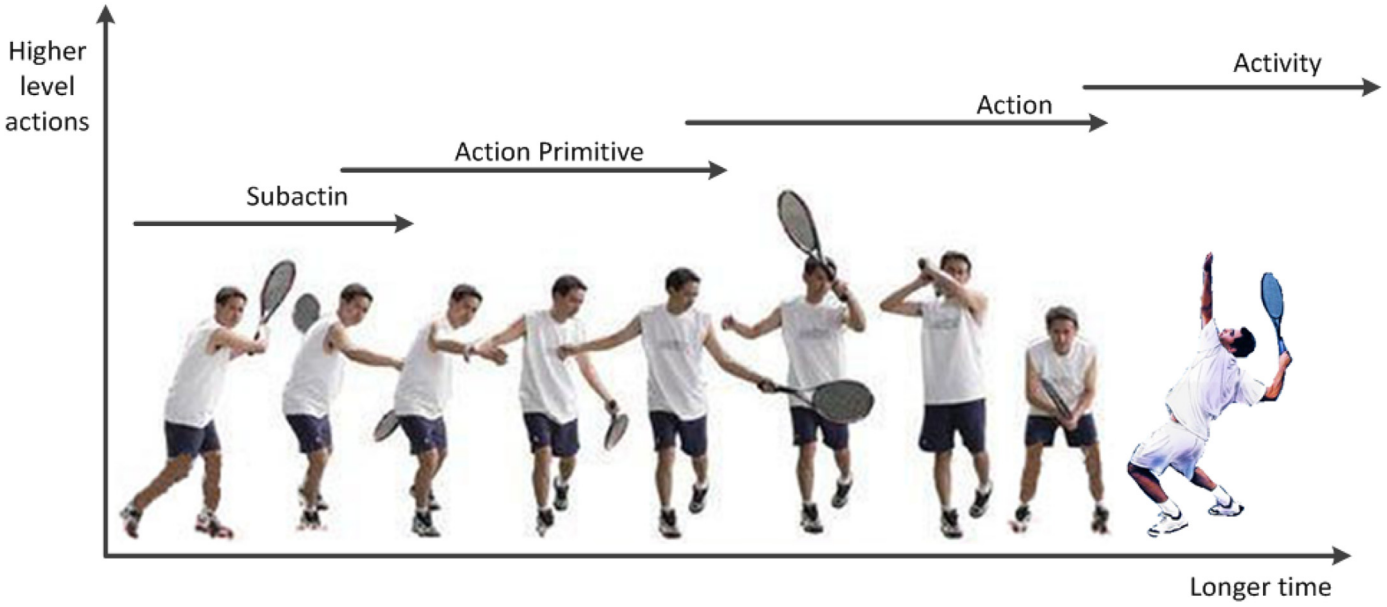


Fig. 2. Different levels of actions, from left to right are “sub-action”, “action primitive”, “action”, “activity”. They are separated by their complexities.

- Space-time shape [8,20];
- Optical flow [23,24];
- Trajectory [11,12,19];
- Local STIPs [4,14].

For instance, Blank et al. [2] treat human actions as 3D shapes induced by the sihouettes in the space-time volume. Efros et al. [5] utilize cross-correlation between optical flow descriptors. Ali et al. [1] apply the trajectories of feet, hand and body into the action recognition framework.

Generally, the first three categories are the global representation methods which rely on the recording circumstances frequently. Meanwhile, as for the last category, Dollar et al. [4] and Leptev et al. [9] concentrate on the local representations of STIPs in the video sequences. In Dollar et al. [4] developed a detector which applied linear 2D Gaussian kernel in the spatial dimensions and 1D Gabor filter in the temporal dimension. The local regions around STIPs are depicted by the proposed descriptors. At last the video sequences are encoded as the frequency histogram of the visual-words. In [9], the Harris

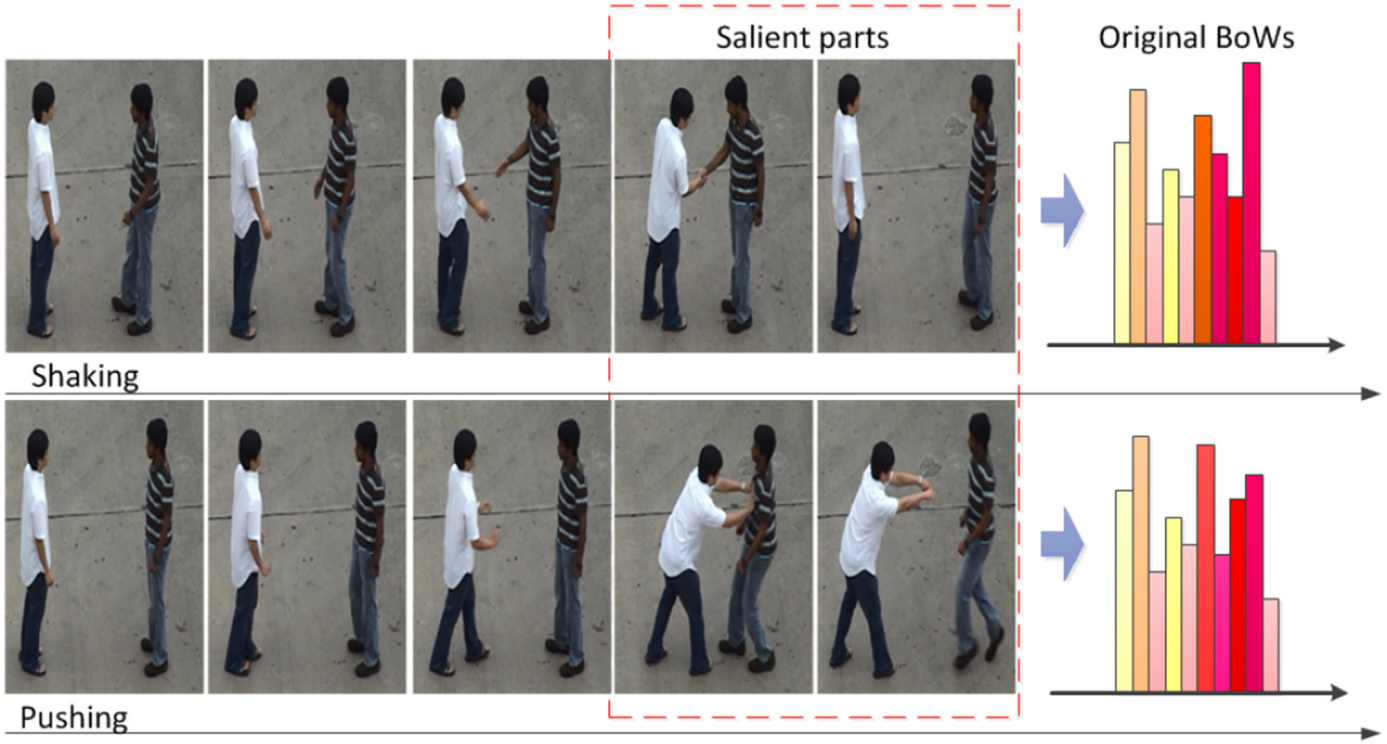


Fig. 3. Examples of two confusing actions: “shaking” and “pushing”. Segmenting them into 5 sections, the former sections are nearly the same. This leads to large ambiguity between their BoWs histograms. But if we focus on the salient parts of the actions, the classification would be much easier.

detector is extended to the space-time domain. Interest points are located using extended Harris detector. The resulting points could be regarded as spatio-temporal corners.

Although BoWs model is popular, it has an essential drawback of only focusing on the number of words but ignoring the spatial-temporal information, which results in ambiguity between different classes of actions. For example, some similar sub-actions happen at different relative period of time, such as standing up at the beginning of an action or at the end. BoWs model is incapable to distinguish them. Moreover, BoWs model handles all the visual words equally, therefore does not lay emphases on the most distinguish parts. The higher proportion of similar sub-actions between classes, the more difficult for classification procedure using original BoWs. Hence, introducing temporal series of BoWs model is of great importance and necessity.

Dollar et al. [4] present the gradient descriptor of STIPs (Spatial-Temporal Interest Points), which obtain better performance compared with other descriptors such as optical flow descriptor, brightness descriptor in their experiments. First, the spatial-temporal cuboids were smoothed in different scales and the image gradients are calculated. Next, the obtained gradient values were connected as the gradient vectors, and the vectors were projected to the lower dimensional space through the Principal Component Analysis (PCA) and finally we got these lower dimensional vectors representation for further applications.

Laptev et al. [10] apply the Histogram of Oriented Gradients (HOG) as the STIPs descriptor. By using HOG descriptor, local object shape and appearance were able to be characterized through the local intensity gradients. Generally, there were two steps in the feature description:

1. The cuboids were divided into small space-time regions;
2. The local 1D-histogram of gradient directions over the pixels for each region were accumulated.

Furthermore, Laptev also combined the Histogram of Optical Flow (HOF) and HOG to form a new descriptor, called HOG-HOF, which outperformed each separate method.

Scovanner et al. [21] proposed the 3-Dimensional Scale Invariant Feature Transform (3D SIFT) descriptor, which accurately captured the spatio-temporal nature of the video sequences and extended the BoW paradigm from 2D to 3D, where the third dimension was time axis. In 3D SIFT descriptor [21]; the gradient magnitude and orientation were represented by $m_{3D}(x,y,t)$, $\theta(x,y,t)$ and $\phi(x,y,t)$. By this means, every angle had been represented by a single unique (θ,ϕ) and each pixel had two values which represented the direction of the gradient in three dimensions. Finally, the orientations were concatenated into a histogram for each 3D sub-region, and the sub-histograms were formed as the desired descriptor.

Zhao et al. [16,22] developed the Local Binary Pattern on Three Orthogonal Planes (LBP-TOP). Delightedly, LBP-TOP had been applied for dynamic texture description and recognition successfully and had obtained good results on facial expression analysis. The proposed algorithm extracted the LBP [15] from three orthogonal *XY*, *XT*, *XT* planes. In this manner, on one hand the spatial information were encoded in *XY* plane. On

the other hand, the spatial temporal co-occurrence information were encoded in *XT* and *YT* planes. Shao et al. [13] proposed to extend the computation of LBP and computed histogramming LBP [7] named Extended LBP-TOP and Extended CSLBP-TOP, respectively. The presented methods could extract more dynamic information inside the spatio-temporal cuboids and be applied on the gradient cuboids.

However, only a small part of previous works focus on capturing visual words' temporal relationships. In some approaches [8,12,20]; spatial-temporal correlations of local features were learned as neighborhoods or correlograms to capture the visual words' spatial-temporal relationships. While these approaches are still too local to capture long-term relationship between words. Other works [11,23,24] counted the co-occurrence between words, but they limited the scope within a small period of time. Recently Ryoo [17] represented an activity as an integral histogram of spatial-temporal features, efficiently modeling how feature distributions change over time, but it could not deal with ambiguities between classes which have similar sub-actions at same relative time. Glaser et al. [6] incorporated temporal context in BoWs models to capture the contextual sequence between words but it still could not focus on the distinctive parts of the actions.

Instead of directly including time information in visual words, we take account of time dimension by segmenting the entire action into small sections. Each section is called a sub-action. The *i*th sub-actions of all actions compose the *i*th sub-section, as shown in Fig. 1.

Then sequential BoWs model is used for video representation, the classification is described as a series of sub-classes classification problems. In this way, we can apply our approach to original BoWs and spatial-improved BoWs, such as approaches in our previous works [11,24]. Satkin et al. [19] extracted the most discriminative portion of a video for training according to the accuracy of a trained classifier. Inspired by this, the discriminative parts of the action is emphasised by assigning them high weights and salience values. The weight indicates the probability of a sub-action belonging to a certain class and the salience means how much a sub-section is distinguished from others. Then the effect of similar sub-actions is minimized, so we can focus on the difference between classes (salient parts in Fig. 3). On the one hand, if similar sub-actions happen at different relative time, they will be in different sub-section and classified separately. On the other hand, if they can not be separated, then low salience values are given to them to reduce their influence. Finally, sub-actions are classified separately and the results will be gathered together through voting. The experiments are implemented on UT-interaction dataset and a more challenging Rochester dataset. The results show our approach can achieve robust and accuracy beyond most related approaches.

The rest of the paper is organized as follows. In Section 2, we introduce the reason for using sub-actions and illustrate the framework of our approach. Section 3 and Section 4 describe the segmentation and classification approach respectively. In Section 5, we conduct experiments on UT-Interaction and Rochester datasets and compare our approach with other BoWs based approaches. Finally, conclusions are drawn in Section 6.

2. The proposed framework

As shown in Fig. 2, human movement can be described at various levels of complexities [16]. Usually, an activity refers a whole expression of a series of actions such as “play tennis”. An action is the element of activity such as “running” or “jumping”. It is often short and represents less motion information. As for action primitive, it is a very short period of action that can not be performed individually. Action primitives are components of actions and actions are components of activities as well. For example, “left leg forward” is an action primitive, whereas “running” is an action. “Jumping hurdles” is an activity that contains starting, jumping and running actions.

In general, action classification is performed at the first two levels. Many different action classes are unavoidable in sharing similar or same action primitives. This largely increases the difficulty to distinguish different classes. However, action classification at primitive level is also impractical. The reason is that there are innumerable action primitives due to the flexibility of human movement. Moreover, different action classes may be composed by different numbers of action primitives, which is not suitable to perform uniform preprocessing for all action classes. To solve this problem, we define sub-action instead of action primitive. Sub-action is also a small period of action, but it is not previously defined or fixed for each action classes. Sub-action is segmented according to a certain action automatically. All the action classes to be classified will have same number of sub-actions. This approach can not only solve the problem mentioned above but also be faster, more flexible and adaptive.

In computer vision field, the Bag-of-Words model (BoWs model) can be applied to image classification, by treating image features as words. In document classification, a bag of words is a sparse vector of occurrence counts of words, that is, a sparse histogram over the vocabulary. In the field of computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features. To represent an image using BoWs model, an image can be treated as a document. Similarly, “words” in images need to be defined too. To achieve this, three steps are usually included:

1. Feature detection;
2. Feature description;
3. Codebook generation.

A definition of the BoWs model can be the “histogram representation based on various features”.

As shown in Fig. 4, first local features are extracted from action videos. The videos are then treated as volumes of visual words. To extract sub-actions, we chop the volumes into small clips according to the intensity of actions. Distances between clips are accumulated. Then segment the accumulated distance into equal parts, *i.e.*, the sub-actions. Sequential histograms are created for each action respectively to describe the action. Before classification, similarity between sub-actions in the same sub-section is figured out with training data. These similarities can be regarded as weights for voting. Meanwhile

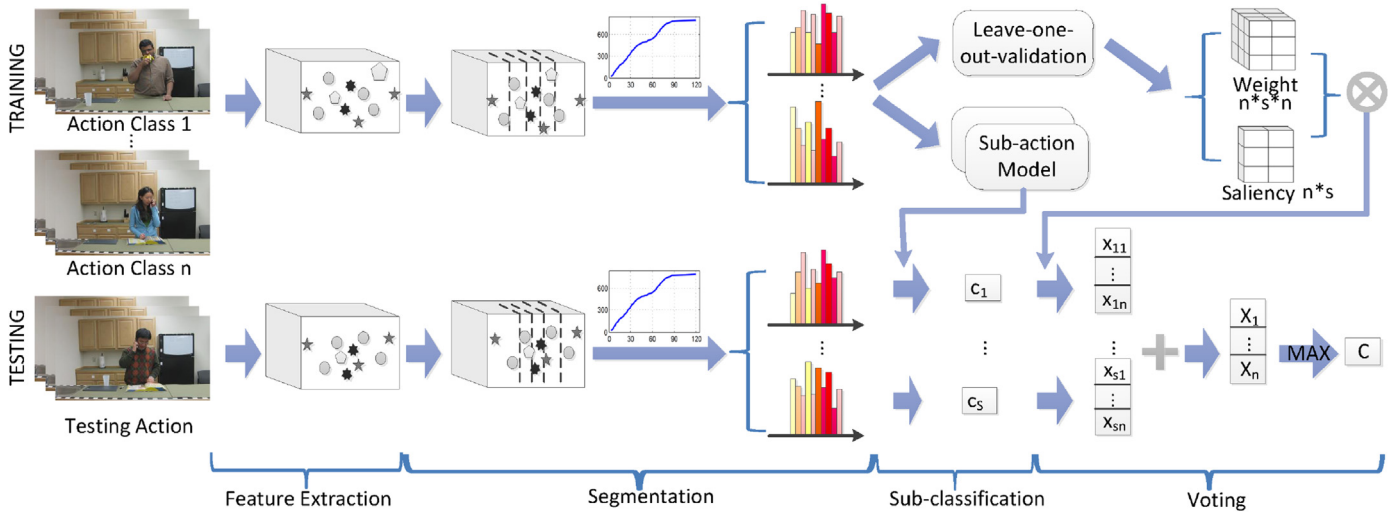


Fig. 4. Framework of our approach. Firstly, videos are transformed into volumes of visual words. Secondly, a two-stage segmentation is conducted. Stage-1 chop volumes into clips according to the density of words, stage-2 divide accumulated distance between clips into equal parts. Thirdly, normal classification is done to each sub-section respectively and results are recorded as c_j. Finally, the final result is decided by a voting process.

salience value for each section is figured out according to the pre-classification accuracy. After sub-section classification, a vote scheme is conducted finally. Category of a test instance A is decided by the equation below:

$$A = \max_{i \in C} \sum_{j=1}^{N_s} \omega_j(i, c_j) s_j(c_j), \tag{1}$$

where C denotes all the possible categories, N_s represents the number of sub-sections, c_j denotes A's sub-classification result in sub-section j and $\omega_j(i,c_j)$ is the weight of the jth sub-action belonging to class i when it is classified to sub-class c_j. The last $s_j(c_j)$ stands for the saliency of the jth sub-action. The instance will be classified to the category with the highest score.

3. Action segmentation

Our approach takes advantage of local spatio-temporal features to represent actions. After extracting local features into visual words from video sequence, we conduct a two-stage segmentation to avoid acquiring inequality sub-actions caused by the scale, range, rate or other individual difference between actors. Optimal segmentation could narrow the classification and disambiguate similar sub-actions occur at different relative time. We briefly introduce the visual words extraction approach in Subsection 3.1, and in Subsection 3.2, our segmentation approach to acquire sub-actions is presented.

3.1. Visual words extraction

As shown in Fig. 5, local feature based representation is widely used in human action analysis for its resistant to clutters and noise. Firstly, Spatial-Temporal Interest Points (STIPs) are detected, small cuboid is extracted around each interest point. Then descriptors are generated to represent local information. There are many different local detectors and descriptors being proposed. Since the average length of sub-action is short, our approach avoid using those detecting approaches which are too sparse. Dense detector [18,25] is among good choices. Specifically, the cuboid detector proposed by Dollar et al. [4] and the scale-invariant detector proposed by Willems et al. [25]. Dollar's descriptor is used in our experiment.

Typically, Dollar [4] proposed an alternative spatio-temporal feature detector for action recognition. The

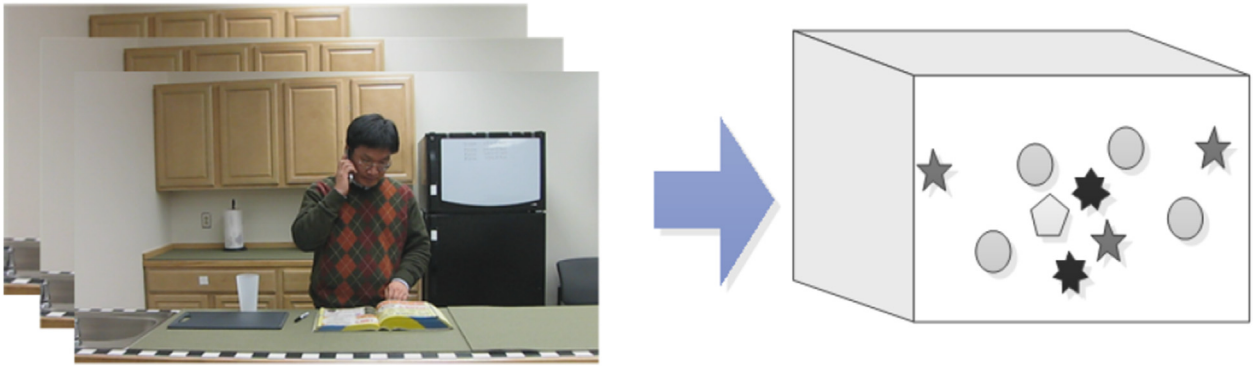


Fig. 5. Extraction of visual words.

detector is based on a set of separable linear filters which regards the spatial and temporal dimensions in different ways. In [4], to extract the STIPs from video sequence I, the response function is given by:

$$R = (I*g*h_{even})^2 + (I*g*h_{odd})^2, \tag{2}$$

where $g(x,y,\sigma)$ is the 2D Gaussian smoothing kernal, applied along the spatial dimensions, and h_{even},h_{odd} are a quadrature pair of 1D Gabor filters applied temporally which are represented as:

$$\begin{cases} h_{even}(t;\tau,\omega) = -cos(2\pi t\omega)e^{\frac{-t^2}{\tau^2}} \\ h_{odd}(t;\tau,\omega) = -sin(2\pi t\omega)e^{\frac{-t^2}{\tau^2}} \end{cases}. \tag{3}$$

In Equation (3), ω represents the underline frequency of cosine in the Gabor filter. When ω=4/τ, the number of parameters in the response function R in Equation (2) is reduced to two: σ, τ, corresponding roughly to the spatial and temporal scale of the detector.

Fig. 6. The key images of “answer phone” acted by two actors. (a), (b), (c), (d), (e), (f) correspond to six segmenting points A, B, C, D, E, F in Fig. 7. (a) and (f) are the start and end of the video. In the first section, actions change from (a) to (b), both actors' hands go forward to get the phone. In the second section (b) to (c), the phones are taken closer. In the third section (c) to (d), actors open the clamshell phones. In the forth section (d) to (e), actors raise the phones. In the final section (e) to (f), actors listen to the phones. Both final sections are longer because actors move little.

Then k-means clustering algorithm is used to build a visual word dictionary and each feature descriptor is assigned to the closest word in the vocabulary. Finally, as shown in Fig. 5, a video with N frames is described as a sequence of frames with visual words:

$$video = [f_1, f_2, \ldots, f_N], \tag{4}$$

where,

$$f_i = [w_{1_i}, w_{2_i}, \ldots, w_{n_i}], \tag{5}$$

w_{n_i} represents the nth visual word in the ith frame.

3.2. Sub-action segmentation

To segment actions efficiently, we should achieve two goals. First, ensure the sub-actions in the same sub-section of the same action class are of the same type, ignoring the speed differences between actors. Second, all sub-actions should capture enough motion information for classification. The two-stage segmentation can reach the above goals nicely, which is detailed below.

3.2.1. Chop clips

In first segmentation stage, the entire video is chopped into normalized clips with approximately equal numbers of feature points. The kth clip ends up in the x_k th frame, x_k should satisfy the equation below:

$$\sum_{i=1}^{x_k-1} n_i < \left(\sum_{i=1}^{N} n_i\right) * k/N_c \leq \sum_{i=1}^{x_k} n_i, \tag{6}$$

where N_c is the clip number.

Here we use point density to chop clips for further segmentation instead of using number of frames to ensure there is enough motion information in all clips. Generally, dense feature points infer strenuous action. The intensity of the action may be changed over the entire action process. Some of the action primitives maybe moderate, so there could be few interesting points in these frames and result in insufficient information. Moreover, it also balances the speed difference between different actors or actions. We compute a histogram of spatial-temporal word occurrence for each video clip, the kth clip is described as:

$$h_k = hist\left(\left[w \middle| w \in f_j, x_{k-1} < j < x_k\right]\right). \tag{7}$$

3.2.2. Segment sub-actions

In second segmentation stage, motion range is calculated to segment sub-actions. The motion range is measured with χ^2 distance between neighbor clips. The distance from clip i to clip j is:

$$dist(i,j) = \sum_{k=i}^{j-1} \chi^2(h_k, h_{k+1}). \tag{8}$$

Then accumulated distance of the whole action is divided into equal parts and the motion range of sub-actions is figured out:

$$T = dist(1, N_c)/N_s, \tag{9}$$

where N_s represents the number of sub-actions. T is used as the threshold to segment clip series. In fact, the distance between clips infers not only the range of motion, but also the changing extent of the action primitives' type. Finally, the sub-actions are segmented by the equation below:

$$dist(i,j) \leq T < dist(i,j+1), \tag{10}$$

where i, j are the beginning and ending clips of the sub-section. A stable segmentation over classes is achieved by concatenating the adjacent clips together.

A segmentation example is illustrated in Fig. 6, and the corresponding accumulated distance curves are shown in Fig. 7. By segmenting action like this, we can eliminate speed and range differences between instances. Therefore, for a certain class, same sub-actions could be basically segmented to the same sub-section ignoring the delicate length difference between instance. Sequential BoWs is formed for each sub-action respectively for weight calculation and sequential classification. All segmentation steps are shown in Algorithm 1. The algorithm focuses on sub-action segmentation to achieve equally distribution among different sections.

Algorithm 1 Sub-action Segmentation

Require: Visual word set W, their location $sub = x, y, t$, number of clips N_c, number of sub-actions N_s
Ensure: Sub-action histograms H

```
Sort W by the order of t in sub
b = sizeof (W)/N_c;
for i = 1 to N_c do
    h(i) = hist (W((i − 1) ∗ b, i ∗ b));
    if i ≠ 1 then
        X(i − 1) = χ²(h(i − 1), h(i));
    end if
end for
T = sum(X)/N_s; temps = 0;is = 1; seg(1) = 1;
for i = 1 to N_c do
    temps = temps + X(i);
    if temps > T ∗ is then
        seg(is + 1) = i; is = is + 1;
    end if
end for
for i = 1 to N_s do
    H(i) = sum(h(seg(i)) : h(seg(i + 1) − 1))
end for
```

4. Sub-classification and voting

A pre-classification using training data is conducted to figure out the weight and salience value for each sub-action. The weight shows the sub-action's discrimination with other

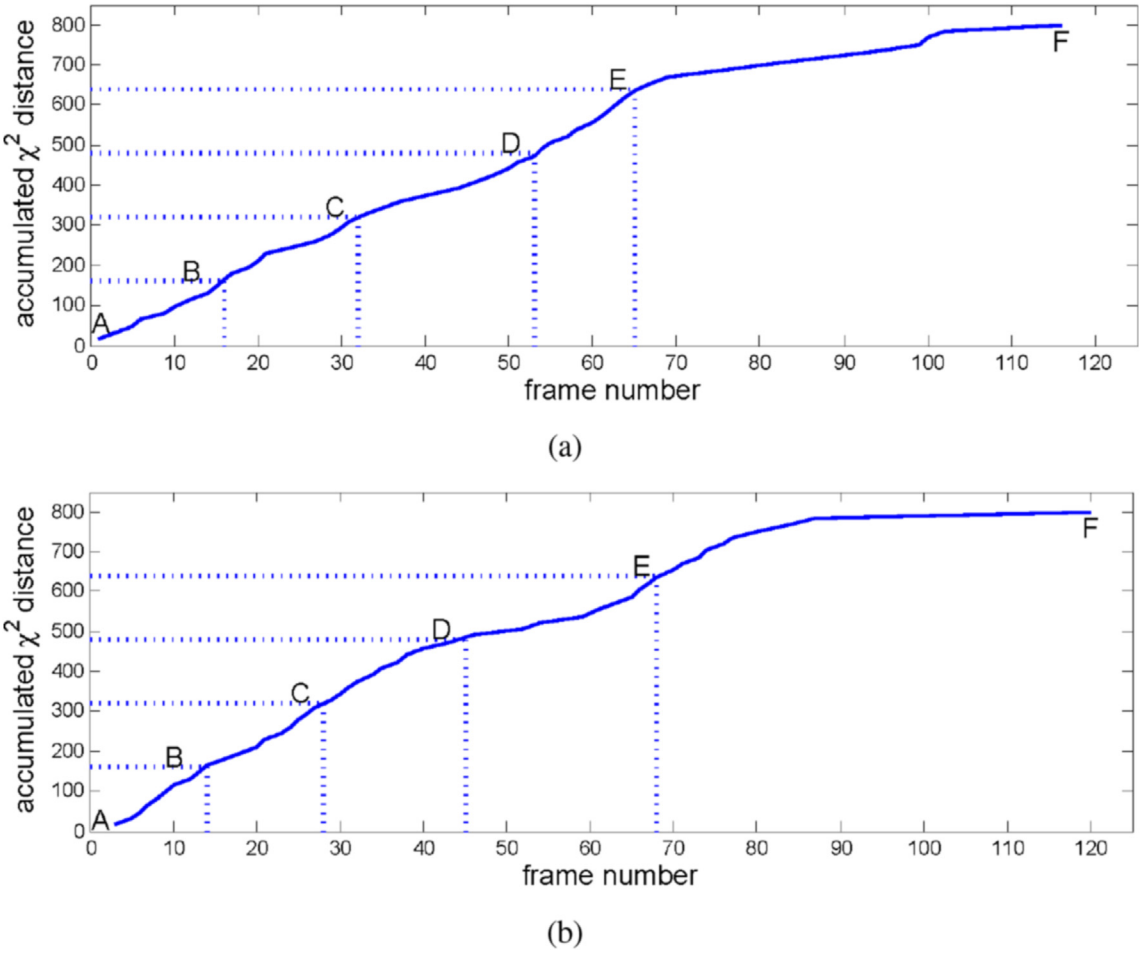


Fig. 7. Example of accumulated χ^2 distance changing over time. (a) represents the upper actor in Fig. 6 and (b) represents the lower one. The accumulated distance curves are similar although they are acted by different actor and rate. B, C, D, E are four equal segmenting points. Their corresponding images are shown in Fig. 6. At each point, the actions are almost executed to the same ratio.

sub-classes, while the salience indicates the sub-action's importance within it's own class.

4.1. Weight calculation

Similar sub-actions may occur in the same sub-section of different classes, hence the result of direct voting would be poor. The pre-classification result $M(i,k)$ represents the percentage of sub-class i be classified to sub-class k. If the jth sub-action is classified as sub-class k, the probability for this sub-action belonging to class i is calculated below:

$$\omega_j(i,k) = P\left(i \middle| c_j = k\right) = \frac{M_j(i,k)}{\sum_{l \in C} M_j(l,k)} \tag{11}$$

This value can be regarded as weight to eliminate the ambiguity between sub-classes. The difference between weights shows the dissimilarity between sub-classes. The smaller the difference is, the higher the similarity shows. Similar sub-actions give approximately equal increase to their own category during voting, so the classification result is barely changed.

4.2. Saliency calculation

The salience value for each sub-action is figured out to differ the importance of each sub-action within the class. In some sub-section, sub-actions are at low similarity so the classification accuracy would be high. We assign high salience scores to sub-actions in such sub-sections. While for some other sub-sections with large ambiguity, classification may be hard. To reduce the effect of these sub-sections, low salience scores are assigned. For a single category, we can figure out its classification accuracy in each sub-section via training sets. But at the testing time, we can not ensure the category for an action, so an average value is used as below:

$$s_j(k) = \sum_{l \in C} \frac{M_j(l,l)}{\sum_{i=1}^{N_s} M_i(l,l)} * \omega_j(l,k) \tag{12}$$

Fig. 9 shows an example of calculated salience maps for UT-Interaction scene-1, scene-2 and the difference between them. To each column, the salience value is proportional to the sub-action's distinctiveness. For example the first column

(a) UT-Interaction scene 1 (b) UT-Interaction scene 2

(c) Rochester

Fig. 8. Examples in datasets, from top to down are UT-Interaction scene 1, UT-Interaction scene 2 and Rochester.

(shake hands), the shaking parts of the action are more salient. But to the fourth column (point), the whole process is quite different from other actions, so its salience is evenly distributed. Although the environment of the two scenes is different, their salience maps are quite similar. The salience ranges from 0 to 1, and the average difference between them is 0.1777. Except the red grid on the right top of (c) caused by overact in scene-2, most difference is less than 0.2. This means our segmentation and salience calculation approach is robust to the inner-class variation and environmental change.

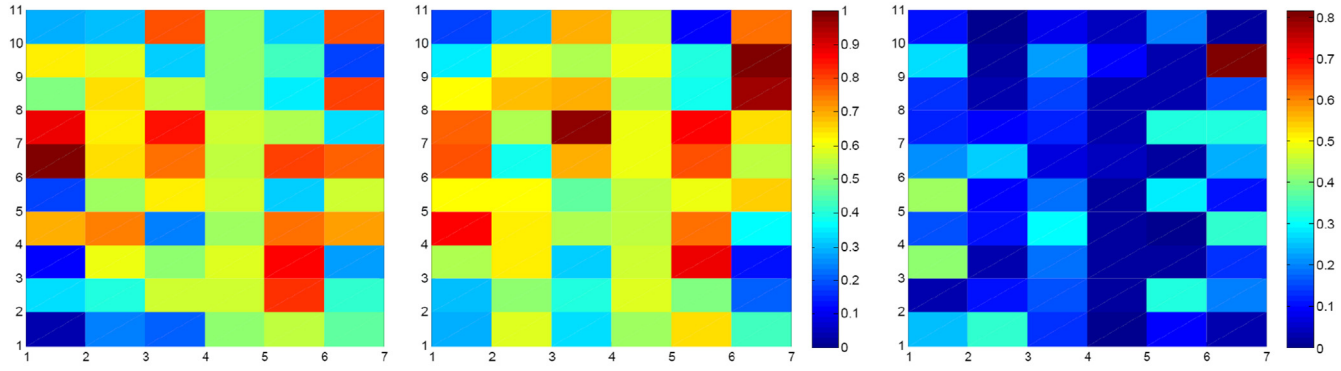
Finally, category of a test instance is decided by the multiplication of those two scores Equation (11), (12) calculated above. By combining these two weights, the effect of ambiguous sub-actions in the same sub-section is weakened, the salient points at temporal dimension are stressed. We can focus the classification on the distinguishing action parts rather than being confused by those similar parts.

5. Experiments and discussions

In this section, the proposed classification approach is implemented and evaluated on two challenging datasets, UT-Interaction [18] and Rochester [14]. Fig. 8 shows some actions from these two datasets. We compare our results to state-of-the-art classification approaches and confirm the advantages of our approach to distinguish similar actions.

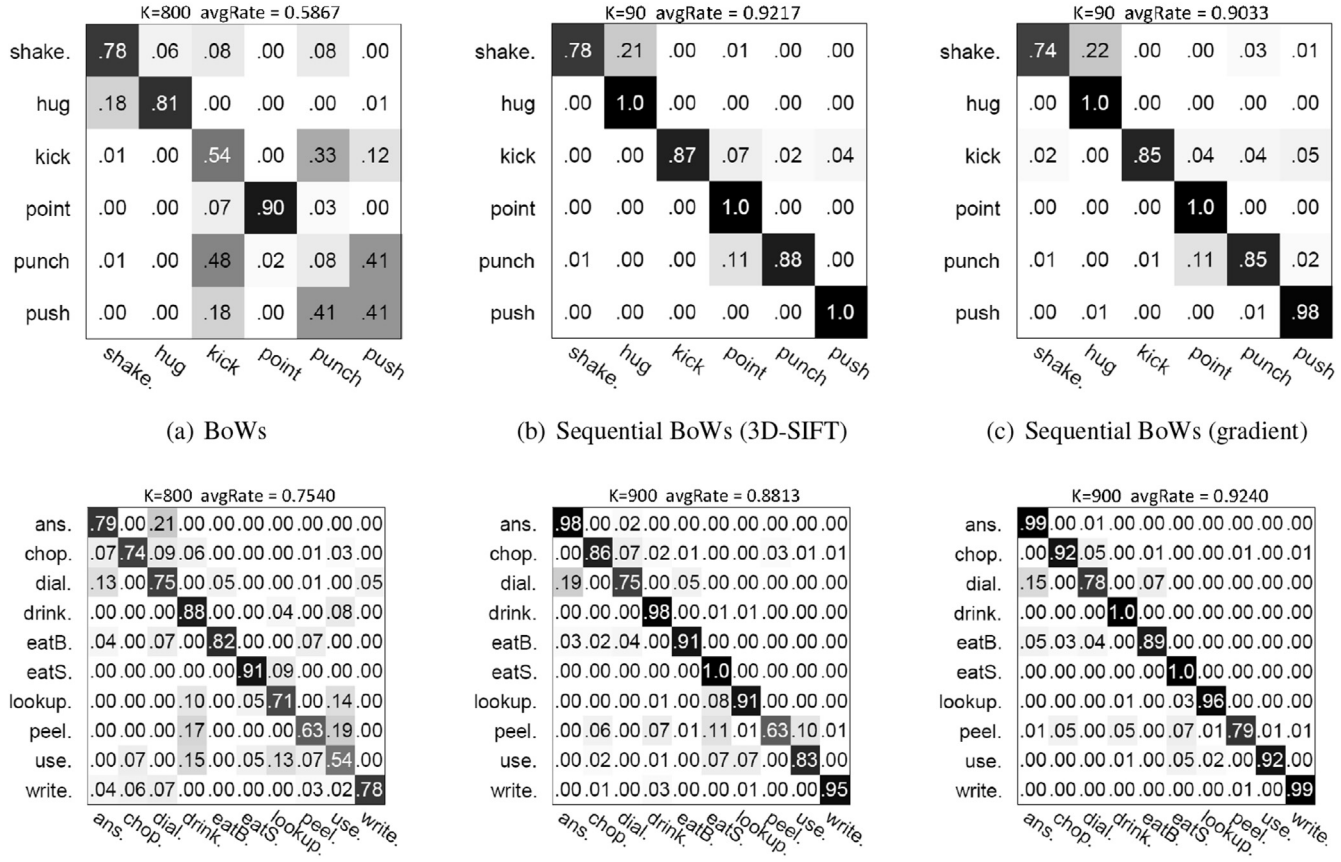
The segmented version of the UT-Interaction dataset contains videos of six types of human actions [3]. All besides "pointing" are interactive activities and are performed by two actors. There are two sets in the dataset performed in different environment (Fig. 8 (a) and (b)). One is relatively simple, which is taken in a parking lot. The other is more complex with moving trees in the background. Each activity is repeated for 10 times per set by different actors, so there are totally 120 videos. The videos are taken with camera jitters and/or passerby and have been tested by several state-of-the-art approaches [11,17]. Rochester dataset contains 150 videos of 10 types of actions. Each category is performed by five actors, repeated for three times in the same scenario. The inter-class ambiguities of these two datasets are both large.

We use the cuboid feature detector [4] as local STIPs detector for its simplicity, fastness and generality [22]. A 100-dimensional Dollar's gradient descriptor [4] and a 640-dimensional 3D-SIFT descriptor [21] are used respectively to describe the STIP-centered cuboids. Other feature detectors and descriptors can also be used to acquire features. For both datasets, the cuboid size is $w = h = 1$ *pixels*, $\tau = 2$ *frames*, threshold is 0.0002 when using gradient descriptor. When using 3D-SIFT, $w = h = 2$ *pixels*, $\tau = 3$ *frames*, the threshold is 0.0001. After extracting features from videos, k-means clustering is used to transform them into visual words. For UT-Interaction dataset, the cluster numbers in two sceneries



(a) Scene-1 (b) Scene-2 (c) Difference

Fig. 9. Salience maps calculated on UT-Interaction scene-1, scene-2 and the difference between them. Each grid represents a sub-action's salience. The horizontal axis represents different classes, from left to right are “shake hands”, “hug”, “kick”, “point”, “punch”, and “push”. The vertical axis represents sub-actions align by time.



(a) BoWs (b) Sequential BoWs (3D-SIFT) (c) Sequential BoWs (gradient)

(d) BoWs (e) Sequential BoWs (3D-SIFT) (f) Sequential BoWs (gradient)

Fig. 10. Confusion matrices for scene-1 of UT-Interaction (upper) and Rochester (lower) are shown in row one and row two. From left to right, BoWs, sequential BoWs with 3D-SIFT descriptor and sequential BoWs with gradient descriptor. K is the cluster number.

are 90 and 140. The cluster number in Rochester is 900. Classification is conducted using 1-NN since the performance of SVM and 1-NN is quite close [22] while 1-NN is faster.

During the segmentation, we use χ^2 distance to measure motion range. The clip number N_c and the sub-action number N_s are decided by iteration tests. N_c is a bit smaller than the frame number of the shortest video, N_s is associated to the action's complexity. We set $N_c = 140, N_s = 90$ for UT-Interaction scnene-1, $N_c = 120, N_s = 90$ for UT-Interaction scnene-2, and $N_c = 150$, $N_s = 20$ for Rochester. Although the average video length of Rochester is longer than that of UT-Interaction, the actions are slow and some moments are even motionless, so the sub-action

number of Rochester is smaller. Pre-classification is done in the training set to figure out the weight and salience. The leave-one-sequence-out cross validation setting is used in all experiments. All confusion matrices are the average of 10 runs since k-means clustering is randomly initialized.

Confusion matrices of UT-Interaction scene-1 and Rochester are shown in Fig. 10. In each column, our sequential BoWs approach using different descriptors is compared with Dollar's original BoWs. The result of our approach is much better than original BoWs. On UT-Interaction scene-1, errors among "kick", "punch" and "push" are most obvious in Fig. 10 (a). These three actions share a lot of same sub-actions. Moreover their unique sub-actions (stick out one's arms/leg to conduct a hit/push) are not only short but also alike to each other. Original BoWs model does not have the ability to capture their differences. Our approach can highlight their difference and solve this problem in an extent. On Rochester dataset, the results of "lookup in phone book", "peel banana" and "use silverware" are unsatisfactory. Our approach can decrease those errors since temporal structure and salient parts are especially considered.

Table 1 compares the classification accuracy of our approach with state-of-the-arts on UT-Interaction and Rochester. Cluster number K is given to indicate the complexity of those approaches as there is no unified comparison approach. "K" is proportional to the algorithm's dimensionality reduction ability. In our approach, Gradient descriptor [4] shows better results on UT-Interaction scene-2 and Rochester than 3D-SIFT because it extracts more feature points to describe the sub-actions more sufficiently. On UT-Interaction scene-1, our result is comparable to [11,17]; but manifest a faster computational speed. Noting that [11] focuses on the spatial structure between visual words, hence it can be combined with our approach. Approach in [17] aims at action prediction, and it conducts a iterate segment matching between classes and is inefficient to action classification. On both UT-Interaction scene-2 and Rochester datasets, our approach shows the best performance.

6. Conclusions

The typical Bag-of-Words (BoWs) model ignore the spatial and temporal structure information of visual words, which brings ambiguities among similar actions. In order to deal with it, in this paper, we present a novel approach called sequential BoWs for efficient human action classification. It reduces the ambiguity between classes sharing similar sub-actions and captures the action's temporal sequential structure by segmenting the action into pieces.

Each sub-action has a tiny movement within a narrow range of action. Then the sequential BoWs are created, in which each sub-action is assigned with a certain weight and salience to highlight the distinguishing sections. It is noted that the weight and salience are figured out in advance according to the sub-actions discrimination evaluated by training data. Salient pieces are stressed by assigning them higher weight and salience. Finally, those sub-actions are used for classification respectively, and voting for united result. Since our approach does not operate on the descriptors directly, it can be combined with many mid-level descriptors. In experiments, the proposed approach is compared with the state-of-the-arts on two challenging datasets, UT-interaction and Rochester datasets. The results demonstrate its higher robustness and accuracy over most state-of-the-art classification approaches especially on complex datasets with cluttered backgrounds and inter-class action ambiguities.

Table 1
Comparing proposed approach with state-of-the-arts.

Method	Scene 1/K	Scene 2/K	Rochester/K
Dollár et al. [4]	58.67%/800	53.33%/800	75.40%/800
Sun and Liu [23]	82.67%/120	79.22%/120	−
Ryoo [17]	88.00%/800	77.00%/800	−
Ryoo et al. [18]	−	−	80.00%/4000
Messing et al. [14]	−	−	89.00%/400
Liu et al. [11]	95.00%/450	86.67%/450	88.00%/500
Ours (3D-SIFT)	92.17%/90	85.83%/140	88.13%/900
Ours (gradient)	90.33%/90	**91.83%**/140	**92.40%**/900

Acknowledgments

The authors acknowledge the financial support by the National Natural Science Foundation of China (NSFC, No. 61340046), the National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, JCYJ20130331144716089) and the Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP, No. 20130001110011). The authors wish to thank the anonymous reviewers for their helpful comments and suggestions.

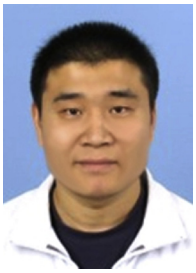
References

[1] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1−8.
[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: IEEE International Conference on Computer Vision (ICCV) vol. 2, 2005, pp. 1395−1402.
[3] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, Elsevier Comput. Vis. Image Underst. (CVIU) 117 (6) (2013) 633−659.
[4] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65−72.
[5] A. Efros, A.C. Berg, G. Mori, J. Malik, et al., Recognizing action at a distance, in: IEEE International Conference on Computer Vision (ICCV), 2003, pp. 726−733.
[6] T. Glaser, L. Zelnik-Manor, Incorporating temporal context in bag-of-words models, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2011, pp. 1562−1569.
[7] M. Heikkilä, M. Pietikäinen, C. Schmid, Springer Comput. Vis. Graph. Image Process. (2006) 58−69.
[8] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2046−2053.
[9] I. Laptev, Springer Int. J. Comput. Vis. (IJCV) 64 (2−3) (2005) 107−123.

[10] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1−8.
[11] H. Liu, M. Liu, Q. Sun, Learning directional co-occurrence for human action classification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1235−1239.
[12] J. Liu, M. Shah, Learning human actions via information maximization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1−8.
[13] R. Mattivi, L. Shao, Springer Comput. Anal. Images Patterns (CAIP) (2009) 740−747.
[14] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: IEEE International Conference on Computer Vision (ICCV), 2009, pp. 104−111.
[15] T. Ojala, M. Pietikäinen, T. Mäenpää, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 24 (7) (2002) 971−987.
[16] R. Poppe, Elsevier Image Vis. Comput. (IVC) 28 (6) (2010) 976−990.
[17] M. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1036−1043.
[18] M. Ryoo, C.-C. Chen, J. Aggarwal, A. Roy-Chowdhury, An overview of contest on semantic description of human activities (sdha), in: Springer Recognizing Patterns in Signals, Speech, Images and Videos, 2010, pp. 270−285.
[19] S. Satkin, M. Hebert, Modeling the temporal extent of actions, in: Springer European Conference on Computer Vision (ECCV), 2010, pp. 536−548.
[20] S. Savarese, A. DelPozo, J.C. Niebles, L. Fei-Fei, Spatial-temporal correlatons for unsupervised action classification, in: IEEE Workshop on Motion and Video Computing (WMVC), 2008, pp. 1−8.
[21] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM International Conference on Multimedia (ACM MM), 2007, pp. 357−360.
[22] L. Shao, R. Mattivi, Feature detector and descriptor evaluation in human action recognition, in: ACM International Conference on Image and Video Retrieval, 2010, pp. 477−484.
[23] Q. Sun, H. Liu, Action disambiguation analysis using normalized google-like distance correlogram, in: Springer Asian Conference on Computer Vision (ACCV), 2013, pp. 425−437.
[24] Q. Sun, H. Liu, Learning spatio-temporal co-occurrence correlograms for efficient human action classification, in: IEEE International Conference on Image Processing (ICIP), 2013, pp. 3220−3224.
[25] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Springer European Conference on Computer Vision (ECCV), 2008, pp. 650−663.

Prof. Hong Liu received the Ph.D. degree in mechanical electronics and automation in 1996, and serves as a Full Professor in the School of EE&CS, Peking University (PKU), China. Prof. Liu has been selected as Chinese Innovation Leading Talent supported by "National High-level Talents Special Support Plan" since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as Pattern Recognition, IEEE Trans. on Signal Processing, and IEEE Trans. on PAMI.

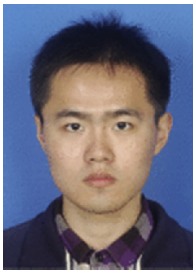
Hao Tang received the B.E. degree in electronics and information engineering in 2013 and is working toward the Master degree at the School of Electronics and Computer Engineering, Peking University, China. His current research interests are image classification, hand gesture recognition, gender recognition, image retrieval, action recognition and deep learning. He has published several articles in ACM Multimedia Conference (MM), IEEE International Conference on Image Processing (ICIP) and International Joint Conference on Artificial Intelligence (IJCAI).



Wei Xiao received his Ph.D in computer science and technology in Tsinghua University, China. He is currently a postdoctoral researcher on Human-Robot Interaction (HRI) in Peking University, China. His research interests include computer vision and HRI. He has published several articles in IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems and ACM Multimedia Conference (MM).

Ziyi Guo received her B.E. degree in digital media technology in 2014, and is working toward the Master Degree in the School of software and microelectronics, Peking University, China. Her research interests include interactive media technologies, interaction design and human action recognition.

Lu Tian received her master degree in computer science and technology in Human-Robot Interaction (HRI) Lab, Peking University Shenzhen Graduate School, China. Her research interest is mainly about human action recognition. She has published articles in IEEE International Conference on Image Processing (ICIP).



Yuan Gao received the B.E. degree in intelligent science and technology from Xidian University in 2012. Then he obtained the M.S. degree in computer applied technology from Peking University in 2015. Currently, he is working toward the Doctor degree in Christian-Albrechts-University of Kiel, Germany. His research interests include object detection, 3D reconstruction, facial expression and gender recognition. He has published articles in IEEE International Conference on Image Processing (ICIP).