LETTER

A Novel 3D Gradient LBP Descriptor for Action Recognition

Zhaoyang GUO[†a)], Xin'an WANG[†], *Nonmembers*, Bo WANG[†], *Member, and* Zheng XIE[†], *Nonmember*

SUMMARY In the field of action recognition, Spatio-Temporal Interest Points (STIPs)-based features have shown high efficiency and robustness. However, most of state-of-the-art work to describe STIPs, they typically focus on 2-dimensions (2D) images, which ignore information in 3D spatio-temporal space. Besides, the compact representation of descriptors should be considered due to the costs of storage and computational time. In this paper, a novel local descriptor named 3D Gradient LBP is proposed, which extends the traditional descriptor Local Binary Patterns (LBP) into 3D spatio-temporal space. The proposed descriptor takes advantage of the neighbourhood information of cuboids in three dimensions, which accounts for its excellent descriptive power for the distribution of grey-level space. Experiments on three challenging datasets (KTH, Weizmann and UT Interaction) validate the effectiveness of our approach in the recognition of human actions.

key words: *action recognition, spatio-temporal interest points, local binary pattern*

1. Introduction

In the action recognition, discriminative descriptors are indispensable and of vital importance. They are not only invariant to the variability disturbed by the noise, but also have the ability which encodes spatio-temporal cues effectively. Among recent work in action recognition, approaches are generally divided into two classes according to [1]: global representation, such as, space-time shape [2], optical flow [3], and trajectory [4]; local representation, for instance, local STIPs [5], [6]. In this paper, we focus on the latter one and propose the 3D Gradient LBP descriptor and perform human action classification along the pipeline of Mattivi and Shao [7]. We observe that LBP-TOP which does not take full advantage of the spatial information of cuboids just with LBP_{XY}, LBP_{XT}, LBP_{YT}. To make up this defect effectively, our method makes good use of the neighbourhood information of cuboids through three dimensions which can describe the distribution of grey-level space better. In addition, the introduced descriptor is 2×2^6 dimensions compared with other descriptors such as the HOG-HOF descriptor which is 162 dimensions.

2. 3D Gradient LBP Descriptor

This section outlines the framework of action recognition using 3D gradient LBP descriptor. The flowchart of action recognition framework is shown in Fig. 1. First, the STIPs are detected using a separable linear filter, and cuboids which represent the local spatial-temporal information are extracted from STIPs [8]. Next, the cuboids are described using the 3D gradient LBP descriptor. The obtained result is a sparse representation of video sequences. In addition, all these data are clustered into a set of visual words. Then, the histogram of these spatial-temporal word occurrence of video sequences can be computed. For classification, the non linear Support Vector Machines (SVM) or 1-Nearest Neighbor (1-NN) is applied separately. Finally, a testing video sequence is processed in similarly flow and finally classified to the best decision.

The traditional LBP operator [9] labels the pixels of an image by thresholding a circular neighborhood region. Under the underlying idea, the $\mathrm{LBP}_{P,R}$ generates 2^P different values on the radius of R, which represent the 2^P different binary patterns respectively. Mattivi and Shao [7] applied and extended LBP and LBP-TOP as a descriptor of small video patches used in local-feature approach for human action recognition, which shown LBP-TOP to be suitable for the description of the spatial-temporal cuboids. LBP-TOP computes the LBP from Three Orthogonal Planes, noted as LBP_{XY}, LBP_{XT}, LBP_{YT}.

The framework of the introduced descriptor is depicted in the Fig. 2. It is roughly divided into four steps to describe a cuboid: (1) dense sampling in each point of a cuboid to get 3D patches; (2) extracting the six planes of every 3D patch and comparing the average value of each plane with local feature in the threshold T; (3) assigning the two decimal values to represent the 3D patch; (4) recoding the two decimal values with histograms for the cuboid respectively and concatenating two histograms to describe every cuboid.

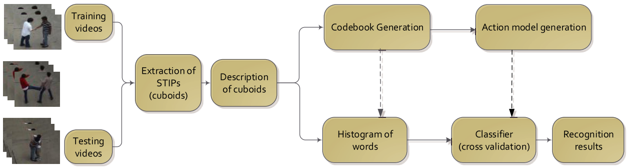


Fig. 1 Flowchart of our action recognition framework, which contains STIPs detection, action description, and classification.

Manuscript received January 10, 2017.
Manuscript revised February 14, 2017.
Manuscript publicized March 2, 2017.

[†]The authors are with Department of Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University, Shenzhen, China.
a) E-mail: guozhaoyang@pku.edu.cn

DOI: 10.1587/transinf.2017EDL8006

After the feature detection, we obtain a set of cuboids, which contain the spatial and temporal information. As the first step, the gradient vectors of the cuboids need to be computed efficiently. Next, dense sampling on every cuboid needs to be conducted to obtain 3×3×3 3D patches. In order to take advantage of the gray-level information of spatial adjacent frames, 3D patches are carved into six planes around central local feature V_0, denoted as $Plane_{front}$, $Plane_{rear}$, $Plane_{left}$, $Plane_{right}$ $Plane_{above}$ and $Plane_{below}$. The next step is to calculate the average pixel values of these six planes which represent the surrounding grey-level information. These above steps correspond to the Step 3 to Step 7 in Algorithm 1.

In order to use the relationship between central local features with around features more efficiently, LBP labels these six planes by thresholding in each 3D patch, as shown the following.

$$LBP_{Plane} = \begin{cases} 1 & \text{if } V_0 - Plane > T; \\ 0 & \text{if } \left|V_0 - Plane\right| \leq T; \\ -1 & \text{if } V_0 - Plane < -T. \end{cases} \quad (1)$$

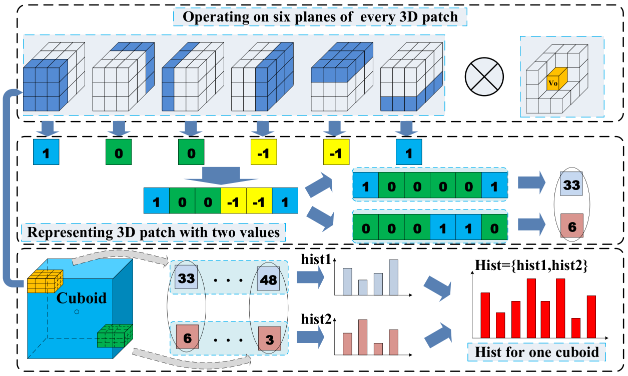


Fig. 2 Framework of 3D Gradient LBP descriptor, which contains dense sampling on a cuboid, representing 3D patches and concatenating histograms for the cuboid.

Algorithm 1 3D Gradient LBP Algorithm

```
Require: Video $V = \{I_t\}_{t=1}^{F}$, frame number $F$, cuboid number $K$, threshold
    value $T$.
Ensure: $Hist$
 1: Detect STIPs: $P = \{p(x,y,t)|(x,y) \in I_t, 1 \leq t \leq F\}$, local features
    $\{h_{p(x,y,t)}\}$;
 2: Calculate the gradient cuboids $g(x,y,t)$ from $p(x,y,t)$;
 3: for $i = 1$ to $K$ do
 4:     Dense sampling on cuboid $i$ to get $3 \times 3 \times 3$ 3D patches $\{Patch\}_m$;
 5:     for $j = 1$ to $m$ do
 6:         for $p = 1$ to 6 do
 7:             $Plane(1,p)$ records average value of six planes for $Patch_j$
                according to the order of front, rear, left, right, above and
                below ;
 8:             if $g(x,y,t) - Plane(1,p) > T$ then
 9:                 $hist(1,p) \leftarrow 1$; $hist(2,p) \leftarrow 0$;
10:             else if $|g(x,y,t) - Plane(1,p)| \leq T$ then
11:                 $hist(1,p) \leftarrow 0$; $hist(2,p) \leftarrow 0$;
12:             else if $g(x,y,t) - Plane(1,p) < -T$ then
13:                 $hist(1,p) \leftarrow 0$; $hist(2,p) \leftarrow 1$;
14:             end if
15:         end for
16:         $hist1_j = hist(1,1) \times 2^6 + ... + hist(1,6) \times 2^0$ ;
17:         $hist2_j = hist(2,1) \times 2^6 + ... + hist(2,6) \times 2^0$ ;
18:     end for
19:     $hist1$ records the histogram of $hist1_j$ $(j = 1 : m)$;
20:     $hist2$ records the histogram of $hist2_j$ $(j = 1 : m)$;
21: end for
22: $Hist \leftarrow \{hist1, hist2\}$.
```

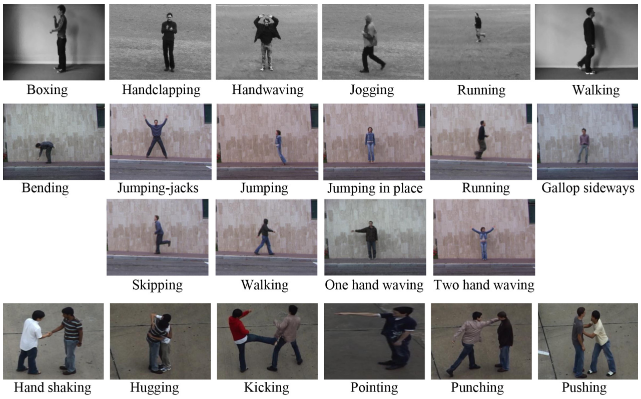
As a result, LBP vector for a 3D patch can be achieved such as $(1,0,0,-1,-1,1)$. However, the storage space for each 3D patch will be 3^6. In order to save storage and not miss any worthy information, the vector can be divided into two part as $(1,0,0,0,0,1)$ and $(0,0,0,1,1,0)$. In this case, the storage space for each 3D patch may be reduced as 2×2^6 effectively. In addition, these two vectors are converted as decimal values as 33 and 6. Next, the histogram statistic of these two sets of the decimal values for one cuboid is applied separately such as $hist1$ and $hist2$ in Fig. 2. Finally, two obtained histograms are concatenated to form the final histogram which described the spatial-temporal cuboid, namely that $Hist = \{hist1, hist2\}$. The part corresponds to the Step 8 to Step 20 in Algorithm 1.

3. Experiments and Analysis

In the following section, we explain more implementation details and parameters to verify the feasibility and superiority of our method. Furthermore, we demonstrate the application of our method to describe the spatial-temporal cuboids in human action recognition.

3.1 Datasets

Our work is verified on three challenging human action datasets: KTH dataset [10], Weizmann dataset [11] and UT-Interaction dataset [12]. KTH dataset [10] in Fig. 3 contains six types of different human action classes: working, jogging, running, boxing, hand waving and hand clapping. Each action class is performed several times by 25 subjects. The video data were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors


Boxing Handclapping Handwaving Jogging Running Walking

Bending Jumping-jacks Jumping Jumping in place Running Gallop sideways

Skipping Walking One hand waving Two hand waving

Hand shaking Hugging Kicking Pointing Punching Pushing

Fig. 3 Datasets: KTH dataset (on the top), Weizmann dataset (on the middle) and UT-interaction dataset (on the bottom).

with different clothes, and indoors. The background is homogeneous and static in most sequences. KTH dataset is divided into two parts: training set (16 peoples) and testing set (9 peoples) following the experimental setup in [10]. Furthermore, the Weizmann dataset [11] in Fig. 3 contains ten different types of human actions classes: bending downwards, running, walking, skipping, jumping-jack, jumping forward, jumping in place, galloping sideways, waving with two hands, and waving with one hand. Each action class is performed once or twice by 9 subjects, and it contains 93 video sequences in total. As suggested by Scovanner et al. [13], 8 subjects are used in training and the remaining subject of the video sequences are used in testing. The UT-Interaction dataset [12] in Fig. 3 contains 20 video sequences of continuous executions of 6 classes of human-human interactions: hand shaking, hugging, kicking, pointing, punching and pushing. This dataset consists of two sets: set 1 is composed of ten video sequences, which were conducted in a parking lot with slightly different zoom rates; set 2 is taken on a lawn in a windy day, where background is moving slightly. In the implements, one among ten video sequences is left in testing and the other nine are used in training.

3.2 Implementation Details

During the detection part, the method proposed by Dollar et al [8] is applied in our framework. The reason is that Dollar's detection method gives much better recognition accuracy compared with other existing methods, such as Laptev's method when only a small number of cuboids are extracted from the original video data. For detection, the parameters that need to be set are σ for the 2D Gaussian smoothing kernel of the response and τ for the quadrature pair of 1D Gabor filter in [8]. The parameter ω of the underline frequency of cosine in the Gabor filter is related to τ in the following relation $\omega = 4/\tau$. By repeated experiments, the parameters $\sigma = 2.8$ and $\tau = 1.6$ may be set to obtain the superior results. Each experiment is repeated 10 times with different random spits of the training and testing sets to obtain the average result, which is more reliable.

Our descriptor is compared with other existing descriptors in experiments. Gradient descriptor [8] is a concatenation of gradient values along the three dimensions. The size of the vector is equal to the number of pixels in the cube multiplied by the number of smoothing scales and the number of gradient directions. For instance, the cuboid are smoothed and the number of cuboid's dimension is three times of the cuboid's volume. 3D SIFT in [13] relied on the number of sub histograms and the number of bins which used to represent θ and φ angles. As suggested in [13], we used $2\times2\times2$ and $4\times4\times4$ configurations of sub-histograms, and 8×4 histograms to represent θ and φ. HOG-HOF descriptor [14] is a concatenation vector of HOG descriptor (72 vector length) and HOF descriptor (90 vector length). In our experiments, we implement the HOG-HOF referring to Laptev's codes [14]. And the 3D HOG descriptor was proposed by Klaser et al. [15]. It is a spatio-temporal descriptor of histograms of 3D gradient orientations and can be seen as an extension of the popular SIFT descriptor to video sequences. Heikkila et al. [16] proposed a center-symmetric LBP (CS-LBP) utilizing both the advantages of SIFT and LBP. This paper extracts the $CS - LBP_{2,8,0.01}$ and takes the uniform weight in 4×4 grid according to the experimental setup in [16].

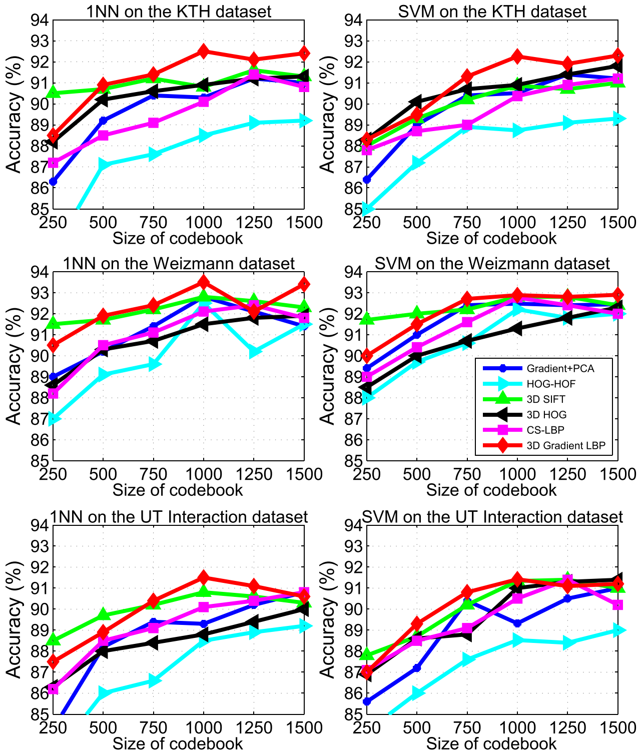


Fig. 4 Results of different descriptors with 1-NN and SVM on the three datasets (KTH, Weizmann and UT-Interaction).

The parameter T in 3D gradient LBP needs to be set in experiments. Abundant self-contrast experiments are conducted to get the superior value T. In this manner, it can achieve the best recognition results when T is set to 6. Moreover, two classifiers are utilized in our framework. First, a 1-Nearest Neighbor (1-NN) classifier with the χ^2 distance is used as suggested in [8]. In addition, non-linear SVM with rbf kernel [17] is applied.

3.3 Analysis

To validate the efficiency of the proposed descriptor, the same detection part may be set with different descriptor. Furthermore, to get a reference which can verify the correctness, the number of cuboids is fixed to 100 for all video sequences. In order to achieve a fair comparison with descriptors of the same length, descriptors are set in their original dimension as well as in lower dimension after PCA.

Experimental results of different descriptors with 1-NN and SVM on the datasets are depicted in Fig. 4. As the results show, the best performance is achieved with a code-

Table 1 Descriptor length, computational time and correctness (%) in 1000 visual words.

Methods	length	Time (s)	KTH	Weizmann	UT
Gradient+PCA	100	0.0060	90.51	92.48	89.32
HOG-HOF	768	0.0139	88.74	92.21	88.53
3D SIFT	640	1.1210	90.86	92.8	91.34
3D HOG	960	0.2256	90.9	91.3	91
CS-LBP	96	0.0115	90.37	92.76	90.5
DL-SFA	-	0.0342	93.1	-	-
HISTF	-	0.0134	93.9	-	-
STLPC	-	0.0251	**95**	-	-
Proposed	**128**	**0.0078**	92.25	**92.88**	**91.42**

book size ranging from 750 to 1250 visual words for the most descriptors. In addition, the results are quite similar when using 1-NN and SVM. However, the performance of the same descriptor are different with various classifiers.

In Table 1, descriptor length and computational time are shown for different descriptors of one cuboid. Among the descriptors, the fastest method is gradient descriptor [8], because the only operation need to be done is the concatenation of the gradient of pixels. 3D SIFT requires the most time among all descriptors, since histograms have to be conducted for different values of θ and Φ [13]. The computational time for HOG-HOF is influenced by the choice of the threshold, and a suitable threshold should be chosen [14]. However, competitive advantages of our method are achieved on descriptor length and computational time compared with other descriptors in Table 1. DL-SFA [18] applied directly to the whole video volume, which is very time-consuming because the video sequences usually have high resolution with a large number of frames. In [19], the authors presented an extension of the Independent Subspace Analysis algorithm to learn hierarchical invariant spatio-temporal features (HISTF) from unlabeled video data. [20] proposed STLPC method, which regards a video sequence as a whole with spatio-temporal features directly extracted from it, which prevents the loss of information in sparse representations while consumes a large amount of time compared with sparse representation methods based on detected local interest points.

To compare with different descriptors, the number of visual words is fixed to 1000, as shown in Table 1. For the majority of descriptors, the correctness on the Weizamann dataset are superior to the KTH dataset and the UT-Interaction dataset, since it is related with the complexity of dataset. On average of 1-NN and SVM, the performance of 3D Gradient LBP is the best on KTH and UT Interaction datasets. However, 3D SIFT has achieved the superior on the Weizmann dataset. It has shown that the proposed method is suitable for the datasets with more complex scenarios.

4. Conclusions

In this paper, we propose a novel local descriptor for human action recognition. In order to make full use of the neighbourhood information of cuboids from three dimensions, we develop the traditional descriptor LBP and extend the descriptor which focused on 2D images to 3D spatio-temporal case. Experiments show that proposed method can achieve the best result considering the accuracy and efficiency. In future work, the mid-level features will be explored to see if they can be combined with the local features aiming at a higher accuracies.

References

[1] H. Liu, H. Tang, W. Xiao, Z. Guo, L. Tian, and Y. Gao, “Sequential bag-of-words model for human action classification,” CAAI Transactions on Intelligence Technology, vol.1, no.2, pp.125–136, 2016.

[2] F. Baumann, A. Ehlers, B. Rosenhahn, and J. Liao, “Recognizing human actions using novel space-time volume binary patterns,” Neurocomputing, vol.173, no.P1, pp.54–63, 2016.

[3] D.-M. Tsai, W.-Y. Chiu, and M.-H. Lee, “Optical flow-motion history image (OF-MHI) for action recognition,” Signal, Image and Video Processing, vol.9, no.8, pp.1897–1906, 2015.

[4] L. Wang, Y. Qiao, X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4305–4314, 2015.

[5] X. Zhen and L. Shao, “Action recognition via spatio-temporal local features: A comprehensive study,” Image and Vision Computing, vol.50, no.C, pp.1–13, 2016.

[6] X. Peng, L. Wang, X. Wang, and Y. Qiao, “Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice,” Computer Vision and Image Understanding, vol.150, pp.109–125, 2016.

[7] R. Mattivi and L. Shao, “Human Action Recognition Using LBP-TOP as Sparse Spatio-Temporal Feature Descriptor,” Computer Analysis of Images and Patterns, International Conference, vol.5702, pp.740–747, 2009.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), pp.65–72, 2005.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol.24, no.7, pp.971–987, 2002.

[10] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach,” Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, pp.32–36, 2004.

[11] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” IEEE International Conference on Computer Vision (ICCV), pp.1395–1402, 2005.

[12] M.S. Ryoo and J.K. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” IEEE International Conference on Computer Vision (ICCV), pp.1593–1600, 2009.

[13] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” ACM International Conference on Multimedia (ACM MM), pp.357–360, 2007.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2008.

[15] A. Klaser, M. Marszalek, and C. Schmid, “A Spatio-Temporal Descriptor Based on 3D-Gradients,” British Machine Vision Conference, Leeds, pp.99.1–99.10, Sept. 2008.

[16] M. Heikkilä, M. Pietikäinen, and C. Schmid, “Description of interest regions with local binary patterns,” Pattern recognition, vol.43, no.3, pp.425–436, 2009.

[17] A. Sargano, P. Angelov, and Z. Habib, “Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines,” Applied Sciences, vol.6, no.10, pp.309–322, 2016.

[18] L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, “Dl-SFA: deeply-learned slow feature analysis for action recognition,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2625–2632, 2014.

[19] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” IEEE Conference on Computer Vision and Pattern Recognition, pp.3361–3368, 2011.

[20] L. Shao, X. Zhen, D. Tao, and X. Li, “Spatio-temporal laplacian pyramid coding for action recognition,” IEEE Transactions on Cybernetics, vol.44, no.6, pp.817–827, 2014.