Action Recognition Using Hybrid Feature Descriptor and VLAD Video Encoding

Dong Xing, Xianzhong Wang, Hongtao Lu

Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Department of Computer Science and Engineering
Shanghai Jiao Tong University, China

Abstract. Human action recognition in video has found widespread applications in many fields. However, this task is still facing many challenges due to the existence of intra-class diversity and inter-class overlaps among different action categories. The key trick of action recognition lies in the extraction of more comprehensive features to cover the action, as well as a compact and discriminative video encoding representation. Based on this observation, in this paper we propose a hybrid feature descriptor, which combines both static descriptor and motional descriptor to cover more action information inside video clips. We also adopt the usage of VLAD encoding method to encapsulate more structural information within the distribution of feature vectors. The recognition effects of our framework are evaluated on three benchmark datasets: KTH, Weizmann, and YouTube. The experimental results demonstrate that the hybrid descriptor, facilitated with VLAD encoding method, outperforms traditional descriptors by a large margin.

1 Introduction

The task of action recognition in video can be divided into five procedures: extracting Space-Time Interest Points (STIPs), describing STIPs, building visual words, encoding video clips and finally classifying action categories. Recent advances in action recognition show that the enhancement of feature description [6, 16–18] as well as video encoding methods [1, 2, 24–26] can significantly improve the correct recognition rate. This paper takes the advantage of both two approaches. A hybrid feature descriptor is built, which combines both the static and motional information inside each STIPs to represent local feature points. Then Vector of Locally Aggregated Descriptor (VLAD) [1, 2] encoding method is adopted, which is proved to be a compact and discriminative encoding method in image representation, to encode the distribution of high-dimensional feature vectors. Figure 1 illustrates the work flow of our action recognition framework.

While in general, motional feature descriptors perform better than static ones in action recognition [3], we strongly believe that motional and static features should be complementary to each other in realistic settings. For example, many two-player ball games, such as badminton and tennis, share similar motional features like waving rackets and jumping. It can be confusing to distinguish

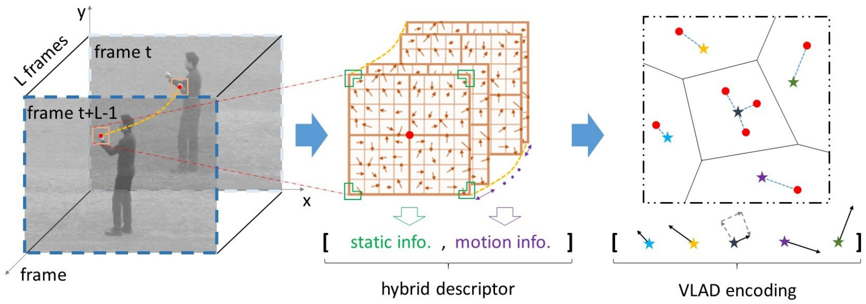


Fig. 1. An illustration of *STIPs extraction* (left), *hybrid feature description* (middle) and *VLAD video encoding* (right). Here we adopt dense trajectory sampling in [6] to extract STIPs. The hybrid descriptor, which is composed of two different descriptors, HOG (static information) and MBH (motional information), is used to cover the features around STIPs. Then, VLAD encoding is adopted to encapsulate the distribution of all hybrid descriptors.

these actions simply by their motional features. However, as we all know, our human vision system can easily recognize these sports even with a single static frame according to their appearance. Yet, for some actions like running and jogging which share similar static appearance, it can be hard to distinguish one from another simply by the static information. In these cases, the motional information is required to represent the features.

Based on the complementary idea mentioned above, two different types of feature descriptors are carefully chosen and combined to form our hybrid descriptor. One is histogram of oriented gradient (HOG) [4], which accumulates the static oriented gradient information inside each frame around the feature point; the other is motion boundary histogram (MBH) [5], which focuses on the dynamic motion boundary information within two or more neighbouring frames. Both of these two descriptors were originally used on the pedestrian detection [4, 5]. However, they also find their place in many other fields respectively. We compare the ability of our hybrid descriptor with the separated individual descriptors on several datasets, and the experimental result shows that our hybrid feature descriptor can achieve a state of the art recognition result.

The impact of different video encoding methods is also considered. Instead of traditional Bag of Words (BoW) encoding, VLAD [1, 2] is chosen to encode the distribution of feature vectors. The idea of VLAD is to aggregate all the differences between feature vectors and their corresponding visual words to form signatures, then concatenate all the signatures to construct the video representation. Although VLAD requires more dimensions than BoW to form the encoding vector of each video clip, the experimental result shows that VLAD encodes more details inside each video, and yields a better result than BoW even with a smaller codebook.

This paper is organized as follows. Section 2 talks about the related work on action recognition. In section 3, the details about hybrid descriptor, VLAD encoding as well as other implementation details in action recognition framework are explained. In section 4, we discuss our experimental result over several public datasets, including YouTube [7], Weizmann [8] and KTH [9]. Finally in section 5, we make a brief summary on action recognition.

2 Related Work

Successful extraction of more comprehensive features from video clips is the precondition for action recognition. Poppe [10] divides feature representations into two categories: global features and local features. Global feature extraction obtains an top-down fashion, which captures the information, such as silhouettes [11], edges [12], shapes [8], optic flows [13] of the human body as a whole, to form an overall representation. Although global features extract much of the action information, this kind of methods generally rely on accurate body localization, foreground extraction or tracking as preprocessing, and the result is sensitive to many environmental disturbance such as various viewpoints, lighting conditions and occlusions. On the other extreme, local feature extraction proceeds in a bottom-up fashion, which describes the observation as a combination of independent local portions. Comparing to global feature extraction, local feature extraction does not strictly rely on the effect of background subtraction or tracking, and is less sensitive to noise and partial occlusion. Due to these advantages, local feature extraction attracts more and more focus in the field of action recognition in recent years.

A wide range of local feature descriptors have been evaluated for action recognition. Based on different extraction methods, local feature descriptors can be divided into two groups: (i) motional feature descriptors, such as motion boundary histogram (MBH) [5] and histogram of optic flow (HOF) [14], which extract information from neighbouring frames through tracking the optic flows or other motional information around feature points; (ii) static feature descriptors, mostly originated from image processing, such as histogram of oriented gradient (HOG) [4] and 2D-SIFT [15], which regard the video clip as a sequence of frames and extract action information inside each frame respectively. Some methods extend the 2D image descriptors into 3D version, such as 3D-HOG [16], 3D-SIFT [17] and eSURF [18], by taking the temporal dimension as the third spatial axis to form a space-time video volume.

All these descriptors, static and motional, have various feature describing emphases, which offers a chance for us to evaluate combinations of different descriptors in order to cover more information about action characteristics. Several previous works [7, 19–21] have shown the effect of combining multiple features or visual cues. However, random combination of different descriptors does not always work. A descriptor of poor quality may drag down the effect of a descriptor of high quality, as our experiment result shows. How to align different descriptors to evoke the potentiality remains a problem, and our approach of combining static and motional descriptors offers a clue to solve this question.

Video encoding encapsulates the distribution of local feature descriptors. Bag of words (BoW) [22, 23] is one of the most popular encoding method, which assigns each feature vector to its nearest neighbouring visual word. The vector frequency of each visual word is accumulated, which is further concatenated directly as the video representation. Although BoW is proved to be a simple but valid encoding method, it omits lots of structural information inside the distribution of high-dimensional feature vectors, which is expected to have the ability of indicating the difference among each action classes to a large extent. Several novel encodings have been proposed to improve the BoW, including locality-constrained linear coding (LLC) [24], improved Fisher encoding [25], super vector encoding [26], VLAD [1, 2] and so on. Among all these encodings, VLAD maintains a simplicity of computational complexity as well as a quality of discrimination.

3 The Proposed Recognition Framework

This section explains in detail the formation of hybrid feature descriptor, the mechanism of VLAD encoding scheme as well as other implementation details in our action recognition framework.

The crux of action recognition lies in the procedure of feature extracting as well as video encoding. Feature extraction should extract features which are relevant to their corresponding action classes from video clips, and video encoding should encapsulate more of the action information inside each video clips into a compact and discriminative representation.

3.1 Hybrid Feature Descriptor

Two different descriptors, HOG [4] (static) and MBH [5] (motional), are combined to form our hybrid feature descriptor directly. Although the idea is simple, we found that this direct combination, facilitated with VLAD encoding, is capable of achieving an advanced recognition result without adding too much complexity.

The essential thought of HOG is to describe action appearance as the distribution of intensity gradients inside each localized portions of the video clip. Gradient values of each pixels inside the local portion are computed firstly frame by frame to describe the local patch appearance, then all the pixels inside each portion cast a weighted vote for an orientation-based histogram according to their amplitudes and orientations.

Unlike HOG or other static descriptors, MBH focuses on the motional information along the boundary of different depth of fields. Optic flows of neighbouring frames are computed first to indicate the motional information. Then a pair of x- and y-derivative differential flow images are obtained, on which large value indicates drastic motion changing. These two differential flow images cast the corresponding orientation-based histograms.

Several spatio-temporal grid combinations of size $n_\sigma \times n_\sigma \times n_\tau$ are evaluated to subdivide the local patch in order to embed the structural information of local portion descriptors. However, denser grid leads to descriptors with more dimensionality and extra computational burden, which should be taken into account when applying action recognition to more realistic situations. Here we set $n_\sigma = 2$ and $n_\tau = 3$ as in [6], which has shown to be the optimal choice on most cases in our experiments, meanwhile maintaining a moderate complexity. The orientations inside each grid are quantized into 8 bins, producing the final 96 dimension HOG descriptor and 192 dimension MBH descriptor.

Some papers [27, 7] also discuss the seamlessly combination over different features. Here we consider the dimensionality balance issue, caused by obvious dimension difference between HOG and MBH. We evaluate the usage of PCA to balance the dimension of different descriptors in order to even the impact of each descriptors. However, the size of million feature points makes PCA not feasible. Picking some dimensions randomly to equalize two descriptors is also tested, which works well in some cases, but the result is not stable and controllable. Therefore, we make a trade-off, and directly combine HOG and MBH to form the hybrid descriptors.

3.2 Video Encoding

VLAD [1] was firstly proposed in 2010 on the application of massive image searching. Unlike traditional BoW encoding, which requires a large size of codebook to achieve a good encoding effect, VLAD can achieve a better result even with a smaller codebook. Besides, VLAD representations are more discriminative than other encoding methods such as local linear constraint (LLC) [24], sparse coding based methods [28], etc.

The idea of VLAD is very simple. A codebook $D = \{\mu_1, \mu_2, ..., \mu_K\}$ of size K is learned using clustering methods (here we adopt k-means clustering). Then for each video clips, the differences between feature vectors and their belonging visual words are aggregated to form the signatures $\{v_1, v_2, ..., v_K\}$ of all visual words. The signature v_i is initialized with zero, and then being accumulated as equation 1 does:

$$v_i = \sum_{x_t:\mathrm{NN}(x_t)=i} x_t - \mu_i \tag{1}$$

where, $\mathrm{NN}(x_t)$ is a function indicating the index of visual words in the codebook D, which should be the nearest neighbour to x_t. The VLAD representation is then further normalized with power-low normalization [25] followed by L2-normalization.

3.3 Other Implementation Details

We adopt regular dense trajectory sampling of space-time features used in [6] to detect the STIPs inside each video clips. Wang *et al.* [3] has proved that dense trajectory sampling outperforms other commonly used feature detectors such as Harris3D [29], Cuboid [19] and Hessian [18] detectors in realistic settings. Meanwhile, dense sampling also maintains a simplicity to scale up the sampling density with a pre-computed dense optic flow fields.

For each sampling point, a list of predefined discrete spatial scale parameters have been covered to maintain the scale-invariant virtue. The trajectory neighbourhood is divided into a spatial-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$, then being described into a vector using our hybrid feature descriptor as well as other describing methods for comparison.

Once we derive all the feature vectors of each video clips in the dataset, we use k-means clustering over the whole feature vectors to quantize the standard of video representation. The centroids produced by k-means clustering is regarded as the visual words, which is further used in VLAD to form the encoding vector. Based on the encoding vector representing each video clips, action classification is finally performed with a one-vs-rest linear SVM classifier [30].

A whole pipeline of our algorithm is illustrated in Algorithm 1.

4 Experiments

In this section, we present a detailed analysis of our action recognition result based on the hybrid feature descriptor as well as VLAD encoding method on several datasets. We evaluate the performance among different descriptors to justify our choices. We also make a comparison between our results and the previously published works.

We choose three publicly available standard action datasets to report our recognition result, which are: YouTube [7], KTH [9] and Weizmann [8] datasets. Figure 2 shows some sample frames from these datasets. For each action classes, mean of average precision is calculated as performance measure. The experimental results show that our action recognition framework is competitive and can achieve a state of the art result.

4.1 Datasets

YouTube: The YouTube dataset [7] contains 1168 video clips from 11 action types, which are: basketball shooting (*B_Shooting*), biking, diving, golf swinging (*G_Swinging*), horse back riding (*H_Riding*), soccer juggling (*S_Juggling*), swinging, tennis swinging (*T_Swinging*), trampoline jumping (*T_Jumping*), volleyball spiking (*V_Spiking*), and walking with a dog (*Walking*). This is one of the challenging datasets due to its wild camera vibrations, cluttered background, viewpoint transitions and complicated recording conditions. Videos for each action type are wrapped into 25 groups, and each group contains four or more

Algorithm 1 Our Algorithm: Hybrid Feature Descriptor with VLAD Encoding

```
Input:
    $TrainVideo : \{a_1, a_2, ..., a_M\}$ is the set of training videos with size $M$;
    $TrainVideoLabel : \{l_1, l_2, ..., l_M\}$ is the set of action label of each training videos;
    $TestVideo : \{a_{M+1}, a_{M+2}, ..., a_{M+N}\}$ is the set of testing videos with size $N$;
Output:
    $TestVideoLabel : \{l_{M+1}, l_{M+2}, ..., l_{M+N}\}$ is the set of action label of each testing videos.
$X := \{X_1, X_2, ..., X_{M+N}\}$
for $d := 1$ to $M + N$ do
    $X_d := \{\}$
    $P := \{p_1, p_2, ..., p_S\}$ is the set of STIPs of video $a_d$ using dense sampling detector
    for $s := 1$ to $size(P)$ do
        $hogVector := HOG(ps)$
        $mbhVector := MBH(ps)$
        $hybridVector := [hogVector,\ mbhVector]$
        $X_d := \{X_d,\ hybridVector\}$
    end for
end for
$D : \{\mu_1, \mu_2, ..., \mu_K\} := kMeans(X, K)$
$V := \{V_1, V_2, ..., V_{M+N}\}$
for $d := 1$ to $M + N$ do
    $Y := X_d$
    for $k := 1$ to $K$ do
        $v_k := 0_d$
    end for
    for $t := 1$ to $size(Y)$ do
        $i := \arg\min_j Dist(Y_t, \mu_j)$
        $v_i := v_i + Y_t - \mu_i$
    end for
    $V_d := [v_1^T, v_2^T, ..., v_K^T]$
    for $k := 1$ to $K$ do
        $v_k := sign(v_k)\,|v_k|^\alpha$
    end for
    $V_d := V_d / \|V_d\|_2$
end for
$SVMClassifier := InitializeSVM(\{< V_1, l_1 >, < V_2, l_2 >, ..., < V_M, l_M >\})$
for $d := M + 1$ to $M + N$ do
    $l_d = SVMClassifier(V_d)$
end for
return $\{l_{M+1}, l_{M+2}, ..., l_{M+N}\}$;
```

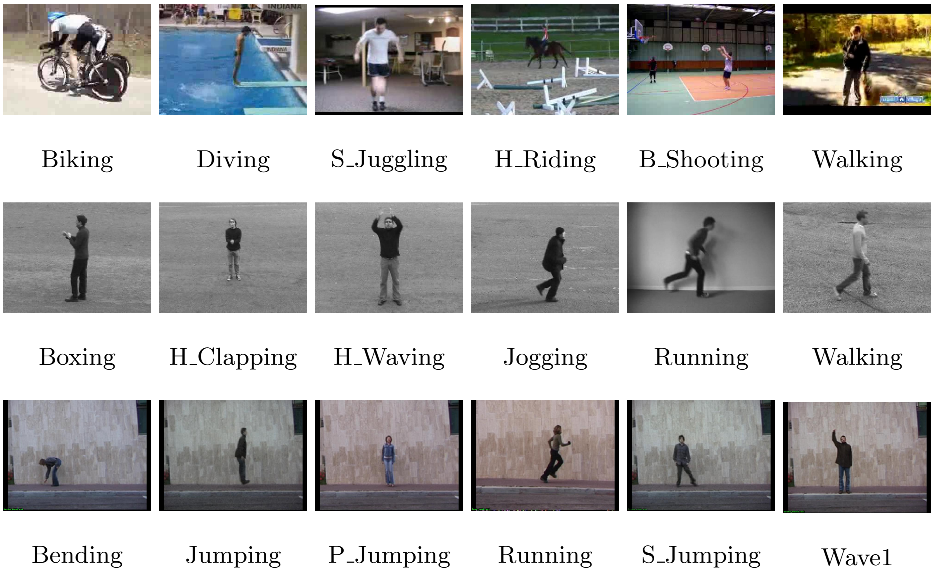


Fig. 2. Some samples from video sequences on *YouTube* (the first row), *KTH* (the second row) and *Weizmann* (the third row) datasets. Among all, YouTube dataset contains large variation, KTH dataset has homogeneous indoor and outdoor backgrounds, and Weizmann dataset records all videos on a static camera.

video clips sharing common features like same actor, similar background and viewpoint. We evaluate the classification accuracy by leave one out cross validation over the predefined 25 groups.

KTH: The KTH dataset [9] contains 600 video clips from six action types, which are: walking, jogging, running, boxing, hand waving and hand clapping. Each action type is performed by 25 persons in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. All sequences are taken over homogeneous backgrounds shot by a static camera, and have an average length of four seconds. We follow the split setup in [9], and choose 16 persons for training, and remaining 9 persons for testing.

Weizmann: The Weizmann dataset [8] contains 93 video clips from 10 action types, which are: bending, jumping jack (*J_Jump*), jumping forward (*F_Jump*), jumping in place (*P_Jump*), jumping sideways (*S_Jump*), skipping, running, walking, waving with two hands (*Wave2*), and waving with one hand (*Wave1*). Each action type is performed by 9 different persons. All the video clips are recorded in homogeneous outdoor background with a static camera, and have an average length of two seconds. We evaluate the classification accuracy by leave one out cross validation over 9 different persons repeatedly, and take the mean of average precision as the final correction rate.

4.2 Experiments on YouTube

We firstly decide the size of visual vocabulary generated by k-means clustering. A list of exponential increasing vocabulary sizes are evaluated over several types of descriptors, including optic Trajectories [6] (motional), HOG (static), MBH (motional) and their pairwise combinations. The result, as figure 3 illustrates, indicates that smaller vocabulary size generally leads to lower accuracy rate. However, picking a vocabulary size too large brings more computational burden, and since we choose dense sampling to find STIPs, the situation is even aggravated because the nearest visual word of all feature points should be required in VLAD. From figure 3 we observe that accuracy changing between 512 and 1024 is very limited, which gives us a chance to choose a size of 512 visual words over all descriptors so as to make a balance between recognition accuracy and computational efficiency.

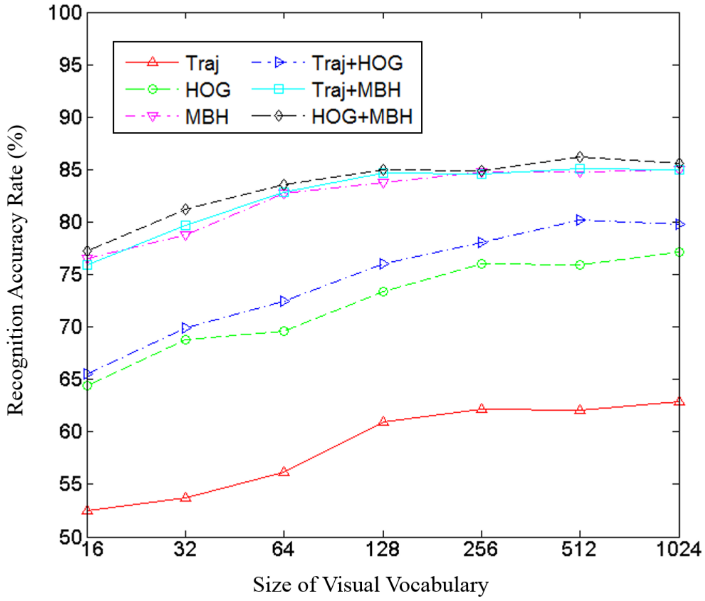


Fig. 3. Comparison among different sizes of visual vocabulary over several descriptors on YouTube dataset. Smaller vocabulary size leads to lower recognition accuracy rate, while larger vocabulary size brings more computational burden. Here we choose 512 visual words, which keeps a balance between recognition accuracy and computational efficiency.

A comparison of action classification effect among HOG, MBH and hybrid descriptors is performed to evaluate the improvement of hybrid descriptor over separated individual descriptors, and the result is shown in figure 4. Among these descriptors, the hybrid descriptor achieves a 86.23% recognition accuracy

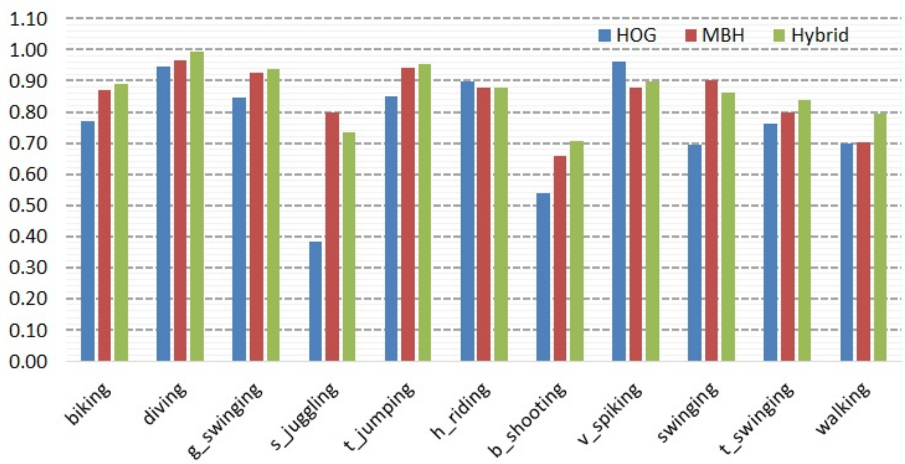


Fig. 4. Comparison of action classification performance over *HOG* (static), *MBH* (motional) and *Hybrid* descriptors on YouTube dataset. The average recognition precision rate for HOG, MBH and Hybrid descriptors are 75.95%, 84.73% and 86.23%, respectively.

rate, which is about 1.50% improvement over MBH, and 10.28% over HOG. Of all eleven actions in YouTube dataset, seven action classes gain improvement of recognition accuracy using hybrid descriptor. We also compare our recognition result to several previously published works [7, 31–35, 6] in Table 1, and the comparison shows our method obtains a state of the art result.

Table 1. Performance comparison between our method and some previously published works on YouTube dataset.

	Proposed	Liu *et al.* [7]	Zhang *et al.* [31]	Reddy *et al.* [32]
mAP	**86.2%**	71.2%	72.9%	73.2%
	Ikizler *et al.* [33]	Le *et al.* [34]	Brendel *et al.* [35]	Wang *et al.* [6]
mAP	75.2%	75.8%	77.8%	84.2%

4.3 Experiments on KTH and Weizmann

The strategy of combining static and motional descriptors is further evaluated on KTH dataset. Figure 5 shows the comparison between hybrid feature descriptors and the separated individual descriptors. We choose four different types of descriptors in [6] to make the combinations, which are HOG (static), TRAJ (motional), HOF (motional) and MBH (motional). Among all the combinations, the ones generated from static and motional descriptors (see row 1 in Figure 5) gain considerable improvement, while we can hardly find this improvement within the ones generated from two motional descriptors (see row 2 in Figure 5).

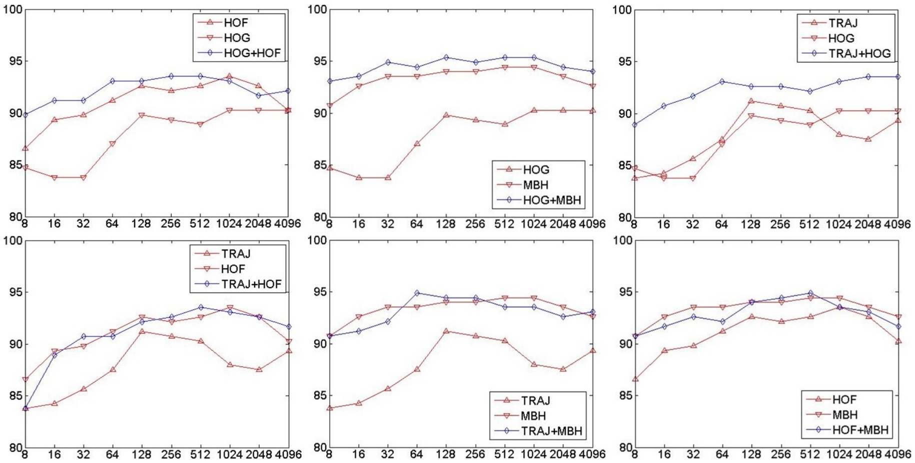
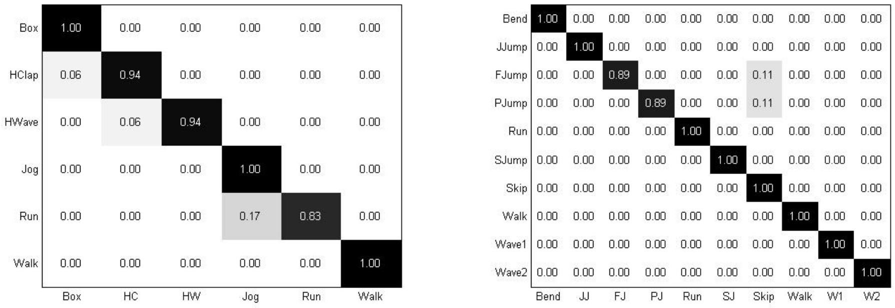


Fig. 5. Comparison between *hybrid feature descriptors* (blue line) and *the separated individual descriptors* (red lines). The x- and y-axis are the size of visual vocabulary and the recognition accuracy rate, respectively. Row 1 contains 3 different combinations of static and motional descriptors, and row 2 contains 3 combinations of two motional descriptors. We can observe that in row 1, hybrid descriptors gain considerable improvement, while in row 2, this improvement can hardly be found.

We perform action recognition on KTH and Weizmann respectively using hybrid feature descriptor and VLAD encoding method, and achieves a recognition accuracy rate of 95.4% on KTH and 97.8% on Weizmann. Figure 6 shows the confusion matrix of our recognition result in these two datasets. From figure 6 we can see that errors in KTH are mostly caused by mislabelling running to jogging, and errors in Weizmann are caused by mislabelling jumping forward and jumping in place to skipping. If the ”skip” action class is expelled from Weizmann dataset, the recognition accuracy rate can be further raised to 100.0%.

Table 2 shows a comparison between our proposal and several previous published works [21, 36–41]. From table 2 we can see that our proposed method achieves a state of the art result on KTH dataset, and maintains a competitive recognition result on Weizmann dataset.



	Box	HC	HW	Jog	Run	Walk
Box	1.00	0.00	0.00	0.00	0.00	0.00
HClap	0.06	0.94	0.00	0.00	0.00	0.00
HWave	0.00	0.06	0.94	0.00	0.00	0.00
Jog	0.00	0.00	0.00	1.00	0.00	0.00
Run	0.00	0.00	0.00	0.17	0.83	0.00
Walk	0.00	0.00	0.00	0.00	0.00	1.00

(a) KTH

	Bend	JJ	FJ	PJ	Run	SJ	Skip	Walk	W1	W2
Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
JJump	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FJump	0.00	0.00	0.89	0.00	0.00	0.00	0.11	0.00	0.00	0.00
PJump	0.00	0.00	0.00	0.89	0.00	0.00	0.11	0.00	0.00	0.00
Run	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
SJump	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Skip	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b) Weizmann

Fig. 6. Confusion matrices for KTH and Weizmann datasets using hybrid feature descriptor and VLAD encoding.

Table 2. Performance comparison between our method and some previously published works on KTH and Weizmann dataset.

	KTH	Weizmann		KTH	Weizmann
Proposed	**95.4%**	**97.8%**	Cao *et al.* [36]	93.5%	94.6%
Grundmann *et al.* [37]	93.5%	94.6%	Fathi *et al.* [38]	90.5%	100.0%
Lin *et al.* [21]	93.4%	100.0%	Schindler *et al.* [39]	92.7%	100.0%
Cai *et al.* [40]	94.2%	98.2%	Liu *et al.* [41]	94.8%	100.0%

5 Conclusion

This paper presents a hybrid descriptor to describe local features extracted from video clips, which takes the advantage of both static and motional information to cover more details inside neighbourhood of spatial-temporal interest points. Besides, VLAD encoding is adopted for each video clips to encapsulate more structural information on the distribution of feature vectors. We evaluate the effect of our action recognition framework over several datasets, and the experimental result shows the usage of hybrid feature descriptor as well as VLAD encoding can significantly improve the average recognition accuracy.

Acknowledgement. This work is supported by NSFC (No.61272247 and 60873133), the Science and Technology Commission of Shanghai Municipality (Grant No.13511500200), 863 (No.2008AA02Z310) in China and the European Union Seventh Frame work Programme (Grant No.247619).

References

1. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3304–3311
2. Arandjelovic, R., Zisserman, A.: All about vlad. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 1578–1585
3. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., et al.: Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British Machine Vision Conference. (2009)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., IEEE (2005) 886–893
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Computer Vision–ECCV 2006. Springer (2006) 428–441
6. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3169–3176
7. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1996–2003
8. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29** (2007) 2247–2253
9. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Volume 3., IEEE (2004) 32–36
10. Poppe, R.: A survey on vision-based human action recognition. Image and vision computing **28** (2010) 976–990
11. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23** (2001) 257–267
12. Carlsson, S., Sullivan, J.: Action recognition by shape matching to key frames. In: Workshop on Models versus Exemplars in Computer Vision. Volume 1. (2001) 18
13. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE (2003) 726–733
14. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
15. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Volume 2., Ieee (1999) 1150–1157
16. Klaser, A., Marszalek, M.: A spatio-temporal descriptor based on 3d-gradients. (2008)
17. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th international conference on Multimedia, ACM (2007) 357–360
18. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Computer Vision–ECCV 2008. Springer (2008) 650–663

19. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, IEEE (2005) 65–72
20. Laptev, I., Pérez, P.: Retrieving actions in movies. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE (2007) 1–8
21. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 444–451
22. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
23. Liu, J., Shah, M.: Learning human actions via information maximization. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
24. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3360–3367
25. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Computer Vision–ECCV 2010. Springer (2010) 143–156
26. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: Computer Vision–ECCV 2010. Springer (2010) 141–154
27. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2046–2053
28. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1794–1801
29. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64** (2005) 107–123
30. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011) 27
31. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: Computer Vision–ECCV 2012. Springer (2012) 707–721
32. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Machine Vision and Applications **24** (2013) 971–981
33. Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition. In: Computer Vision–ECCV 2010. Springer (2010) 494–507
34. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3361–3368
35. Brendel, W., Todorovic, S.: Activities as time series of human postures. In: Computer Vision–ECCV 2010. Springer (2010) 721–734
36. Cao, X., Zhang, H., Deng, C., Liu, Q., Liu, H.: Action recognition using 3d daisy descriptor. Machine Vision and Applications **25** (2014) 159–171
37. Grundmann, M., Meier, F., Essa, I.: 3d shape context and distance transform for action recognition. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE (2008) 1–4

38. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
39. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
40. Cai, Q., Yin, Y., Man, H.: Learning spatio-temporal dependencies for action recognition, ICIP (2013)
41. Liu, L., Shao, L., Zhen, X., Li, X.: Learning discriminative key poses for action recognition. (2013)