A New Dataset and Evaluation for Infrared Action Recognition

Chenqiang Gao[1](✉), Yinhe Du[1], Jiang Liu[1], Luyu Yang[1], and Deyu Meng[2]

[1] Chongqing Key Laboratory of Signal and Information Processing,
Chongqing University of Posts and Telecommunications, Chongqing, China
gaocq@cqupt.edu.cn

[2] Institute for Information and System Sciences
and Ministry of Education Key Lab of Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, China

Abstract. Action recognition (AR) is one of the most important tasks in computer vision and there are a large number of related research works along this line. While most of these works are investigated on AR datasets collected from the visible spectrum, the AR problem on infrared scenarios still has not attracted much attention, and there is even few public infrared datasets available for supporting this research. This study aims to emphasize the importance of the infrared AR problem in real applications and arouse researchers' attention on this task. Specifically, we construct a new infrared action dataset and evaluate the state-of-the-art AR pipeline, including widely-used low-level local descriptors, coding methods and fusion strategies, on it. Through these evaluations, we find some interesting results. E.g., dense trajectory feature can achieve the best performance while the appearance features, e.g., HOG, has relatively poorer performance; the coding method of vector of locally aggregated descriptors is evidently better than that of the widely-used fisher vector; the late fusion facilitates a better performance than early fusion. Furthermore, the best performance achieved on our dataset is 70%, leaving a relative large space for promoting new methods on this infrared AR task.

Keywords: Infrared action dataset · Action recognition · Local descriptors · Feature fusion

1 Introduction

Action recognition (AR) is one of the most important tasks in computer vision. Its potential applications include video surveillance, video indexing, human-computer interaction (HCI), etc. [1]. Over the past decades, human action recognition has attracted extensive attention and a number of methods have been proposed to address this task [24]. Basically, most of the efforts have been put into visible imaging videos and many existing methods follow the pipeline: raw feature extraction, feature coding and classifier learning. Generally speaking, the description ability of the adopted features is very important to the performance

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 302–312, 2015.
DOI: 10.1007/978-3-662-48570-5_30

of the method. So far, many good feature descriptors have been widely used for action recognition, such as STIP [18], HOG3D [14], 3DSIFT [23], etc.

The development of feature descriptors needs to be refined and substantiated on proper AR datasets. Recently, many AR datasets have been constructed to research purposes, such as KTH [22], UCF sports [26], HMDB51 [16],WEB-interaction [8], etc. The recently proposed AR datasets [15] more and more simulate real scenarios. While benefited from these datasets, recently designed methods for AR can better adapt real applications, these methods still often encounter great challenges, such as illumination change, shadow, background clutter, occlusion of the object, etc. Actually, these challenges also make other computer vision tasks, like object detection, very hard to be effective only based on the provided visible information.

Compared to visible spectrum imaging, the infrared thermal imaging have many complementary advantages over the aforementioned challenges [10], e.g., the infrared imaging is able to work well under poor light condition, like imaging at night. These advantages have been utilized in pedestrian detection [30], face recognition [13] and other computer vision tasks, but still have not been attracted much attention in AR community [9]. Especially,to the best of our knowledge, there is still no public dataset available for AR research purpose so far.

To the aforementioned issues, we set up a new infrared dataset, called infrared action dataset (IAD), for the infrared AR task. Following the approach of construction of existing AR datasets in visible spectrum [3], the new dataset collects 12 kinds of common human actions. The samples vary from simple to complex scenes. We further evaluate the state-of-the-art AR pipeline on our dataset. Specifically, our evaluation emphasis is put on widely-used low-level local descriptors, the coding strategies and the fusion strategies. This work is expected to establish a benchmark and baseline for infrared AR research, like KTH dataset for AR of visible spectrum.

The rest of this paper is organized as follows: Section 2 introduces details of the newly constructed dataset. Section 3 introduces the employed local descriptors and the utilized evaluation methods. Section 4 presents experimental setup and evaluation results on the dataset. The conclusion is drawn in Section 5.

2 Infrared Action Dataset(IAD)

Following the approach to construct a AR dataset from the visible spectrum [3], we collect 12 common human actions from infrared videos. As shown in Fig. 1, the action types include one hand wave(wave1), multiple hands wave(wave2), handclapping, walk, jog, jump, skip, handshake, hug, push, punch and fight. Each action type has 30 video clips. All actions are performed by 25 different volunteers. The videos are captured by a handled infrared camera IR300A. Each clip lasts 4 seconds in average. The frame rate is 25 frames per second and the resolution is 293×256. Each video contains one person or several persons performing one action or more actions. Some of them are interactions between multiple persons. Table 1 lists the detailed information of our IAD and some known existing visible AR datasets.

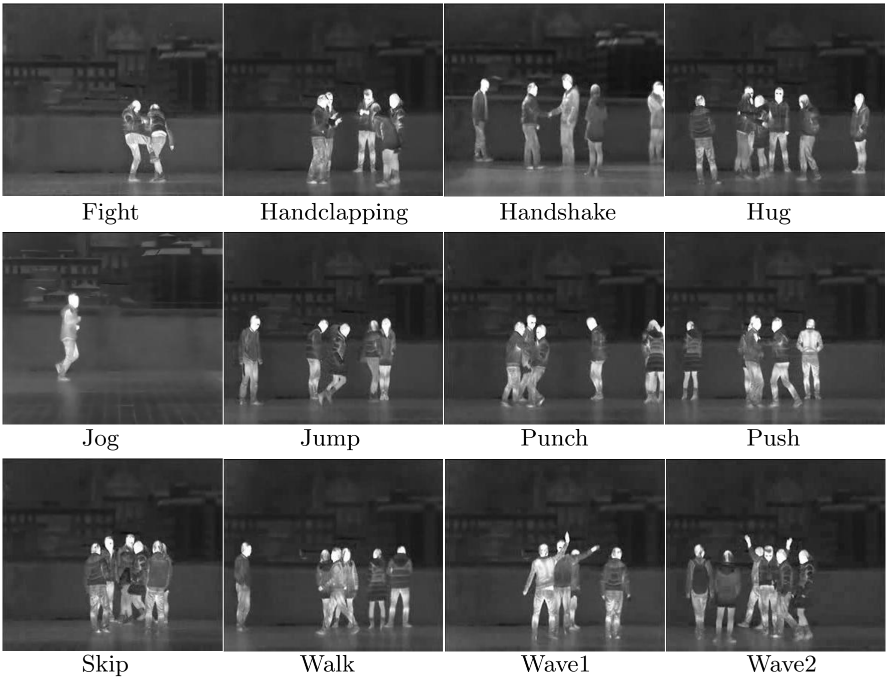
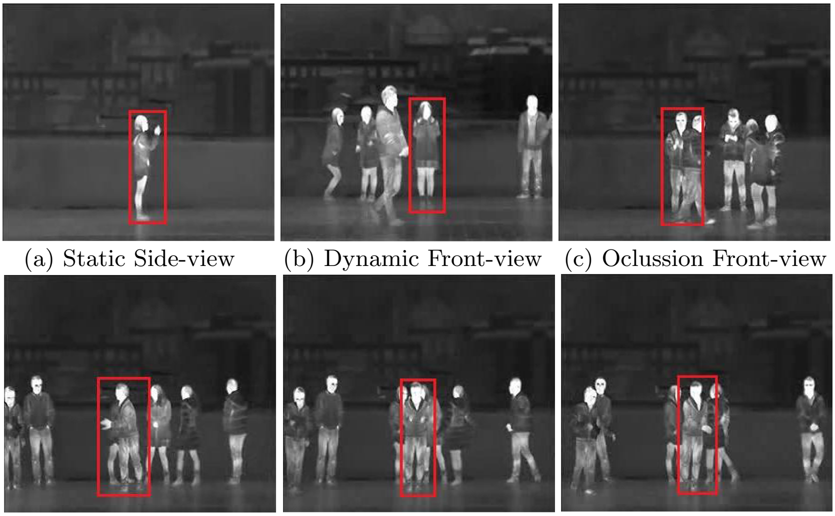


Fig. 1. 12 actions of the newly constructed infrared AR dataset.

Table 1. Comparison of existing AR datasets and the new IAD dataset.

	KTH	IXMAS	UCF sports	HMDB51	IAD
Video clips	600	2236	156	(min)5151	360
Action Class	6	13	13	51	12
Resolution	160×120	390×291	480×360	320×240	293×256
Frame Rate	25	25	25	30	25
Average Length(s)	4	3	3	3	4
Data Type	visible	visible	visible	visible	infrared
Reference	[29]	[28]	[21]	[29]	-

In order to make our dataset more representative for real-world varying scenarios, we consider four intra-class variations: **(a)** The background varies from simple scene (clean background) to complex one (including real-life background with moving humans). For some clips with simple background, there are only the person performing actions with clean background, as shown in Fig. 2(a). On the contrary, for some other clips with complex background, there are interrupting pedestrian activities concurring with the action, as shown in Fig. 2(b)-(f). **(b)** We specify 2-3 video clips with over 50% occlusion in each class, as shown in Fig. 2(c). **(c)** The pose variation is considered even for the same person, as shown in Fig. 2(d)-(f). **(d)** The viewpoint variation is also considered. Around


(a) Static Side-view (b) Dynamic Front-view (c) Oclussion Front-view

(d) Posture1 (e) Posture2 (f) Posture3

Fig. 2. Examples of intra-class variations of the handclapping action. (a) the ideal case with side-view angle and single person. (b) and (c) two cases with dynamic and occlusion variations in the background. (d), (e) and (f) cases with different postures.

20 video clips are taken under the front-view, and the remaining are taken under side-view, as shown in Fig. 2(a) and (b)-(c).

3 Evaluation Pipeline

3.1 Local Descriptors

Seven widely-employed low-level local descriptors are extracted from infrared video, including STIP [18], HOG3D [14], 3DSIFT [23], trajectory feature TRAJ [27], appearance feature HOG [4], and motion features HOF [19] and MBH [5]. Combination of TRAJ, HOG, HOF and MBH forms the dense trajectory feature [27], denoted as Dense-traj. In order to further introduce our evaluation settings, we briefly review these descriptors below.

STIP: The spatio-temporal interest points (STIP) is proposed by Laptev et al. [18] based on the idea from the Harris and Forstner interest point operators [11], which is widely used as a video representation to handle videos with complex and dynamic background recently [31]. As actions often have characteristic extending both in spatial and temporal domain, Laptev el al. extended the notion of interest points into the spatio-temporal domain and adapted both spatial and temporal scales of the detected features. In our experiment, we use the off-the-shelf binary package available online to extract this feature.

HOG3D: This feature is the local descriptor proposed by Klaser et al. [14], which is based on histograms of oriented 3D spatio-temporal gradients. It computes 3D gradient from an integral video representation. Then regular polyhedrons are used to quantize orientation of 3D gradient. The author divided a 3D patch from videos into $n_x \times n_y \times n_t$. The corresponding descriptor concatenates gradient histograms of all cells and is then normalized. In our paper, we firstly detect interest points with Harris corner detector, and then represent them with the HOG3D descriptor. We use executable programs from the author website and apply their recommend parameter setting as: $n_x = n_y = 4, n_t = 3$, where n_x, n_y and n_t are numbers of spatial and temporal cells, respectively.

3DSIFT: The 3 Dimensional Scale-Invariance Feature Transform (3DSIFT) descriptor was proposed by Scovanner el at. [23] which encodes gradient characteristics in 3 dimensional space. First the gradient magnitude and orientations in 3D are computed, and then A sub-histogram is created by sampling subregions surrounding the interest points. For each sub-region the orientations are accumulated into sub-histogram. The final descriptor is a vectorization of the sub-histogram. Here we detect interest points using the Harris Corner detector. We use the publicly available code from the Scovanner's website with the suggested parameter settings.

TRAJ: Want et al. [27] put forward a trajectory descriptor which encodes local motion of the densely-sampled interest points. The trajectory shapre is described by the sequence of the relative motion between every two consecutive frames, and the feature points are sampled on the trajectory with a fixed number of frames. This feature can well capture the motion characteristics of the video, which is significant in action recognition.

HOG/HOF: The Histogram of Gradients (HOG) and the Histogram of Optical flow (HOF) descriptors are introduced by Laptev et al. [19]. The authors compute histograms of spatial gradient and optical flow accumulated in space-time neighborhoods of detected points, which can be detected using any interest point detectors [7,18]. In our experiment, these points are selected along the motion trajectory [27] and features are computed within a $N \times N$ volume around these points. Each volume is subdivided into a space-time grid of size $n_\sigma \times n_\sigma \times n_T$. The default parameters for our experiments are $N = 32, n_\sigma = 2, n_T = 3$.

MBH: The Motion Boundary Histogram (MBH) descriptor is proposed in the work of Dalal et al. [5], where derivatives are computed separately for the horizontal and vertical components of the optical flow. The descriptor encodes the relative motion between pixels. Since MBH represents the gradient of the optical flow, constant motion information is suppressed and only the information concerning changes in the flow field (i.e., motion boundaries) is kept. This descriptor yields good performance when combined with other local descriptors. In our evaluation, we used the same MBH parameters as used in the work of Wang et al. [27].

3.2 Feature Encoding Methods

In this paper, two encoding methods, namely the Fisher Vector [20] and the Vector of Locally Aggregated Descriptors (VLAD) [12], are used. The former utilizes Maximum Likelihood (ML) estimation to train a Gaussian mixture model (GMM), which is later employed to form the description of low level features. The latter, however, utilizes the k-means technique to assign each feature to the closest cluster of a vocabulary with size k.

3.3 Fusion Strategy

At present, early-fusion and late-fusion are the basic feature fusion strategies. Early fusion [25] combines multiple features before classification. In our work, the concatenation of multiple features is employed since it is a commonly-used way in early fusion. Late fusion [17] requires more computation, since it combines the outputs of each type of feature. In our work, the average of the output scores is adopted for late-fusion.

4 Experiments

In this section, we first describe the implementation details, and then present the evaluation results of low-level local descriptors, including the encoding strategies on our IAD dataset. Besides, the fusion strategies are also evaluated.

4.1 Implementation Details

In our experiments, we follow the widely-used pipeline of raw feature extraction, feature encoding and classifier training in general AR systems. Basically, the raw feature extraction adopted the off-the-shelf coding and the default parameter configure as aforementioned. For the Fisher Vector, the number K of the Gaussian distributions model is an important parameter. We tested many values and empirically found that $K = 90$ can get relative better performance. For the VLAD, the size K of the codebook is also empirically determined as 500. We adopted the SVM [6] as the classifier and the libSVM [2] software in our experiments. We tested two kernels of Linear kernel and RBF kernel for two coding methods. The corresponding optimal parameters C and γ are obtained using 5 fold cross validations with a grid searching algorithm. Using the Fisher Vector and VLAD encoding methods, the searching results are as follows: For the linear kernel the optimal C is 30 and 80, respectively, and for the RBF kernel, the optimal C is 50 and 8, and γ is 2.7×10^{-3} and 5.0×10^{-1}, respectively.

4.2 Evaluations on Low-Level Local Descriptors

We evaluate 8 local feature descriptors as aforementioned with respect to different coding methods and different classifier kernels. For each evaluation, we randomly select 20 samples as the training set out of a sample set containing 30 samples and the rest 10 samples are used as the test set. We conduct the experiments of the same settings five times and the average is used as the final result. All evaluation results are shown in Table 2.

Table 2. The average precision (%) of different local descriptors with different kernels and coding methods.

Descriptor	Fisher Vector		VLAD	
	Linear	RBF	Linear	RBF
TRAJ	55.74	51.66	62.5	60.49
Dense-Traj	**68.15**	65.83	**74.83**	72
HOG	48.61	43.66	52.5	50.83
HOF	66.94	65.5	69.16	69.16
MBH	64.53	64.16	70	68.83
STIP	62.66	45.66	61.83	58.16
HOG3D	57.16	37.83	56.66	56
3DSIFT	53.66	20.33	58.33	54.16

From Table 2, we can observe that the best performance is from the Dense-traj [27]. This is basically in accordance with the situation on some other available datasets of visible spectrum [3]. Overall, the performance of the HOG is relatively bad among these descriptors. The reason may be caused by the lack of local texture information in infrared images (please see Fig. 2). Since the HOG descriptor is good at appearance description, its strength may not be revealed in the situations where local texture is relatively weak.

It can be also observed that the performance of linear kernel is much better than the RBF one, especially for those descriptors with higher feature dimensions (e.g., HOG3D, 3DSIFT) under the Fisher vector coding strategy. One possible reason is that the RBF kernel causes over-fitting in our task. It is also interesting to see that the performance of VLAD is better than Fisher Vector. This may be the fact that the codewords generated by HOG3D, MBH, etc. are not suitable to be described as the GMM model.

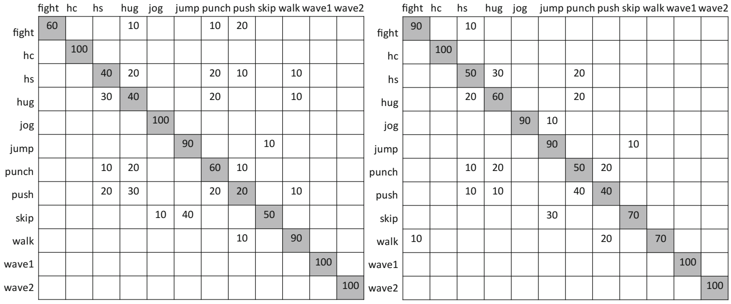
4.3 Early Fusion and Late Fusion

The early fusion and late fusion strategies are evaluated on the IAD dataset. We mainly consider STIP, TRAJ, HOG, HOF and MBH descriptors since they are of different and complementary types. The combinations of different numbers of features are tested, respectively, and the results are shown in Table 3. We can observe that the late fusion benefits more to the overall performance than the early fusion. Besides, the number of features for fusion does not determinate the final performance. In our case, the best performance for both fusion strategies is obtained when using STIP, HOF and MBH.

In order to further explore the classification performance for each action, we illustrate two confusion matrices shown in Fig. 3. The left one is the result of early fusion of STIP and HOF, and the right one is the result of late fusion of STIP and MBH. It can be seen that four actions of handshaking, hugging, punching and pushing have relative lower precisions. These actions are easily confused with other actions, e.g., handshaking and hugging, punching and pushing, pushing and hugging, etc. Fig. 4 shows two pairs of frames from four action videos. From the left pair

Table 3. The evaluation results (AP) of Early vs. Late Fusion with the same coding method of fisher vector.

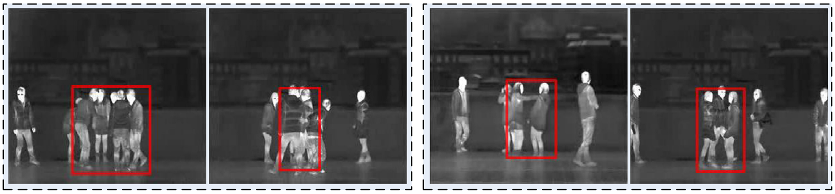
Fusion Type	Descriptor	Early Fusion	Late Fusion
Two features fusion	STIP+TRAJ	63.33	62.5
	STIP+HOG	58.33	61.67
	STIP+HOF	70.83	74.17
	STIP+MBH	69.16	75.83
Three features fusion	STIP+TRAJ+HOG	63.33	65
	STIP+HOG+HOF	70	72.5
	STIP+HOF+MBH	78.33	77.50
Four feature fusion	STIP+TRAJ+HOG+HOF	70.83	71.67
	STIP+HOG+HOF+MBH	72.5	72.5
Five feature fusion	STIP+TRAJ+HOG+HOF+MBH	73.33	72.5



	fight	hc	hs	hug	jog	jump	punch	push	skip	walk	wave1	wave2
fight	60			10			10	20				
hc		100										
hs			40	20			20	10		10		
hug			30	40			20			10		
jog					100							
jump						90			10			
punch			10	20			60	10				
push			20	30			20	20		10		
skip					10	40			50			
walk								10		90		
wave1											100	
wave2												100

	fight	hc	hs	hug	jog	jump	punch	push	skip	walk	wave1	wave2
fight	90		10									
hc		100										
hs			50	30			20					
hug			20	60			20					
jog					90	10						
jump						90			10			
punch			10	20			50	20				
push			10	10			40	40				
skip						30			70			
walk	10							20		70		
wave1											100	
wave2												100

Fig. 3. The comparative results of two fusion strategies, where the left is from early fusion strategy of STIP and HOF, while the right is from the late fusion strategy of STIP and MBH. Note that "hc" stands for "handclapping", "hs" stands for "handshake".



handshaking hugging punching pushing

Fig. 4. Two pairs of easily confused actions. The left shows two actions of handshaking and hugging with heavy occlusion, while the right shows two similar actions of punching and pushing.

of frames, we can see that the handshaking and hugging actions are both occluded by crowded persons around. These background clutter would bring big confusion. From the right pair, we can see that punching and pushing are so similar that it may even be deceitful for human eyes to recognize.

5 Conclusion

In this paper, we introduce a new infrared action dataset and evaluate the state-of-the-art AR pipeline on it. The evaluation results reveal that the dense trajectory feature can achieve the best performance on our dataset and the appearance features have relative poorer performance. Besides, the coding method of vector of locally aggregated descriptors is better than the widely-used fisher vector, and the late fusion benefits more to performance than early fusion. In addition, the best average precision on our infrared action dataset is around 70%, which leaves sufficient space for promoting new infrared-oriented AR methods.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61102131, 61571071), the Natural Science Foundation of Chongqing Science and Technology Commission (No. cstc2014jcyjA40048), Cooperation of Industry, Education and Academy of Chongqing University of Posts and Telecommunications (No. WF201404).

References

1. Aggarwal, J., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) **43**(3), 16 (2011)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm
3. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding **117**(6), 633–659 (2013)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
6. Dikmen, M., Ning, H., Lin, D.J., Cao, L., Le, V., Tsai, S.F., Lin, K.H., Li, Z., Yang, J., Huang, T.S., et al.: Surveillance event detection. In: TRECVID (2008)
7. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE (2005)
8. Gao, C., Yang, L., Du, Y., Feng, Z., Liu, J.: From constrained to unconstrained datasets: an evaluation of local action descriptors and fusion strategies for interaction recognition. In: World Wide Web, pp. 1–12 (2015)
9. Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, CVPR Workshops 2005, p. 17. IEEE (2005)
10. Han, J., Bhanu, B.: Fusion of color and infrared video for moving human detection. Pattern Recognition **40**(6), 1771–1784 (2007)

11. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, UK, vol. 15, p. 50 (1988)
12. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3304–3311. IEEE (2010)
13. Klare, B.F., Jain, A.K.: Heterogeneous face recognition using kernel prototype similarities. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(6), 1410–1422 (2013)
14. Klaser, A., Marszalek, M.: A spatio-temporal descriptor based on 3d-gradients (2008)
15. Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T.: Hmdb51: A large video database for human motion recognition. In: High Performance Computing in Science and Engineering 2012, pp. 571–582. Springer (2013)
16. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2556–2563. IEEE (2011)
17. Lan, Z., Bao, L., Yu, S.-I., Liu, W., Hauptmann, A.G.: Double fusion for multimedia event detection. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 173–185. Springer, Heidelberg (2012)
18. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2–3), 107–123 (2005)
19. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
20. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
21. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1234–1241. IEEE (2012)
22. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, pp. 32–36. IEEE (2004)
23. 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, pp. 357–360. ACM (2007)
24. Shao, L., Zhen, X., Tao, D., Li, X.: Spatio-temporal laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics **44**(6), 817–827 (2014)
25. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402. ACM (2005)
26. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
27. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2011)
28. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1), 60–79 (2013)

29. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3551–3558. IEEE (2013)
30. Wang, J.T., Chen, D.B., Chen, H.Y., Yang, J.Y.: On pedestrian detection and tracking in infrared videos. Pattern Recognition Letters **33**(6), 775–785 (2012)
31. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2834–2841. IEEE (2013)