Nazir, S., Yousaf, M.H. y Velastin, S.A. (2018). Feature Similarity and Frequency-Based Weighted Visual Words Codebook Learning Scheme for Human Action Recognition. In *PSIVT 2017 Image and Video Technology*,10749, pp 326-336.

DOI: https://doi.org/10.1007/978-3-319-75786-5_27

Feature Similarity and Frequency-Based Weighted Visual Words Codebook Learning Scheme for Human Action Recognition

Saima Nazir[1], Muhammad Haroon Yousaf[1](✉), and Sergio A. Velastin[2]

[1] University of Engineering and Technology Taxila, Taxila, Pakistan
saima_nazir_91@yahoo.com, haroon.yousaf@uettaxila.edu.pk
[2] Universidad Carlos III de Madrid, Madrid, Spain
sergio.velastin@ieee.org

Abstract. Human action recognition has become a popular field for computer vision researchers in the recent decade. This paper presents a human action recognition scheme based on a textual information concept inspired by document retrieval systems. Videos are represented using a commonly used local feature representation. In addition, we formulate a new weighted class specific dictionary learning scheme to reflect the importance of visual words for a particular action class. Weighted class specific dictionary learning enriches the scheme to learn a sparse representation for a particular action class. To evaluate our scheme on realistic and complex scenarios, we have tested it on UCF Sports and UCF11 benchmark datasets. This paper reports experimental results that outperform recent state-of-the-art methods for the UCF Sports and the UCF11 dataset i.e. 98.93% and 93.88% in terms of average accuracy respectively. To the best of our knowledge, this contribution is first to apply a weighted class specific dictionary learning method on realistic human action recognition datasets.

Keywords: Human action recognition · Bag of visual words
Spatio-temporal features · UCF Sports

1 Introduction

Human action recognition is an emerging research area in computer vision, aiming at automatic classification of action present in a video. It has numerous applications such as intelligent surveillance systems, video search and retrieval, video indexing and human-computer interaction. Regardless of its popularity, it is one of the challenging problems in computer vision. Challenges include inter and intra class variation, changing viewpoint, cluttered background, camera motion etc. The low quality and high dimension of video data typically add difficulty to develop efficient and robust human action recognition algorithm.

Spatio-temporal features are extensively used for recognizing human actions [2,11,18,20] and have gained the state-of-the-art recognition performance on many challenging action recognition datasets. These approaches do not need to detect human body, rather they treat the action volume as a rigid 3D-object and extract appropriate features to describe the patterns of each 3D volume. They are robust to illumination changes, background clutter, and noise [20].

The Bag of visual words (BoVW) approach along with local features representation and its variations [15,20] have have proved to be effective for human action recognition especially for realistic datasets and it is popular due to its simplic-ity and computational efficiency. BoVW approach for human action recognition consists of four steps in general i.e. feature representation, codebook generation, feature encoding, and action classification. In each step, many efforts have been made for improvement.

Many local features representation approaches has been presented in literature [16]. Popular feature detectors include Dense Trajectories [23], STIPs [5] etc. and feature descriptors includes 3D SIFT [19], MBH [23], HOF [6], HOG for action representation. For visual word codebook generation, k-mean is a popular approach used for providing a partition for local descriptors in local feature space [4]. For feature encoding many methods are available for effective and efficient representation (Peng et al. [16] for detail study).

How to make decision in each step to obtain the best variation of BoVW for action recognition still remains unknown and needs to be extensively explored. In this work, we present a scheme to recognize human action using weighted visual word codebook learning. The work is based on our preliminary work [13], where general bag of visual words approach has been evaluated for recognizing human action in realistic and complex scenarios using spatio-temporal features.

Our previous work shows that performance can be significantly improved in complex and realistic scenarios by incorporating spatio-temporal domain information to represent an action in form of visual features. We have used the state-of-the art space time interest point detector and descriptor to capture the maximum possible information to represent an action. It represents video by utilizing characteristic shape and motion, independent of space time shifts. No prior segmentation like individual segmentation is needed for this approach. Our method is general and shows better results on different type of human action recognition datasets. We have also performed comparison with the state-of-art result for three different human action recognition datasets i.e. KTH, UCF Sports and Hollywood2.

We extend our proposed model from [13] and learn a concatenated weighted class specific dictionary. The proposed scheme assigns weights to each visual word with respect to its feature similarity and occurrence frequency *within* its action class so as t have a more efficient representation of an action class. For evaluation, we analyze our method on human action recognition benchmark datasets and show that our model outperforms the traditional bag-of-words approach.

The rest of paper is organized as follows. In Sect. 2 we describe our proposed weighted visual word codebook learning scheme. We introduce the weighting mechanism used for weight assignment by incorporating the importance of each visual word with respect to its feature similarity and occurrence frequency within its action class. In Sect. 3 we evaluate the importance of each parameter for our proposed approach and also compares our work with recent state-of-the-art methods. Section 4 presents the main conclusions and state the potential future direction for the proposed scheme.

2 Human Action Recognition Using Weighted Visual Words Codebook

Figure 1 provides an overview of the proposed scheme used for recognition of human action using weighted visual words codebook. Firstly, features are represented using a local feature representation approach followed by learning a weighted class specific dictionary. Feature encoding is performed to represent each input video using the histogram of weighted visual words. Finally, this histogram representation for training videos is used to train a supervised classifier. Similarly, during the testing phase, feature representation is obtained for unlabeled videos and quantized using the weighted codebook generated during the training phase. Feature encoding is performed to obtain a histogram of weighted visual words for testing videos and passed to trained classifier to obtain action labels. These processes are explained in more detail in the following sub-sections.

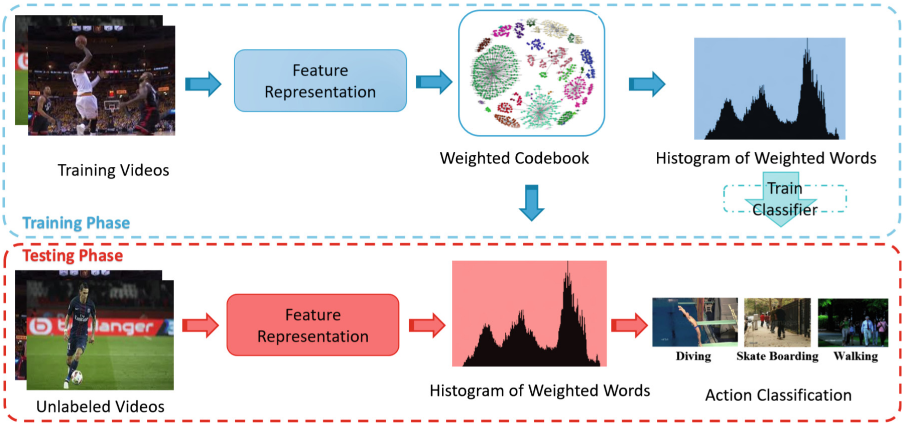


Fig. 1. Human action recognition using weighted visual words codebook scheme.

2.1 Feature Representation

A variety of feature representation methods exists for video representation. As proposed in [12], we use the popular 3D Harris interest point detector [5] to detect well-localized interest points in the spatio-temporal domain. These detected interest points are represented as $P = \{P_i \,|\, P_i \in (x,\, y,\, t)^n{}_{i=1}\}$. Detected interest points are shown (in space domain only) in Fig. 2. Further, we used the 3D SIFT descriptor [19] to describe these detected spatio-temporal interest points. 3D SIFT provides robustness to noise and orientation by encoding information in both space and time domains. Here feature vector is represented as $F = \{f_i\}^n{}_{i=1}$.

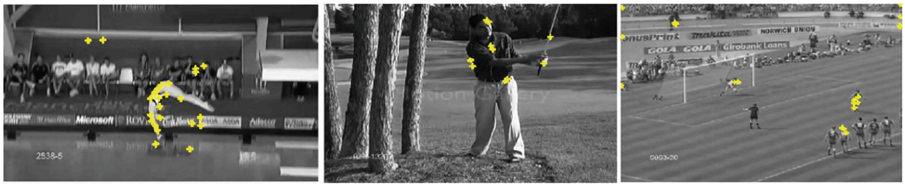


Fig. 2. Few detected space time interest points on example videos for UCF Sports dataset.

2.2 Weighted Class Specific Dictionary Learning

In the next step, we learn a weighted class specific dictionary which is discriminative enough to differentiate between each action class, by considering the relevance of each visual word within its action class. Consider the feature representation of an action class grouped together as $PF_l = \{P_i, F_i\}^{n_l}{}_{i=1}$, where n_l is the total no. of features l action class. We applied the popular k-means clustering algorithm [4] for the feature set PF_l and divide it into k clusters. Each cluster center is associated with a visual word. We formulated a weighting scheme for previously learned class specific dictionary based on a textual infor-mation concept. According to this concept words with high frequency describes a document better while the words occurrence frequency also depends on the length of a document [21]. Based on this concept, we assign a weight parameter WW_x to each visual word, μ_x (see below) is used to measure the similarity of a particular visual word with its action class. As a result, weight is calculated as

$$Weight(w_x) = \mu_x + WW_x \tag{1}$$

where μ_x is defined as

$$\mu_x = \frac{K(Fw_x, F_l)}{||wp_x - P_l||^2} \tag{2}$$

$K(Fw_x, F_l)$ is the cosine similarity measure between feature description of visual word w_x and set of feature vector F_l for class l. μ_x becomes low for the words that are too far and high for the words that are too close because of the distance value for wp_x and P_l, resulting in providing insignificant information in both cases. In such case we can ignore the value of μ_x.

In Eq. 8. WW_x provide an additional parameter to highlight the importance of a visual word within its action class according to its occurrence frequency.

We emphasize the visual words with high frequency by normalizing their occurrence frequency with the sum of visual words occurrence frequency in a particular action class. WW_x is defined as

$$WW_x = \frac{freq(w_x)}{\sum_{i=1}^{k} freq(w_i)} \tag{3}$$

Figure 3 shows the weighted visual words assigned to each action class for the UCF Sports action dataset using a graphical representation. There are a few action classes that have high weighted visual words representation e.g. Golf Swing, Riding Horse, Skate Boarding, Walking and Running as compared to other action classes. On the other hand, some action classes from the same dataset have less weighted visual words e.g. Lifting, Swing bench and Swing side. This weighted visual representation is highly dependent on the feature similarity of each visual word as well as its occurrence frequency within its relevant action class. Action classes with the high-weighted visual words representation tend to have more similar and high-frequency visual words representation with respect to other action classes.

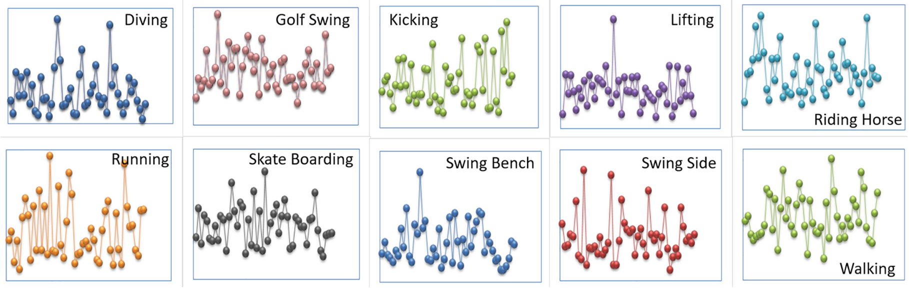


Fig. 3. Weighted visual words representation for UCF Sports dataset

2.3 Feature Encoding

In this step, the main focus is encoding feature representation for each video using weighted visual word codebook. Let $Feat = \{f_1, f_3, f_3, ..., f_z\}$ represents the features for each video. For each feature fm the codebook word wj can be viewed as function of f and is defined as:

$$A(f) = \begin{cases} Weight(w_j), \arg\min_j ||f_m - w_j||_2 \\ 0, \qquad\qquad\qquad otherwise \end{cases} \tag{4}$$

Each feature vector votes for only its nearest codebook word. The weighted occurrence of the votes are stored in histogram for each video.

2.4 Action Classification

For action classification, we used the popular supervised support vector machine (SVM) classification algorithm as proposed in [12]. SVMs is the state-of-the-art large margin classifier, which has recently gain popularity for human action recognition. We used a multi-class non-linear support vector machine trained using *c-1* binary SVM and *ordinal* coding design scheme. SVM used Gaussian kernel for learning which is defined as:

$$G(x_1, x_2) = exp(||x_1 - x_2||^2) \tag{5}$$

3 Performance Evaluation and Results

To test our scheme, we performed a number of experiments on publicly available datasets i.e. UCF Sports and UCF11. All experiments were carried out on an Intel Core i7-6500U CPU with 2.50 Ghz, and the proposed scheme was implemented in MATLAB 2015R(a).

3.1 Human Action Recognition Datasets

UCF Sports contains sports action videos captured in realistic environments. It contains 10 sports actions e.g. walking, diving, kicking, horseback riding etc. Sample frames from UCF Sports action dataset are shown in Fig. 4. It contains 150 video clips and the total duration of videos is 958 s. The average duration of action clips have great similarities across different classes, therefore it would not affect the performance of our proposed scheme. UCF Sports action videos have a large number of intra and inter-class variation typical of many real life environments. We used leave one out cross validation method as proposed in [17].

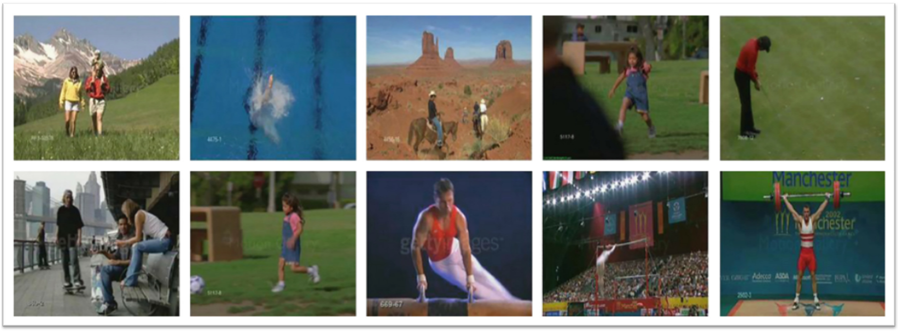


Fig. 4. Sample frames from the UCF Sports actions sequences.

UCF11, previously known as YouTube action dataset, is captured in realistic environments with large variation in viewpoints, backgrounds, camera motions, object appearances and poses. Sample frames from UCF11 action dataset are shown in Fig. 5. It contains 11 action categories and, for each action class, video clips are grouped into 25 groups each containing at least 4 video clips. Each group shares some similar features, like similar environment, same actor and similar viewpoints. As proposed by Rodriguez et al. [17], Leave One Group Out evaluation method is used for evaluation purposes.

Fig. 5. Sample frames from the UCF11 actions sequences.

3.2 Parameter Evaluation and Discussion

For constructing a class-specific dictionary, we performed k-means clustering using different no. of visual words for each action class. As shown in Table 1 accuracy and computational time increase with respect to an increase in the number of visual words till $k = 300$.

For evaluation of the weighted visual scheme, we performed different experiments and evaluated our approach on UCF Sports datasets. We studied the impact of both parameters used for weight assignment. As discussed in Sect. 2, the proposed weight is dependent on two different parameters and is defined as:

$$Weight(w_x) = \mu_x + WW_x \tag{6}$$

Table 1 shows the result of evaluation of effect of each individual parameter for UCF Sports dataset. By defining weight as:

$$Weight(w_x) = \mu_x \tag{7}$$

average accuracy is 97.5%. Here μ_x shows the feature similarity based importance of each visual words within its action class. The accuracy for second parameter is 96.5%. For which weight is defined as:

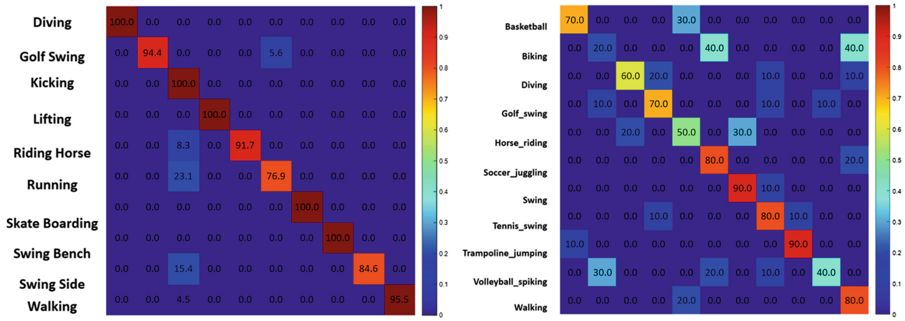
$$Weight(w_x) = WW_x \tag{8}$$

Here WW_x emphasize the importance of a visual word within its action class according to its occurrence frequency. This shows that both parameter have significant contribution in assigning the relevant weights to visual words representation. Finally, we evaluated our approach by combining both parameters and observed that it results in improved performance as compared to using only one parameter. We have also presented the performance of proposed scheme with $Weight(w_x) = 1$ to compare the effectiveness of proposed weighting scheme. Result shows that the proposed weighting scheme outperforms the base framework.

Table 1. Weighted visual words codebook parameter evaluation for UCF Sports dataset.

Weight's parameter	Accuracy
$Weight(w_x) = 1$	96.15%
$Weight(w_x) = \mu_x$	97.50%
$Weight(w_x) = WW_x$	96.50%
$Weight(w_x) = \mu_x + WWx$	**98.93%**

Figure 6(a) shows the performance of our scheme on the UCF sports dataset in the form of confusion matrix. Slight confusion between 'kicking' action class and a few other action classes is observed. Kicking is confused with Riding Horse, Running and Swing Side actions. Figure 6(b) shows confusion between a few action classes for the UCF11 dataset. As both datasets are captured in realistic scenarios some unwanted action in the background can mislead a classifier.



Diving	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Golf Swing	0.0	94.4	0.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0
Kicking	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lifting	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Riding Horse	0.0	0.0	8.3	0.0	91.7	0.0	0.0	0.0	0.0	0.0
Running	0.0	0.0	23.1	0.0	0.0	76.9	0.0	0.0	0.0	0.0
Skate Boarding	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Swing Bench	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Swing Side	0.0	0.0	15.4	0.0	0.0	0.0	0.0	0.0	84.6	0.0
Walking	0.0	0.0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	95.5

(a)

Basketball	70.0	0.0	0.0	0.0	30.0	0.0	0.0	0.0	0.0	0.0	0.0
Biking	0.0	20.0	0.0	0.0	0.0	40.0	0.0	0.0	0.0	0.0	40.0
Diving	0.0	0.0	60.0	20.0	0.0	0.0	0.0	10.0	0.0	0.0	10.0
Golf_swing	0.0	10.0	0.0	70.0	0.0	0.0	0.0	10.0	0.0	10.0	0.0
Horse_riding	0.0	0.0	20.0	0.0	50.0	0.0	30.0	0.0	0.0	0.0	0.0
Soccer_juggling	0.0	0.0	0.0	0.0	0.0	80.0	0.0	0.0	0.0	0.0	20.0
Swing	0.0	0.0	0.0	0.0	0.0	0.0	90.0	10.0	0.0	0.0	0.0
Tennis_swing	0.0	0.0	0.0	10.0	0.0	0.0	0.0	80.0	10.0	0.0	0.0
Trampoline_jumping	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.0	0.0	0.0
Volleyball_spiking	0.0	30.0	0.0	0.0	0.0	20.0	0.0	10.0	0.0	40.0	0.0
Walking	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	80.0

(b)

Fig. 6. Confusion matrix for UCF Sports and UCF11 datasets.

3.3 Comparison with State-of-the-Art Methods

Table 2 shows the comparison of our scheme with few recent approaches for human action recognition for UCF Sports and UCF11 datasets. Our proposed

Table 2. Comparison with state-of-the-arts methods for UCF-Sports and UCF11 Dataset.

Dataset	Paper	Method	Results
UCF Sports	**Our**	Weighted class specific dictionary learning scheme	**98.93%**
	[14]	Multi region two stream R-CNN	95.74%
	[1]	Bag of visual words	90.90%
	[23]	Dense trajectories and motion boundary descriptor	88.00%
	[3]	CNN + rank pooling	87.20%
	[9]	hk-means and TF-IDF scoring based vocabulary construction	78.40%
	[7]	Independent sub space analysis	86.50%
UCF11	**Our**	Weighted class specific dictionary learning scheme	**93.88%**
	[22]	Dense trajectories	84.20%
	[24]	Motion boundaries and dense trajectories	91.30%
	[10]	Tenser motion descriptor	75.40%
	[8]	Bag of visual words	71.20%

scheme outperforms other mentioned methods in term of average accuracy. We significantly improve the recognition performance for UCF Sports dataset presented in Peng and Schmid [14] by approximately 3%. Our scheme also showed significant improvement when compares with recent bag of visual words approach presented by Abdulmunem et al. in [1]. They have used saliency guided 3D SIFT-HOOF (SGSH) feature for feature representation. Performance is improved by around 3% for UCF Sports and for UCF11, we significantly improve the results presented in [24] by around 2%.

4 Conclusion and Future Work

In this paper, we enhance the performance of traditional bag of visual words approach for human action recognition. We have used local feature representation approach to represent complex action in realistic scenarios. Then, we exploit the importance of visual words for a particular action class. Our scheme is based on the concept of textual information present for document retrieval systems. We learned a class specific dictionary, further we assigned weight to each visual words based on its similarity with the respective action class and its occurrence frequency. Lastly, we have used these visual words weights to encode each videos. Our results showed improved performance for human action recognition in realistic and complex scenarios.

Future work includes several research directions. We can exploit the use of convolutional neural network instead for representing videos with handcrafted features such SIFT, HOF, 3D Harris etc. We can also improve the performance of bag of visual words by incorporating the neighboring information in a spatio-temporal grid for assigning weight to each visual word. Finally, the proposed scheme can be extended to work in complex and realistic scenarios i.e. Hollywood2 dataset.

Acknowledgements. Sergio A Velastin acknowledges funding by the Universidad Carlos III de Madrid, the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement n 600371, el Ministerio de Econom y Competitividad (COFUND2013-51509) and Banco Santander. Authors also acknowledges support from the Directorate of ASR and TD, University of Engineering and Technology Taxila, Pakistan.

References

1. Abdulmunem, A., Lai, Y., Sun, X.: Saliency guided local and global descriptors for effective action recognition. Comput. Vis. Media **2**(1), 97–106 (2016)
2. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2005 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE (2005)
3. Fernando, B., Gould, S.: Learning end-to-end video classification with rank-pooling. In: Proceedings of the International Conference on Machine Learning (ICML) (2016)
4. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. **31**(8), 651–666 (2010)
5. Laptev, I., Lindeberg, T.: Space-time interest points. In: proceedings of 9th IEEE International Conference on Computer Vision, Nice, France, pp. 432–439 (2003)
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
7. 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: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3361–3368. IEEE (2011)
8. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1996–2003. IEEE (2009)
9. Markatopoulou, F., Moumtzidou, A., Tzelepis, C., Avgerinakis, K., Gkalelis, N., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: ITI-CERTH participation to TRECVID 2013. In: TRECVID 2013 Workshop, Gaithersburg (2013)
10. Mota, V.F., Souza, J.I., Araújo, A.D.A., Vieira, M.B.: Combining orientation tensors for human action recognition. In: 2013 26th SIBGRAPI-Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 328–333. IEEE (2013)
11. Murthy, O., Goecke, R.: Ordered trajectories for large scale human action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 412–419 (2013)
12. Nazir, S., Haroon, M., Velastin, S.A.: Inter and intra class correlation analysis (IICCA) for human action recognition in realistic scenarios. In: International Conference of Pattern Recognition Systems (ICPRS) (2017, to appear)
13. Nazir, S., Haroon Yousaf, M., Velastin, S.A.: Evaluating bag of visual features (BoVF) approach using spatio temporal features for action recognition. In: Computer and Electrical Engineering (2017, submitted)
14. Peng, X., Schmid, C.: Multi-region two-stream R-CNN for action detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016 Part IV. LNCS, vol. 9908, pp. 744–759. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0 45

15. Peng, X., Wang, L., Cai, Z., Qiao, Y., Peng, Q.: Hybrid super vector with improved dense trajectories for action recognition. In: ICCV Workshops, vol. 13 (2013)
16. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput. Vis. Image Underst. **150**, 109–125 (2016)
17. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
18. 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)
19. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia, pp. 357–360. ACM (2007)
20. Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2004–2011. IEEE (2009)
21. Tirilly, P., Claveau, V., Gros, P.: A review of weighting schemes for bag of visual words image retrieval. Research report PI 1927, p. 47 (2009)
22. Wang, H., Kläser, 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)
23. Wang, H., Kläser, A., Schmid, C., Liu, C.-L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
24. Yadav, G.K., Shukla, P., Sethfi, A.: Action recognition using interest points capturing differential motion information. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1881–1885. IEEE (2016)