uc3m | Universidad Carlos III de Madrid

e-Archivo

This is a postprint version of the published document at:

Nida, N., Yousaf, M.H., Irtaza, A. y Velastin, S.A. (2019). Bag of Deep Features for Instructor Activity Recognition in Lecture Room. In *MultiMedia Modeling*, *LNCS,* 11296, pp. 481-492.

DOI: https://doi.org/10.1007/978-3-030-05716-9_39

© Springer Nature Switzerland AG 2019.

Bag of Deep Features for Instructor Activity Recognition in Lecture Room

Nudrat Nida[1] , Muhammad Haroon Yousaf[1(✉)] , Aun Irtaza[2], and Sergio A. Velastin[3,4,5]

[1] Department of Computer Engineering,
University of Engineering and Technology, Taxila, Pakistan
{16F-PHD-CP-53,haroon.yousaf}@uettaxila.edu.pk
[2] Department of Computer Science,
University of Engineering and Technology, Taxila, Pakistan
aun.irtaza@uettaxila.edu.pk
[3] Department of Computer Science, Applied Artificial Intelligence Research Group,
University Carlos III de Madrid, 28270 Madrid, Spain
[4] Cortexica Vision Systems Ltd., London SE1 9LQ, UK
[5] School of Electronic Engineering and Computer Science,
Queen Mary University of London, London E1 4NS, UK
sergio.velastin@ieee.org

Abstract. This research aims to explore contextual visual information in the lecture room, to assist an instructor to articulate the effectiveness of the delivered lecture. The objective is to enable a self-evaluation mechanism for the instructor to improve lecture productivity by understanding their activities. Teacher's effectiveness has a remarkable impact on uplifting students performance to make them succeed academically and professionally. Therefore, the process of lecture evaluation can significantly contribute to improve academic quality and governance. In this paper, we propose a vision-based framework to recognize the activities of the instructor for self-evaluation of the delivered lectures. The proposed approach uses motion templates of instructor activities and describes them through a Bag-of-Deep features (BoDF) representation. Deep spatio-temporal features extracted from motion templates are utilized to compile a visual vocabulary. The visual vocabulary for instructor activity recognition is quantized to optimize the learning model. A Support Vector Machine classifier is used to generate the model and predict the instructor activities. We evaluated the proposed scheme on a self-captured lecture room dataset, IAVID-1. Eight instructor activities: pointing towards the student, pointing towards board or screen, idle, interacting, sitting, walking, using a mobile phone and using a laptop, are recognized with an 85.41% accuracy. As a result, the proposed framework enables instructor activity recognition without human intervention.

Keywords: Human activity recognition
Instructor activity recognition · Motion templates
Academic quality assurance

1 Introduction

An effective teacher is a source of inspiration and responsible for students achievement. The students' academic performance is highly dependent upon the teacher's effectiveness and their behavioral traits. In educational institutes, students are surveyed for the evaluation of instructor effectiveness and quality of the delivered lecture. In this regard, institutes are evaluating the effectiveness of teachers and lectures mainly through student feedback. This feedback is based on the standard survey mechanism "Student's Evaluating Teaching (SET)". The prime motive to examine feedback is to improve the quality of the lecture. Unfortunately, in SET a teacher's instructional and behavioral skills are evaluated on a smaller scale. Moreover, SET feedback are usually collected at the end of the semester which is not beneficial for students enrolled in a current semester. These performance evaluation statistics have being collected for a long time and empirically have been fount to have no significant impact on ratings of teaching [5]. The root cause of inaccurate instructor's performance insight is due to poorly designed questionnaires, personal biases, and non-serious student's response. Hence, an alternative system is required to support the evaluation process that can provide a consistent view of the quality of lecture and effectiveness of the teachers.

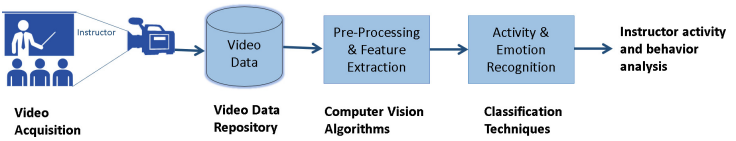

Fig. 1. The visual classroom information for the development of intelligent application using actions and emotion recognition techniques.

Everyday numerous hours of video data are recorded in academic institutes across the world. However, the majority of institutes rarely analyze the recorded or live stream videos for instructor performance evaluation or estimation of the lecture effectiveness, as illustrated in Fig. 1. Such real-time classroom data has an enormous potential to explore various problem domains within a classroom and to provide a solution to academicians to understand visual semantics using computer vision and pattern recognition techniques as shown in Fig. 1. Recently, some computer vision findings have been reported to automatically estimate instructor performance using pose, gesture and activity recognition [15, 19, 20]. In [15], vision-based instructor activity recognition uses silhouette representation to train a Hidden Markov Model (HMM). The system was able to identify five activities: walking, writing, pointing towards the board, standing and pointing towards presentations with a recognition accuracy of 90%. Similarly, in [19] face recognition and pose estimation techniques were applied to analyze academic performance within a lecture room. Structure and texture features were used to

localize objects and the instructor within the classroom. A Bayesian classifier was used to model five activities: standing, walking, pointing, writing and addressing. The achieved accuracy was reported to be 96%. In [20], instructor activities were used to record a lecture automatically through localization of classroom objects and instructor. Then morphological features were used to generate fuzzy rules for instructor activities recognition. Such techniques [15,19,20] use spatial data for activity recognition but have ignored temporal information and have had to resort to their own datasets because of the unavailability of standard activ-ity datasets for instructor evaluation. Consequently, there is a need to develop standard datasets for researchers to compare and improve results.

Especially during the last decade, understanding visual information using computer vision techniques and systems has been useful to recognize human activities and behaviors in real-time in applications such as in surveillance, robot navigation, elderly home care, etc. The literature on human action recognition can be grouped into two broad categories: handcrafted and deep learning techniques. Activity recognition based on handcrafted features can be further categorized into spatio-temporal [10], motion template [14] and action trajectory information [18]. Spatio-temporal features tend to be sparse and not of fixed length, which affects accuracy [10,21]. The varying length of spatio-temporal features may be overcome through time evolution of actor silhouettes [14,21]. However, accurate segmentation of actor silhouettes is a challenging problem. Action trajectory information [18] is very effective but it is computationally expensive to capture temporal movement information of actor which is also sensitive to occlusion and noise. Handcrafted action representations target specific applications and thus fail to provide generic solutions [21]. Recently, deep learning based techniques have been shown to outperform traditional methods in most recognition tasks and that has motivated researchers to explore its capabilities for action recognition especially using spatio-temporal data. In the action recognition domain, deep learning solutions can be categorized as frame learning [2,12], transformed frame learning [1], handcrafted features with deep representation [4], 3D convolutional neural network [3] and hybrid models [7,16]. Deep learning based solutions can find generalized models for real-time application but suffer from the scarcity of standard video datasets [21]. Frame learning tries to predict action recognition without learning temporal information [2,12] optimizes the models by tuning the weights. A limitation of the frame learning technique is that resolution and number of frames are fixed for all action sequences, yet realistic action videos are not of fixed length. Transformed frame learning overcomes limitations of frame learning by incorporating temporal information from adjacent frames. However, it works best on smaller resolution video frames that makes it inadequate for high resolution action prediction. Recurrent neural networks are used to learn the sequential action information and predict activities [7]. Deep models for action representation using handcrafted features as input data may be appropriate for human action recognition, assuming that adequate features can be found [4]. 3D CNN [3] techniques use modified 2D CNN to embed temporal information. However, prediction results are not much

better than 2D frame learning techniques [21]. Deep learning techniques are data hungry methods that require large-scale data representation and powerful computational resources. Among these techniques, a fusion of handcrafted and deep learning features might offer promising recognition results, as compared to state of the art techniques, due to higher dimensional action representation [21].

This research work proposes a technique for action recognition that uses a fusion of handcrafted and deep learning features to generate Bag-of-Deep-Features (BoDF) for instructor activity recognition. Such holds a high dimensional discriminative power to recognize different objects [21], and performs promisingly to recognize instructor activities. We evaluated this technique on a newly created video dataset: "Instructor Activity Video-I (IAVID-I)". Our contributions are: (i) To utilize a computer vision technique for understanding the visual semantics of classroom for academic quality assurance, (ii) Proposing a novel Bag-of-Deep-Features technique for instructor activity recognition, (iii) The proposed technique has the potential to solvie action recognition irrespective of the application domain, as the motion template generated from human silhouette captures the spatio-temporal representation of an actor that is benefical for accurate prediction of activities, (iv) To make available to other researchers a new dataset and a baseline set of results.

2 Proposed Methodology

Bag of Feature (BoF) is one of the most effective frameworks for various image and video classification applications [11]. The BoF for action recognition fol-lows a generic pipeline: (i) extraction of 3D feature detector and descriptor,(ii) Construction of visual vocabulary, (iii) Quantization of visual vocabulary,(iv) Generation of a training model for action prediction, (v) Testing. Boradly, our technique is based on the fusion of handcrafted features (i.e. MT) and deep features for construction of a BoDF representation. We believe that the fusion of handcrafted motion templates of instructor and deep representation is capable

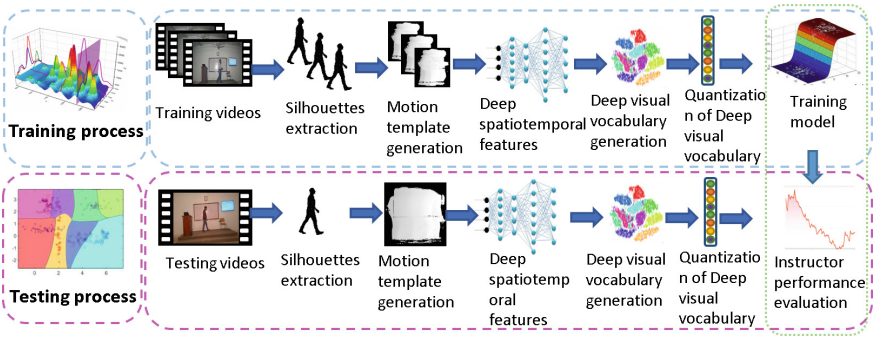

Fig. 2. The visual classroom information for the development of intelligent application using actions and emotion recognition techniques.

to predict the instructor's activities. To validate this hypothesis, handcrafted motion templates of instructor's activities and deep features from different lay-ers of a CNN network [6] are fused together and each video sequence is repre-sented by a feature vector of fixed size. The proposed technique is based on a generic BoF [13] representation with some modifications at the feature compu-tation step. Since deep learning techniques have improved image classification performance [6], therefore it is amenable to utilize deep features for action rep-resentation using BoF paradigm. In BoDF, the deep features are computed as $Z = z_i, i\epsilon\,(1, \ldots, N)$, where $z_i \epsilon DST\ F^{\ D}$ is deep spatio-temporal action descriptor computed from action video sequences and N is the length of DSTF of fixed size D. Where D is 4096 or 1000 depending upon the fully connected layer used for feature extraction. Then, encoding these DSTF to obtain an optimal represen-tation of action videos through function, $f\colon DSTF^D \Rightarrow DSTF^K$. The coding function f maps the deep spatio-temporal $DSTF$ into deep visual dictionary representation into K clusters. Then, the histogram h is used to quantize the deep visual vocabulary to train a SVM model for instructor action prediction. The vocabulary size is varied to examine the behavior of the model. The expe-riential results are discussed in the results and findings section. The following subsection will explain the methodology in detail.

2.1 Preprocessing

Initially, RGB video action sequences of classroom lectures are presented to the framework as input. Then, instructor silhouettes are extracted using graph cut segmentation. In minimum graph cut segmentation, each pixel of the instructor activity video frame is represented by a graph node. Each node is connected to each other through a vertex. In our technique, Gaussian distribution is used to assign probability weights to each vertex and segment the instructor silhouetted $s(x, y, t)$ from the classroom static background based on associated probabil-ity weights. The minimum probability at the vertex is responsible to segment the instructor silhouettes from the static background. We extract instructor sil-houettes to encode instructor motion information of the entire video sequence through motion templates. On application of graph cut segmentation, instructor silhouettes $s(x, y, t)$ are obtained for each video frame, i.e. the spatial location of the instructor as a binary representation.

2.2 Motion Template (MT) Generation

The instructor binary silhouettes generated from the video sequence are pro-cessed further to form motion templates, as shown in Fig. 2. These templates hold the spatio-temporal representation of action sequences and computed for all the training and testing video sequences. The motion template (MT) is a func-tion of intensity for holding information on the most recent spatial location of motion [8]. A brighter pixel indicates recent motion location of instructor within

the classroom. MT is computed using Eq. 1, where MT is spatio-temporal template generated from a silhouette frame $s(x, y, t)$ represents the object of interest, i.e. instructor at time t at location (x, y) as shown in expression (1).

$$MT = \begin{cases} \tau & if s(x, y, t) = 1 \\ max(0, s_{t-1}(x, y) - 1) & otherwise \end{cases} \quad (1)$$

Here, τ is total number of frames used for generation of MT for every action sequence in a similar way. The benefit of using MT is to reduce the spatial and computational complexity of action recognition. The resultant MT is cumulative greyscale motion representation of the instructor in an action video sequence, as 3D instructor spatial and temporal information are mapped into 2D greyscale MT. All the videos MT are normalized, wrapped to 227×227 or 224×224 dimension and centered to reduce redundant information. The wrapping and centering processes are applied to overcome constraints of spatial location, viewpoint variation, and scale, as motion templates are sensitive to spatial location and viewpoint.

2.3 Deep Spatiotemporal Features (DSTF)

Then, these spatio-temporal MTs are described through deep features from a pre-trained AlexNet CNN network [6] and VGG19 [17] to form deep visual words ($DV W$). The aim is to obtain higher dimensional spatio-temporal instructor action representation of MT using deep visual words at different network depth. Visual patches of motion templates are represented as deep numerical vectors to represent each type of instructor activities. The input layer in the CNN receives the MT and passes it to the convolutional layer. The convolutional layer performs convolution of MT at the smaller region with weights to generate a y feature map of neurons. Assume that we have spatio-temporal template MT of dimension MxM and present into CNN to extract $DSTF$, ultimately forming the DVW for instructor activity recognition. The receptive field or kernel of size is $r \times r$ and w is the number of the kernel, the convolutional layer will generate an output neuron volume of $(M - r + 1) \times (M - r + 1)$ in Eq. 2.

$$DSTF_{mn} = \sum_{\alpha=0}^{r-1} \sum_{\beta=0}^{r-1} wy_{x+i,y+i}^{r-1} \quad (2)$$

We have computed the $DV W$ from a 25 layered Alexnet [6] at different network depths, such as deep features are extracted from fully connected layer 17(DVW_{17}), 20(DVW_{20}) and 23(DVW_{23}) respectively. There are 196,608 deep visual words generated when DVW_{17} and DVW_{20} are used for feature extraction and 43,000 deep visual words are generated when DVW_{23} is used for feature computation. Similarly, we have computed the DVW from 47 layered VGG19 [17] at different network depths, and in this case deep features are extracted from fully connected layer 39(DVW_{39}), 42(DVW_{42}) and 45(DVW_{45}) respectively.

There are 196,608 deep visual words generated when DVW_{39} and DVW_{42} are used for feature extraction and 43,000 deep visual words generated when DVW_{45} is used for feature computation. We have explored the deep features capability to represent actions through BoDF representation, as it was not yet explored for action representation. We argue that the deep representation of motion templates is a major factor for precise action recognition, due to higher dimensional feature representation.

2.4 Deep Visual Vocabulary Generation and Quantization

Then, as illustrated in Fig. 2, the next step includes generation of visual vocabulary through unsupervised clustering of deep visual words (DVW) by K-means clustering algorithm. Suppose there are N activities of the instructor that are divided into K clusters, such that all the DVW are assigned to centroids of the cluster through minimizing the distance between the cluster centroid and $DSTF$. The K or vocabulary size is varied from 100 to 500 and the performance of the proposed technique was analyzed. The deep visual vocabulary represents the DVW of each instructor activities as the frequency of occurrence of DVW. The visual vocabulary is beneficial for estimating the instructor activities through DVW histogram to quantize the deep visual vocabulary. We have selected the 40% strongest DVW for quantization of deep visual vocabulary. These DVW are divided into 4,800 bins of final histogram.

2.5 The Training and Testing Video for Instructor Action Recognition

A Support Vector Machine (SVM) classifier is used to train the instructor activity recognition model from the quantized deep visual vocabulary representation of instructor activities. The SVM classifier defines decision boundaries separating the set of instructor actions having different class memberships. SVM performs classification through the generation of hyperplanes across multidimensional space that discriminate video samples of different instructor action classes. Later on, the test video is represented by a histogram of DVW and an SVM used to predict the instructor's activity.

3 Results and Findings

In this section, we describe a series of experiments performed on the Instructor Activity Video (IAVID-I) dataset, for evaluation of our system at various deep CNN depth for feature learning and computational cost is also estimated. Hardware and software specification of our system is an NVIDIA GTX-950 GPU card, Windows 10, an Intel i7-7700K processor with 4.5 GHz and 12 GB memory. The framework was implemented using MATLAB's Deep Learning Toolbox.

3.1 IAVID-I Dataset

We have recorded video dataset, Instructor Activity Video (IAVID-I), to recognize the activities of the instructor using real-time classroom video data. The environmental condition remains the same during recording, 12 actors participated in the acquisition phase focusing on stage. There are 100 videos having 854×480 high-resolution RGB 24 bit videos. There are eight actions in this dataset, i.e. interacting or idle, pointing towards the board, pointing towards the screen, using a mobile phone, using a laptop, sitting, walking and writing on the board, as illustrated in Fig. 3.

Table 1. Evaluation of deep BoDF for instructor activity recognition at various network depth and vocabulary size.

sr.no	DVW	DSTF	DVW dimension	Accuracy
1	DVW_{17} [6]	4096	196608	84.32%
2	DVW_{20} [6]	4096	196608	85.41%
3	DVW_{23} [6]	1000	48000	83.33%
4	DVW_{39} [17]	4096	196608	70.00%
5	DVW_{42} [17]	4096	196608	75.56%
6	DVW_{45} [17]	1000	48000	66.67%

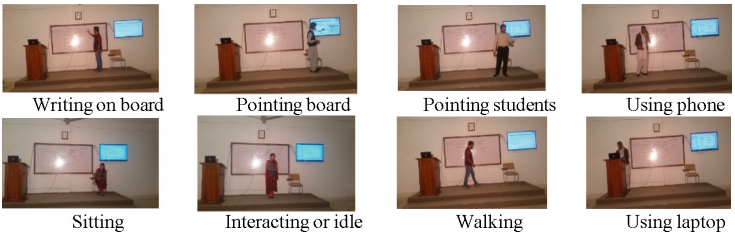

Fig. 3. The visual classroom information for the development of intelligent application using actions and emotion recognition techniques.

3.2 Evaluation of Deep BoDF at Various Network Depth

We followed a cross validation scheme on the IAVID-I dataset by randomly holding half of the video sequences for testing, while the other half was used for training the model. The hyper-parameters were set as per the pre-trained network Alexnet [6] and VGG19 [17]. Deep features are extracted from fully connected layers at network depth of 17, 20, 23, 39, 42, and 45. The prediction accuracy is the most reliable measure to estimate the robustness of the learned model. Therefore, Table 1 summarizes the performance of the deep BoDF model in terms

of prediction accuracy. It is notable that DVW_{20} performed better than DVW_{17} and DVW_{23} at visual vocabulary size of 100, due to higher dimensional feature representation as compared to DVW_{23}. DVW_{17}, computed from shallower fully connected layer reduces performance by 1.09% compared with DVW_{20}, due to the fact that some features at shallower layers correlate with each other, and resulted in slight variation in prediction accuracy. From Table 1, it is observed that the greater the number of visual words, the better will be the represen-tation of action classes. The DVW extracted from Alexnet performed better than VGG19, as hyper-parameter and network architecture varies. However, this adoption needs a lot of experimentation i.e. architecture configuration and extension to make it able to be used as an real time system giving a performance comparable to humans in an efficient way. Thus, experimental results portray that the optimal choice of DVW computation is from 20 layers of Alexnet pre-trained network [6]. We have computed the confusion matrix, as shown in Fig. 4, using a visual vocabulary size of 100 generated from DVW_{20}, achieving 85.41% accuracy. From the confusion matrix it can be observed that the lowest predic-tion accuracy occurs for writing on board, due to the fact that in some sequences actors were walking while writing on board, therefore the writing action class is confused with walking. Similarly, instructor action pointing towards the student is also confused with pointing towards the board because of the visual similarity of motion templates of two action classes. To further examine the performance of the proposed method, we have plotted a box and whisker plot to analyze the spread of prediction accuracy, as shown in Fig. 5. The box and whisker plot presents the distribution of accuracy across the number line and divides it into four quartiles, and median accuracy. From the spread of plot, it is notable that visual words DVW_{20} perform better than DVW_{17} and DVW_{23} at all vocabulary size and the accuracy spectrum is position at upper half of the box and whisker plot, representing higher prediction score. The deep features hold higher discrim-ination representation among the classes, therefore at minimum vocabulary size, DVW performed well to fitting data into suitable class boundaries for precise prediction of instructor activities. Moreover, shallower fully connected layers of CNN networks holds higher dimensional deep features to generate discriminative DVW for BoDF representation. Table 2 describes the computational cost of the proposed method. In the preprocessing step, instructor silhouette extraction and MT generation required 1 min per sequence on average for 30 fps. The computa-tional cost is averaged for all the task. It is concluded that deep BoDF requires a smaller vocabulary size to learn prediction model at a lower computational time. On average, the minimum time required for prediction of action is 0.43 s. In the IAVID-I dataset, there is a total of 8 action classes having 100 videos sequences and each frame is 854×480 resolution. Using the cross-validation scheme, train-ing and testing of 100 video sequences takes 30 s, i.e. on average it takes 0.43 s per sequence at a frame rate of 139.53 frames/second (FPS). It is concluded that BoDF requires a small vocabulary size to learn prediction model at the lower computational time.

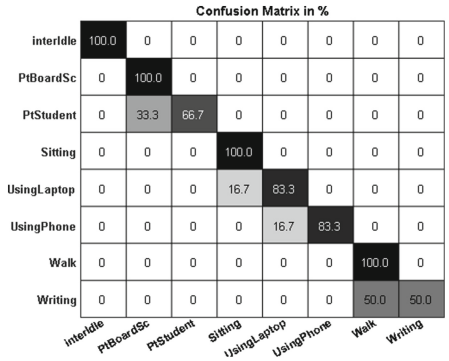

Fig. 4. Confusion matrix achieved from the deep BoDF representation of DVW_{20} [6] at vocabulary size of 100.

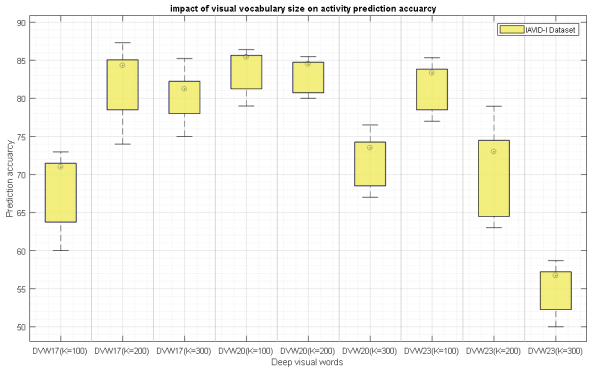

Fig. 5. Impact of visual vocabulary size on prediction accuracy.

Table 2. Evaluation of deep BoDF for instructor activity recognition at various network depth and vocabulary size.

	Propose technique	Time (hh:mm:ss)
1	Preprocessing and motion template generation	2:00:00
2	BoDF (training)	0:05:00
3	BoDF (testing)	0.43 s

4 Conclusion

In this paper, we have presented a BoDF method for instructor activity recognition. The deep model learns instructor activities through spatio-temporal deep features to form deep visual words. These deep visual words enable an SVM clas-

sifier to recognize the activities of the instructor. Such application is significant to understand the classroom contextual information and helpful for instructor self-evaluation. Through empirical analysis on network depth and different type of CNN model, reveals that AlexNet performs better than VGG19. The goal of our work is improved academic performance for societal gain rather than solely profit gain. As future work, we are focusing on instructor activities for self-evaluation of the instructor, and later on, we will analyze the behaviors, emotions of the instructor along with audience engagement for a more comprehensive evaluation of lecture effectiveness. For real time action recognition we will explore temporal action segmentation method [9], as instructors perform multiple activities sequentially. In conclusion, the availability of a commercial lecture effectiveness tool will enhance teachers' effectiveness and lifelong learning of instructors to overcome many classroom challenges.

Acknowledgements. Sergio A Velastin has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2014-51509) el Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander.

References

1. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
2. Ijjina, E.P., Chalavadi, K.M.: Human action recognition using genetic algorithms and convolutional neural networks. Pattern Recognit. **59**, 199–212 (2016)
3. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
4. Kim, H.-J., Lee, J.S., Yang, H.-S.: Human action recognition using a modified convolutional neural network. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) ISNN 2007. LNCS, vol. 4492, pp. 715–723. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72393-6_85
5. Knol, M.H., Dolan, C.V., Mellenbergh, G.J., van der Maas, H.L.: Measuring the quality of university lectures: development and validation of the instructional skills questionnaire (ISQ). PloS One **11**(2), e0149163 (2016)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
7. Li, W., Wen, L., Chang, M.C., Lim, S.N., Lyu, S.: Adaptive RNN tree for large-scale human action recognition. In: ICCV, pp. 1453–1461 (2017)
8. Murtaza, F., Yousaf, M.H., Velastin, S.A.: Multi-view human action recognition using 2D motion templates based on MHIS and their hog description. IET Comput. Vis. **10**(7), 758–767 (2016)
9. Murtaza, F., Yousaf, M.H., Velastin, S.A.: PMHI: proposals from motion history images for temporal segmentation of long uncut videos. IEEE Signal Process. Lett. **25**(2), 179–183 (2018)
10. Nazir, S., Yousaf, M.H., Nebel, J.C., Velastin, S.A.: A bag of expression framework for improved human action recognition. Pattern Recognit. Lett. **103**, 39–45 (2018)

11. Nazir, S., Yousaf, M.H., Velastin, S.A.: Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. Computers & Electrical Engineering (2018)
12. Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E.: Toward automatic phenotyping of developing embryos from videos. IEEE Trans. Image Process. **14**(9), 1360–1371 (2005)
13. O'Hara, S., Draper, B.A.: Introduction to the bag of features paradigm for image classification and retrieval. arXiv preprint arXiv:1101.3354 (2011)
14. Orrite, C., Rodriguez, M., Herrero, E., Rogez, G., Velastin, S.A.: Automatic segmentation and recognition of human actions in monocular sequences. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 4218–4223. IEEE (2014)
15. Raza, A., Yousaf, M.H., Sial, H.A., Raja, G.: HMM-based scheme for smart instructor activity recognition in a lecture room environment. SmartCR **5**(6), 578–590 (2015)
16. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1762–1774 (2009)
19. Yousaf, M.H., Azhar, K., Sial, H.A.: A novel vision based approach for instructor's performance and behavior analysis. In: 2015 International Conference on Communications, Signal Processing, and Their Applications (ICCSPA), pp. 1–6. IEEE (2015)
20. Yousaf, M.H., Habib, H.A., Azhar, K.: Fuzzy classification of instructor morphological features for autonomous lecture recording system. Inf. J. **16**(8), 6367 (2013)
21. Zhu, F., Shao, L., Xie, J., Fang, Y.: From handcrafted to learned representations for human action recognition: a survey. Image Vis. Comput. **55**, 42–52 (2016)