Research Article

The Recognition of Teacher Behavior Based on Multimodal Information Fusion

Dongli Wu,^{1,2} Jia Chen ,^{1,2} Wei Deng,^{1,2} Yantao Wei,^{1,2} Heng Luo,^{1,2} and Yangyu Wei^{1,2}

¹School of Educational Information Technology, Central China Normal University, Wuhan 430079, Hubei, China

²Hubei Research Center for Educational Informationization, Wuhan 430079, Hubei, China

Correspondence should be addressed to Jia Chen; jc@mail.ccnu.edu.cn

Received 21 August 2020; Revised 17 September 2020; Accepted 24 September 2020; Published 13 October 2020

Academic Editor: Jianbing Ma

Copyright © 2020 Dongli Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Teaching reflection based on videos is the main method in teacher education and professional development. However, it takes a long time to analyse videos, and teachers are easy to fall into the state of information overload. With the development of “AI + education,” automatic recognition of teacher behavior to support teaching reflection has become an important research topic. In this paper, taking online open classroom teaching video as the data source, we collected and constructed a teacher behavior dataset. Using this dataset, we explored the behavior recognition methods based on RGB video and skeleton information, and the information fusion between them is carried out to improve the recognition accuracy. The experimental results show that the fusion of RGB information and skeleton information can improve the recognition accuracy, and the early-fusion effect is better than the late-fusion effect. This study helps to solve the problems of time-consumption and information overload in teaching reflection and then helps teachers to optimize the teaching strategies and improve the teaching efficiency.

1. Introduction

In the past 20 years, because video can truly reproduce classroom teaching and provide detailed materials for teachers’ teaching reflection, various methods based on video have been used for teaching reflection. Although video is considered as a very effective tool for teacher professional development, due to the large amount of information contained in the real scene of the classroom, teachers have problems in the selection of videos, the focus of reflection, and the energy consumed, which makes them easy to fall into the state of information overload [1–4]. Some empirical studies also show that a considerable number of teachers have not effectively improved their teaching through teaching reflection. One of the main reasons is that teachers’ teaching tasks are heavy. They usually think briefly in their minds or write down a few notes occasionally. And, it is almost impossible to analyse teaching videos and review teaching scenes. In view of this, some researchers have developed a framework to guide teachers to choose videos and watch video, such as five heuristic studies on video

viewing by novice teachers [5] and the principle of watching teaching video proposed [3]. These studies focus on teaching principles and strategies, but few have been done to alleviate the problem of information overload based on video.

In recent years, with the development of artificial intelligence theory and technology, human action recognition, as one of its branches, has achieved good results in many fields such as video monitoring, retrieval, human-computer interaction, virtual reality, and body-sense games. In the field of education, human action recognition has also been applied to a certain extent. For example, the virtual reality system applied to dance teaching can provide score feedback on the completion degree of each part of the learners and find out the wrong parts by comparing with the dance movements of the virtual teacher [6]; the teaching can be conducted in a more natural way by recognizing the teacher’s gestures instead of blackboard chalk, keyboard, and mouse [7]; real-time classroom attendance and classroom performance were obtained by student behavior detection [8]. However, there is little research in the field of teacher professional development. Teachers’ gestures and behaviors

in teaching play an important role in emphasizing and demonstrating, which can enhance the appeal of teaching atmosphere, strengthen the teaching information, and influence the students' beliefs, so it is an important basis for reflecting the state of teachers' teaching [9, 10]. From the perspective of the development of teachers' professional ability, it is particularly important to collect and analyse the data of teacher behavior in the classroom to support teachers' teaching reflection. However, the types of teacher behavior and student behavior are different and teachers have been in a state of free walking. It leads to a high degree of background variation and great differences in behaviors within and between classes, which makes it difficult to recognize behaviors.

To sum up, this paper researches the recognition of teacher behavior based on classroom video, in order to improve the recognition accuracy of teachers' classroom behavior, help teachers to quickly analyse classroom teaching video, and reduce the problem of information overload, so as to support teaching reflection, promote teachers' professional development, and help teaching reform.

2. Related Work

2.1. Teacher Behavior Based on Video. As early as the 1860s, Stanford University used video to carry out microteaching and simplify the teaching process, which was mainly used for fore-service training of teachers. At the end of the 20th century, the video case method gradually rose and was used in classroom teaching observation research, such as the global video research project TIMSS. After entering the 21st century, people pay more and more attention to the application of classroom teaching video. Video can stimulate teachers' memory of the classroom, overcome the shortcomings of traditional recall reflection, and enable teachers to think more clearly about classroom teaching, so it has gradually become the main teaching reflection method.

To effectively apply video in teaching reflection, some guidance is needed, including the whole process of video recording, editing, and analysis [1]. Video recording needs to set the camera position to photograph the classroom's picture, generally including two cameras for the teacher and students. And then, select the appropriate screen according to the actual situation. For video editing, it needs to select the appropriate video clips from different perspective videos based on the teaching process, while at the same time preventing the skips and stutters with video. The process is to analyse and reflect on a certain video clip from the perspective of teaching. It can be seen that this series of processes needs a lot of time for teachers. If the computer can automatically recognize teacher behavior and sort the behavior in a structured way, teachers can understand their performance in the classroom from the macrolevel and can quickly and accurately locate the places they are interested in. It can reduce the burden of teachers and alleviate the information overload problem in the paper.

2.2. Human Action Recognition. The task of human action recognition is to classify the actions in a series of frames (or videos), which can be divided into three stages: feature extraction, action representation, and action classification. Feature extraction is to extract features related to human action with pattern invariance and discrimination; action representation refers to the action representation of video obtained from the distribution or state change of statistical features in each video; classification is to classify the action representation of video into a certain type of human action by using a classification algorithm. In recent years, the performance of deep learning methods in action recognition is gradually better than traditional methods because it no longer needs manual design and extraction of features but through a large number of data training automatic learning features. However, the method based on deep learning needs a lot of data support. This paper studies the traditional methods.

2.2.1. Feature Extraction. Early human action recognition is mainly for simple scenes, and the main method is to extract global features, such as human contour, human skeleton, and human motion field [11–14]. With the continuous expansion of the application field, it is necessary to recognize more and more complex human actions, and the background of video recording is becoming complex. It is difficult to extract reliable global features from video, which makes the human action recognition method based on global features difficult to meet the performance requirements of applications. Subsequently, more and more attention has been paid to human action recognition based on local features [15–17]. Compared with the global features, local features describe the local space-time region through the feature descriptor with certain pattern invariance, which is more robust to the complex background caused by the changes of view angle, illumination and scale, and so on. To solve the problem of human action recognition in complex scenes, it is not enough to detect the grey changes in the temporal and spatial regions. Therefore, researchers have proposed many feature extraction methods based on feature point tracking [18, 19]. These methods first detect the feature points in the spatiotemporal region of the video, then track these feature points frame by frame, and connect them to form the trajectory of the feature points. Then, the trajectory and its spatiotemporal neighbourhood are described by feature descriptors.

2.2.2. Action Representation. It is to model the action features. Of course, some early scholars directly classified the features [20]. Although this is simple, it ignores the connection between video frames, and the effect is not good. The commonly used methods are the template matching method [21], state space model [12, 22], bag-of-words model [19], and so on. Based on the template matching method, the action pattern is represented by one or a group of static templates, and the possibility that the video belongs to a specific action is determined by calculating the distance between the video to be recognized and the action template. Its advantages are low computational complexity and simple implementation process. However, since the action template is

a static fixed mode and the intraclass divergence is large, it is difficult for a single template to express the action completely and its robustness is poor. Since human action can be regarded as a time-varying data sequence, the state space method can be used to model the changing law of time-varying data, which has been widely concerned. The common state space models include the hidden Markov model, dynamic Bayesian network, and conditional random field. The bag-of-words model was first applied in the field of natural language processing. The number of times a word appears in a document is counted as a document representation. Compared with the early methods, the bag-of-words model is more robust to occlusion and complex background and has achieved great success in human action recognition. Later, some scholars proposed Fisher vector (FV) for action recognition [18]. It is a coding method similar to the bag-of-words model, such as extracting SIFT features of images and constructing visual dictionaries through vector quantization. FV uses Gaussian mixture model to construct codebooks. And, FV not only stores the frequency of visual dictionaries in an image but also counts the differences between visual dictionaries and local features.

3. Dataset and Methods

3.1. Dataset Construction. There is no open dataset suitable for teachers' classroom behavior recognition at present. Through the national education resources public service platform, we select the classroom recording videos of special-grade teachers and ministry level "excellent class" from the activities of "one excellent class for one teacher, one teacher for one class" to collect the video data of teacher behavior.

Teachers' gestures in class are usually divided into three categories [23, 24] (1) indicative gesture, which refers to an object or position, usually extending a finger or a hand; (2) descriptive gesture, which enhances semantic content through the shape or motion track of the hand; (3) rhythmic gesture, a simple rhythmic action up and down, does not describe semantic content, but is consistent with the rhythm or discourse structure of speech. Some studies have shown that indicative or descriptive gestures are more conducive to learners' learning than rhythmic gestures and no gestures [25]. The classic classroom teaching analysis methods FIAS and S-T also have relevant descriptions of teacher behavior, including interpretation, demonstration, blackboard-writing, media demonstration, questioning, roll call, and inspection. Combined with the above analysis, we define six types of common teacher behaviors: blackboard-writing, questioning, displaying, instructing, describing, and nongesture behavior.

Most of the behavior recognition in the field of computer is also known as action recognition, to recognize the action after cutting, which is not in line with the actual teaching application scene. For the uncut video, action detection is to determine whether and where the action occurs in a long video, and the target action generally only accounts for a small part of the video. Therefore, action detection can be used as abnormal action detection in class, but it is different from the purpose of this study. Based on the literature review of teaching analysis, we use the FIAS analysis model to divide the video into 3s segments and label them with meaningful categories.

Therefore, there are some special cases in each video clip, such as the lack of behavior initiation or inconsistency between behavior initiation and video initiation, which is different from action recognition in the field of computer vision. Obviously, this will increase the difficulty of recognition.

Since this study focuses on teacher behavior, the videos collected take teachers as the main body and adopt the perspective of positive platform. The collected video is uniformly sized to the resolution of 640×480 , as shown in Figure 1.

3.2. The Proposed Method. The information modes of human action recognition include RGB information, depth information, and skeleton information. RGB information can represent the color and texture of objects and human body surface, which provides sufficient discriminant information for the model. However, because of its rich color information, it is easy to be affected by noise, which will interfere with the representation of action. Depth information contains three-dimensional depth information. We can also see the contour information of human body through depth image. It is like a binary RGB image, but it cannot display the apparent texture and color information. For skeleton data, it is an abstract representation of human joint combination, which can eliminate the interference of complex background but cannot represent the background information. It can be seen that the single modal feature has inherent difficulties that are difficult to overcome, and the fusion of multimodal heterogeneous features can often express more sufficient distinguishing ability. The depth information needs special shooting equipment (such as Kinect), while the ordinary camera is usually used to shoot classroom teaching video, so the modal information does not conform to the application scenario of this paper. Generally, there are two methods to obtain skeleton information, one is based on sensor, which can get more accurate skeleton; the other is image-based method, which has a lower cost and is easy to implement. According to the actual application scenario of this paper, the second method is adopted, which uses the popular OpenPose algorithm [26] to extract skeleton information. The overall research framework is shown in Figure 2.

3.2.1. Based on RGB Information

(1) Dense Sampling. To ensure the comprehensiveness of feature sampling, the dense sampling [17] method meshes the images at multiple scales and performs intensive sampling separately. Usually, eight spatial scales are used (if the image is large, it can be increased appropriately). The idea of dense sampling is to track the detected feature points in time dimension to form a feature trajectory. Since it is difficult to track the feature points in the region where the change is not obvious, it is necessary to filter the detected feature points.

The optical flow method is used to track the feature points. After sampling the coordinates of a feature point in a frame, the motion direction of the feature point is obtained by calculating the optical flow in the neighbourhood of the feature point, and then the position of the feature point in

FIGURE 1: Representations of six common teacher behaviors: (a) blackboard-writing, (b) questioning, (c) displaying, (d) instructing, (e) describing, and (f) nongesture behavior.

FIGURE 2: The overall research framework.

the next frame is calculated. This method can form a feature track on a continuous L -frame image. Due to the drift phenomenon of tracking for a long time, dense sampling and tracking are needed for every L frame. In this paper, $L = 15$ is chosen as the research object. The visualization effect of dense trajectory sampling is shown in Figure 3.

(2) *Feature Description.* For a feature point, we select the neighbourhood with the size of $n \times n$ in each frame along its trajectory to form a space-time volume. In the spatial dimension of each time space body, it is divided into w parts along each direction, while in the time dimension, t parts are evenly selected, and there are $w \times w \times t$ blocks. For each region, HOG, HOF, and MBH are used as feature descriptors to extract features. HOG calculates histogram of gradient of grey image, HOF calculates histogram of optical flow, and MBH calculates histogram of gradient of optical flow, which can be regarded as HOG feature of optical flow image.

(3) *Feature Coding.* For a video, there are a lot of trajectories, each track corresponds to a group of features, and the number of features obtained by different videos is different.

Therefore, it is necessary to encode these feature groups to obtain a fixed length coding feature for the final video classification. The common methods are bag-of-words and Fisher vector, and Fisher vector is adopted here. Fisher vector is a kind of coding method similar to the bag-of-words model, which uses the Gaussian mixture model to construct a visual dictionary, namely, codebook.

During training the codebook, we first randomly sample 200 features of each video segment to construct the codebook. Then, PCA is used to reduce the dimension of the extracted dense track features, and the dimension of each feature is halved. We use GMM to code, the number of Gaussian clustering is set to 200, and then a coding template is formed for each video segment.

(4) *Classification using SVM.* Support vector machine (SVM) is a kind of commonly used binary classification model. Its basic idea is to find the hyperplane which makes the samples have the largest space in the feature space. It is a classical linear classifier. For the multiclassification problem, the objective function can be modified directly, but the complexity is too high; the other is the indirect method, which realizes multiclassification by combining multiple two

FIGURE 3: The visualization effect of dense trajectory sampling.

classifiers. There are two kinds of indirect methods: one-versus-one and one-versus-rest. In application, the one-versus-one method needs to cost more, so it is often used for multiclassification task. In this paper, the one-versus-rest method is used in the experiment.

3.2.2. Based on Skeleton Information

(1) *Skeleton Information Extraction.* The complex background in the video will affect the behavior recognition, and the skeleton information can eliminate the interference of the background and get the more essential characteristics of the behavior. There are two commonly used skeleton information extraction methods, one is based on the sensor, which can get more accurate skeleton information; the other is based on image, which needs lower cost and is easy to implement. Therefore, we adopt the second method, using the popular OpenPose algorithm [18] to extract skeleton information.

OpenPose algorithm is an algorithm proposed by Carnegie Mellon University (CMU) based on convolutional neural network and supervised learning. It has developed an open-source library supporting the third party and is one of the commonly used skeleton extraction tools. OpenPose can extract the coordinates and confidence of 25 joint points of human body, as shown in Figure 4(a). In this paper, we only need to extract the skeleton information of teachers, but students' figure will appear in the video, which will affect the extraction of skeleton information. Therefore, we set the parameter number people max to 1 in OpenPose, and the effect is shown in Figure 4(b).

(2) *Features Construction.* There is no direct relationship between the teacher's behavior and the lower half of teacher's body, and the information of the teacher's lower

body will be missing in some pictures because of the limitation of shooting pictures. Therefore, we delete 10 joints (10, 11, 13, 14, and 19~24) of the teacher's lower body. Because the actual recorded video will be affected by camera, lighting, angle, and other factors, the detection of joint points will be missing and biased. It is necessary to detect anomalies and smooth the data. For the missing value, this paper sets it to the intermediate value between the previous frame and the next frame. If not, it is set to the average value of the joint points. Finally, 45 skeletons are uniformly sampled from each video for subsequent model training and evaluation.

3.2.3. *Fusion of RGB Information and Skeleton Information.* The features of different modal information are independent and complementary. After effective information fusion, the discriminant power of the model can be improved. The common multimodal information fusion methods generally include early fusion and late fusion. In this paper, we use the two fusion methods for experiments.

In early fusion, the independent modal information is fused into a feature vector and then the classifier is used for training and testing. It is also called feature fusion. The strength of feature fusion is that the gain of the recognition after feature fusion may be greatly improved rather than simply permutation and combination. This is because more feature information can be obtained by extracting multiple features of the target object, which may be complementary to each other. Through rule fusion, it can be more capable than the way that represents object with single feature. Therefore, to describe the target object more accurately, multifeature fusion is beneficial to improve the recognition performance. Based on this, we extracted features from RGB information

FIGURE 4: (a) 25 joint points of human body in OpenPose; (b) the effect of skeleton information extraction when setting the parameter number_people_max is 1.

and skeleton information and fused them in series, that is, RGB features and skeleton features were connected to form a longer vector as a description of teacher behavior.

While the late fusion inputs different modal information into the independent classifier, respectively, finally, the results of different classifiers are fused. The common typical representative is ensemble learning. We took stacking to our experiment. The stacking is a general method of combining individual learners with training learners, in which the individual learner is called the first-level learner, and the combiner is called the second-level learner or the meta-learner. The stacking trains a first-level learner using the original training dataset. Then, the output of the first-level learner is used as the input feature, and the corresponding original marker is used as the new marker to form a new dataset for the training of the second-level learner. Finally, the fusion results of different models are obtained.

4. Experimental Results

This section shows the experimental results of only using RGB information for recognition and fusion of RGB information and skeleton information. The main equipment of this experiment is TITAN Xp (GPU), i7-7700 (CPU). We divided the dataset into training set and test set according to the ratio of 3:1 for experiments.

4.1. Results Based on RGB Information. When the dense trajectory method is used to extract features of RGB video, it is necessary to select the appropriate feature descriptor. HOG features mainly detect the edge information of image and then represent the appearance and shape of the local target. HOF is a common feature in motion recognition, which can overcome the sensitivity of optical flow to scale and direction of motion information and can represent the temporal motion information. The effect of MBH feature is to extract the boundary information of moving objects. In this paper, the above three descriptors are tested separately and in combination. Table 1 shows the effect of different feature descriptors on teacher behavior recognition. It can be seen that the HOF effect is the worst when using the above features alone, because some video clips contain more redundant behavior, and HOF is more sensitive to different behaviors. The table shows that the highest accuracy is obtained by using both HOG and MBH, followed by the simultaneous use of three feature descriptors.

4.2. Results Based on Fusion of RGB Information and Skeleton Information. Based on the result of Section 4.1, HOG + MBH and HOG + HOF + MBH are used as candidate features for RGB information fusion, and then fusion experiments are conducted with skeleton information. Table 2 shows that the

TABLE 1: Accuracy based on RGB information.

Information mode	Feature	Accuracy (%)
RGB	HOG	72.76
	HOF	58.75
	MBH	75.49
	HOG + HOF	73.15
	HOG + MBH	78.99
	HOF + MBH	73.93
	HOG + HOF + MBH	76.65

TABLE 2: Accuracy based on the fusion of RGB information and skeleton information.

Information mode	Feature	Early fusion (%)	Late fusion (%)
RGB + skeleton	HOG + MBH + joints	82.49	79.0
	HOG + HOF + MBH + joints	83.27	76.65

effect of early fusion is better than the late-fusion effect, and it is better than using single modal information. This indicates that RGB information and skeleton information can complement each other and achieve the effect of $1 + 1 > 2$, which can better reflect the essence of teacher's behavior. However, the accuracy of late fusion in the decision-making level is not improved compared with that before fusion. Through the experiment, we can see that the best effect is to use HOG + HOF + MBH and skeleton information for early fusion, and the accuracy reaches 83.27%, which is higher than the accuracy of using both alone.

5. Conclusions

The use of human action recognition technology to automatically recognize teacher behavior can help teachers to quickly understand the self-expression in the classroom and can also accurately locate the areas of their interest, which plays an important role in reducing the burden of teachers' self-reflection and information overload. Based on this, we explored the application of human action recognition in teacher behavior recognition, mainly aiming at multimodal information fusion from RGB video recorded in class. Based on RGB video, we used gesture estimation tools to get the skeleton information of teachers and integrated it with RGB information, which improved the recognition accuracy of teacher behavior. The experimental results show that the early-fusion method can effectively improve the recognition accuracy for teacher behavior.

For the follow-up research, on the one hand, we will explore the application of deep learning methods in this field to get higher recognition accuracy; on the other hand, we will comprehensively analyse classroom behavior combined with student behavior, so as to better support teaching reflection and improve teaching effect.

Data Availability

The teacher behavior dataset used to support the findings of this study is currently under embargo while the research

findings are commercialized. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61605054), the National Key Technology Research Program of the Ministry of Science and Technology of China (2015BAK33B02), the Fundamental Research Funds for the Central Universities (CCNU19QD007 and CCNU19TD007), the Open Fund of Hubei Research Center for Educational Informationization, and the Central China Normal University (HRCEI2020F0205).

References

- [1] D. M. LeFevre, "Designing for teacher learning: videobased curriculum design," in *Using Video in Teacher Education*, J. Brophy, Ed., Emerald Group Publishing, Bingley, UK, pp. 235–258, 2004.
- [2] G. Cyrille and C. Sébastien, "Video viewing in teacher education and professional development: a literature review," *Educational Research Review*, vol. 16, pp. 41–67, 2015.
- [3] K. Hosun and E. A. Van Es, "Articulating design principles for productive use of video in preservice education," *Journal of Teacher Education*, vol. 70, no. 3, pp. 237–250, 2018.
- [4] G. A. Wright, "How does video analysis impact teacher reflection-for-action?" Doctoral Dissertation, Brigham Young University, Provo, Utah, 2008.
- [5] G. Blomberg, A. Renkl, M. Gamoran Sherin, H. Borko, and T. Seidel, "Five research-based heuristics for using video in pre-service teacher education," *Journal for Educational Research Online*, vol. 5, no. 1, pp. 90–114, 2013.
- [6] J. C. P. Chan, H. Leung, J. K. T. Tang, and T. Komura, "A virtual reality dance training system using motion capture technology," *IEEE Transactions on Learning Technologies*, vol. 4, no. 2, pp. 187–195, 2011.

- [7] C. Jing, *Kinect-Based Gesture Recognition and Applying It in Teaching*, Shanghai Jiaotong University, Shanghai, China, 2013.
- [8] J. Xu, W. Deng, and Y. Wei, "Automatic recognition of student's classroom behaviours based on human skeleton information extraction," *Modern Educational Technology*, vol. 30, no. 5, pp. 108–113, 2020.
- [9] H. A. Smith, "Nonverbal communication in teaching," *Review of Educational Research*, vol. 49, no. 4, pp. 631–672, 1979.
- [10] E. Brey and K. Pauker, "Teachers' nonverbal behaviors influence children's stereotypic beliefs," *Journal of Experimental Child Psychology*, vol. 188, Article ID 104671, 2019.
- [11] L. Wang and D. Suter, "Informative shape representations for human action recognition," in *Proceedings of 18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, China, August 2006.
- [12] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 171–178, New York, NY, USA, October 2006.
- [13] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on R transform," in *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 2007.
- [14] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.
- [15] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proceedings of the BMVC 2008—19th British Machine Vision Conference*, Leeds, UK, September 2008.
- [16] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proceedings of the Computer Vision—ECCV 2008, 10th European Conference on Computer Vision*, Marseille, France, October 2008.
- [17] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.
- [18] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, December 2013.
- [19] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.
- [20] C. Schödl, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004*, Cambridge, UK, September 2004.
- [21] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007.
- [22] R. Zhao, W. Xu, H. Su, and Q. Ji, "Bayesian hierarchical dynamic model for human action recognition," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.
- [23] Z. Pi, Y. Zhang, J. Yang, and W. Hu, "All roads lead to rome: instructors' pointing and depictive gestures in video lectures promote learning through different patterns of attention allocation," *Journal of Nonverbal Behavior*, vol. 43, no. 2, pp. 549–559, 2019.
- [24] M. W. Alibali, M. J. Nathan, M. S. Wolfgram et al., "How teachers link ideas in mathematics instruction using speech and gesture: a corpus analysis," *Cognition and Instruction*, vol. 32, no. 1, pp. 65–100, 2014.
- [25] K. Seokmin, G. L. Hallman, L. K. Son, and J. B. Black, "The different benefits from different gestures in understanding a concept," *Journal of Science Education and Technology*, vol. 22, no. 6, pp. 825–837, 2013.
- [26] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299, Honolulu, HI, USA, June 2016.