Dynamic Group Interactions in Collaborative Learning Videos

Wenjing Shi[1], Marios S. Pattichis[1], Sylvia Celedón-Pattichis[2] and Carlos LópezLeiva[2]
{wshi, pattichi, sceledon, callopez}@unm.edu
[1] image and video Processing and Communications Lab (ivpcl.unm.edu)
Dept. of Electrical and Computer Engineering
University of New Mexico, United States.
[2] Dept. of Language, Literacy, and Sociocultural Studies
University of New Mexico, United States.

***Abstract*—We introduce a new method to detect student group interactions in collaborative learning videos. We consider the following video activities: (i) human to human, (ii) human to others, and (iii) lack of any interaction. The system uses multi-dimensional AM-FM methods to detect student faces, hair, and then use the results to detect possible interactions. We use dynamic graphs to represent group interactions within each video. We tested our methods with 15 videos and achieved an 84% accuracy for students facing the camera and 76% for students facing both towards and away from the camera.**

***Index Terms*—Human activity detection; video analysis.**

I. Introduction

We study the problem of detecting group interactions in videos of students learning how to program. Here, we are interested in detecting dynamic interactions between humans (e.g., human-to-human (H2H)), humans and other objects (e.g., human looking at a monitor (H2O)), or the lack of any interaction (NI). We present an example in Fig. 1.

Our problem requires that we first develop methods for human activity recognition. Recently, video activity recognition has been dominated by neural-networks methods that require large datasets for training. For example, in [1], the authors detect human activity using a Recursive Neural Net (RNN) with a probabilistic inference model. A variety of neural network models were explored in [2]. In [2], the authors did not find significant improvements via the use of optical flow in feature extraction. In [3], the authors used Trajectory-Pooled Deep-Convolutional Descriptors (TDD) for action recognition in digital videos. By combining TDD with improved trajectories (iDT) [4], [5], they obtained an accuracy of 65.9% on the HMDB51 dataset and an accuracy of 91.5% on the UCF101 dataset. More recently, in [6], the authors developed a new approach based on the use of both spatial and temporal Convolutional Neural Networks (ConvNets) with a variety of input modalities: RGB, RGB difference, Optical Flow, Warped Flow, and different combinations. On the UCF101 dataset, they obtained 91.7% accuracy by combining Optical Flow, Warped Flow, and RGB and 93.5% using Temporal Segment Networks (Table 5 in [6]).

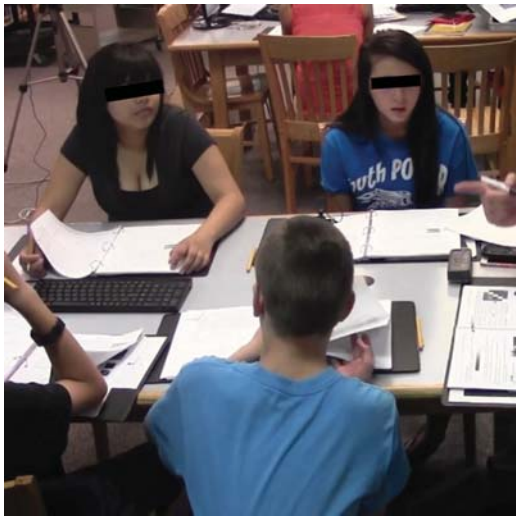

Fig. 1: Group activity interactions in a collaborative learning video. In this example, the two girls are looking to the right towards their facilitator (H2O). The young boy is looking towards the left girl but there is no interaction between them in the sense that she is not looking back at him (NI).

In a related paper by our group [7], we presented a method for detecting writing, typing, and talking using motion vectors and deep learning methods. In [7], we did not consider human interactions. Instead, our current approach relies on the use of multiscale AM-FM decomposition and intuitive video analysis methods that can be trained using small datasets. The current paper represents a significant extension over our recent work presented in [8]. In [8], we presented our results on person detection. In this paper, we establish group interactions and construct dynamic graphs that use links to describe the interaction type (H2H, H2O, NI) between the detected persons.

In the rest of our paper summary, we provide a summary of the method in Section II. We present the results in Section III and provide concluding remarks in Section IV.

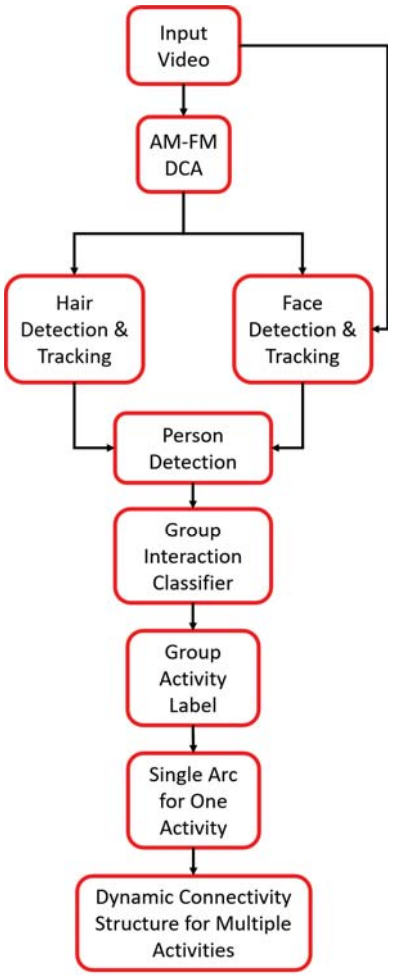

Fig. 2: Group interaction classification system.

II. Methodology

We present an overview of our method in Fig. 2. Initially, we process each video image using a Gabor filterbank to compute the dominant AM-FM component [9], [10], [8]:

$$I(x,y) \approx a(x,y)\cos\phi(x,y). \quad (1)$$

We present the 2D frequency response of the Gabor filterbank (after the application of 1D Hilbert filters) in Fig. 3(a). In Fig. 3(a), the directional filterbank is shown to tightly cover the 2D frequency plane with the ellipsoidal support elongated along each ray. The corresponding FM image in Fig. 3(b) has very clearly defined facial features that are clearly visible and do not suffer from image intensity variations.

Face detection combines the results from two face detectors. The first detector uses a simple color model based on HSV values (see [11]) to detect human skin color. The second detector processes the FM component using a simple K-NN classifier (K=3). For the FM detector, the input image is divided into 60×60 blocks with 50% overlap between

(a) (b)

Fig. 3: FM image example. (a) Frequency magnitude response of daisy-petal Gabor filterbank (54 channels), after the application of 1D Hilbert filters along each row. (b) FM component estimate.

blocks. Here, we note that K-NN classifier was trained on a small, independent training set. The final result is generated by *anding* the results from the two face detectors. We then use the detected face as a template and rely on cross-correlation to track the face through 10 video frames.

For hair detection, we combine the results from the AM and FM components. The AM detector uses a simple threshold to determine the darker regions that correspond to the hair components. We then apply a Canny edge detector to the FM component to identify sharp changes. Here, we note that the processed FM image will have strong vertical components that correspond to the hair strands (see Fig. 3(b)). To detect the hair strands, we *and* the resulting image with the AM-detector results so as to isolate the darker regions of the image. Then, we sum-up the surviving pixels along each column and identify the hair regions as the top 60 vertical sums. In other words, our vertical sums serve as hair-strand detectors. To localize the hair detector, we simply look for the highest density block within the resulting image. As for face tracking, we use the detected hair region as a template and rely on cross-correlation to track the hair through 10 video frames.

We then combine the results from the face and hair detection to detect each person in the video as described in [8]. To detect the human activities, we develop a group interaction classifier that process the resulting faces to detect where people are looking (left, right, or forward). Here, we find the centroid of each face region and use simple KNN $K = 3$ for 60×60 image regions over 2,000 training samples. Here, we note that the results were validated over an independent testing set that did not include our training samples.

In order to construct the dynamic graphs, we define nodes for each human, and also introduce a right-node for persons looking to the right, and a left-node for persons looking to the right. Here, we did not yet fully develop the methods for detecting specific objects that the students are looking at. As we shall see in the examples, the students may be looking at objects that are not seen in the video image. Then, we define dynamic links that represent links between the nodes. For each person, we first determine where they are looking

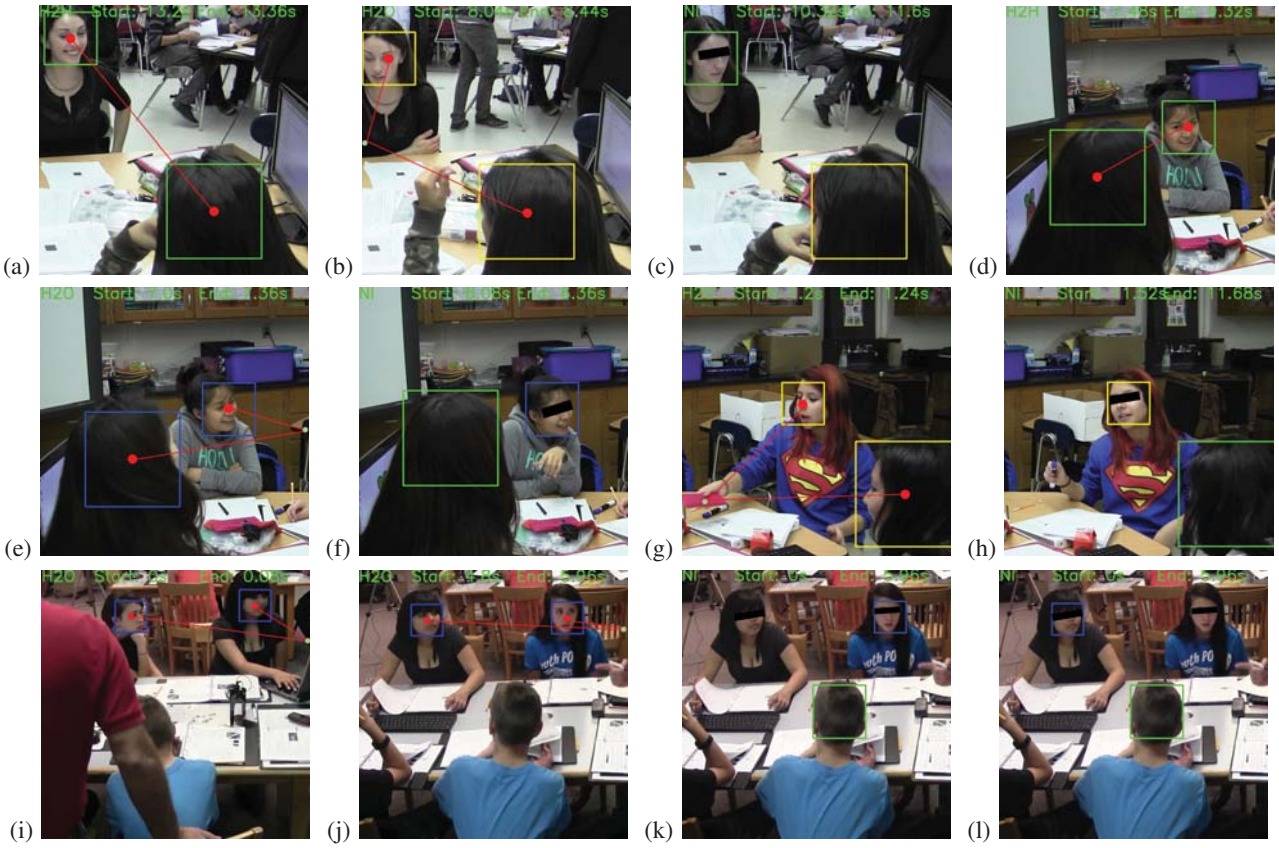

Fig. 4: Single human activity detection. In this example, we show how to construct the links (link by link) for the dynamic graphs. In each example, we only consider activities involving two humans (two nodes) at a time. Activity abbreviations: human to human (H2H), human to other (H2O), and no interaction (NI). We show bounding boxes around each detected person and a link for each detected interaction. At the top of each video frame, we provide the detected activity, the start time and end time. Our examples come from four videos. **Video 6 from Table II:** (a) Student to student interaction where one of the faces is not visible (H2H). (b) Student to other interaction with partial face occlusion (H2O). (c) Students looking away from each other (NI). **Video 8 from Table II:** (d) Student to student interaction where one of the faces is not visible (H2H). (e) Student to other interaction with partial face occlusion (H2O). (f) Students looking away from each other (NI). **Video 3 from Table II:** (g) Student to other interaction with partial face occlusion (H2O). (h) Students looking away from each other (NI). **Video 5 from Table I:** (i) Student to other (monitor) interaction (H2O). (j) Student to other interaction where faces are visible (H2O). (k) Students looking away from each other (NI). (l) Students looking away from each other (NI).

(left, right, or forward). Second, for the same person, we determine whether they are looking at another person who is looking back at them (H2H). Then, if two persons are looking at different directions, we determine that there is no interaction between them (NI). On the other hand, if any two humans are looking into the same direction, but not at another human, we have two human-to-other (H2O) links.

To make the graphs dynamic, each node is associated with a start time and and an end time and the type of human activity that it describes (H2H, H2O, or NI). Then, to fully capture group activities, we go through all of the dynamic arcs and construct a static graph of the links that are active at any given time.

III. Results

We use 15 different video segments to test our methods. We present examples of human activity detection in Fig 4.

We note the strong variations in detecting students who are facing towards and away from the camera. In many cases, we cannot even see their faces. Figures 4 (a) and (d) show student to student interaction (H2H). Figures 4 (b), (e), (g), (i) and (j) show students both focus on the same object or other person (H2O). Fig 4 (c), (f), (h), (k) and (l) present the link with no interaction.

Based on the complexities of the videos, we summarize the results in Tables II and I. For Table II, human interactions based on front-face detection. We report the results based on

the number of detected activities per frame. For this case, we have achieved an average accuracy of 84%. On the other hand, for Table I, we require both front-face and back of the head detection. For this case, the performance dropped to an average accuracy of 76%.

As described in our methodology, to establish group interactions, we will need to consider all possible links. To understand this step, we return to the example in Fig. 1. We analyze all possible pairs of interactions among humans in: (1) Fig. 4(j) for girls 1 and 2, (2) Fig. 4(k) for girl 1 and boy 1, and (3) Fig. 4(l) for girl 2 and boy 1. Note that for Figs. 4(k) and (l), we correctly label the frame as having no interaction. Thus, the only detected interaction is associated with the two girls looking to the right (H2O) as shown in Fig. 4(j).

TABLE I: Group activity detection for videos of humans facing the camera. Activity abbreviations: human to human (H2H), human to other (H2O), and no interaction (NI). Links represent human activities.

Dominant Video Activity	Time (s)	Accuracy	Persons	Links
V1: NI	16	79.3%	3	5
V2: NI	16	83.5%	3	4
V3: NI	6	83.9%	3	3
V4: H2O	20	85.6%	3	6
V5: H2O	8	86.2%	3	2
V6: H2O	6	87.3%	3	3

TABLE II: Group activity detection for videos containing a mixture of people facing the camera and people looking away from the camera (requiring back of the head detection). Activity abbreviations: human to human (H2H), human to other (H2O), and no interaction (NI). Links represent human activities.

Dominant Video Activity	Time (s)	Accuracy	Persons	Links
V1: NI	23	61.7%	2	3
V2: H2O	20	63.6%	2	3
V3: H2O	12	68.7%	2	2
V4: H2H	20	69.2%	2	3
V5: NI	12	69.4%	2	0
V6: NI	20	71.5%	2	2
V7: H2O	18	79.8%	2	2
V8: H2H	18	97.7%	2	1
V9: NI	20	98.4%	2	0

IV. Conclusion

In our paper, we presented a method for constructing dynamic graphs that describe human activities in a collaborative learning environment. To detect the activities, we introduce the use of robust human texture detection using AM-FM representations derived from a tightly constructed Gabor filterbank Currently, our system relies on human gaze detection in complex settings where a person may be facing or looking away from the camera. In ongoing work, we are considering extending our results to include object detection and also to incorporate our research in detecting other human activity (e.g., writing, typing, and talking). Furthermore, we are also considering fast implementations of the underlying filterbanks as described in [12], [13] based on the fast computation of the Discrete Periodic Radon Transform [14], [15].

V. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1613637 and No. 1842220.

References

[1] A. Richard, H. Kuehne, and J. Gall, "Weakly supervised action learning with rnn based fine-to-coarse modeling," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 3.

[2] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[3] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[4] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[5] ——, "Lear-inria submission for the thumos workshop," in *ICCV workshop on action recognition with a large number of classes*, vol. 2, no. 7, 2013, p. 8.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[7] A. Jacoby, M. S. Pattichis, S. Celedon-Pattichis, and C. LopezLeiva, "Context-sensitive human activity classification in collaborative learning environments," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, in press, 2018.

[8] W. Shi, M. S. Pattichis, S. Celedon-Pattichis, and C. LopezLeiva, "Robust head detection in collaborative learning environments using am-fm representations," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, in press, 2018.

[9] M. S. Pattichis and A. C. Bovik, "Analyzing image structure by multidimensional frequency modulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 753–766, 2007.

[10] V. Murray, P. Rodríguez, and M. S. Pattichis, "Multiscale am-fm demodulation and image reconstruction methods with improved accuracy," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1138–1152, 2010.

[11] W. Shi, "Human Attention Detection Using AM-FM Representations," Master's thesis, the University of New Mexico, Albuquerque, New Mexico, 2016.

[12] D. Llamocca and M. Pattichis, "Dynamic energy, performance, and accuracy optimization and management using automatically generated constraints for separable 2d fir filtering for digital video processing," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, vol. 7, no. 4, p. 31, 2015.

[13] C. Carranza, D. Llamocca, and M. Pattichis, "Fast 2d convolutions and cross-correlations using scalable architectures," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2230–2245, 2017.

[14] ——, "Fast and scalable computation of the forward and inverse discrete periodic radon transform," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 119–133, 2016.

[15] C. Carranza, M. Pattichis, and D. Llamocca, "Fast and parallel computation of the discrete periodic radon transform on gpus, multicore cpus and fpgas," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4158–4162.