Visualizing the Lesson Process in Active Learning Classes

Masahiro Toyoura† Mayato Sakaguchi† Xiaoyang Mao† Masanori Hanawa† Masayuki Murakami‡
† University of Yamanashi
Kofu, Yamanashi, Japan
‡ Kyoto University of Foreign Studies
Kyoto, Japan

***Abstract*—Active learning classes, which aim at increasing student participation in class, demand more management skills from the instructor than a conventional lecture class does. However, the instructor rarely recognizes how his/her lessons are different from those of others. The instructor cannot know exactly how one of his/her lessons is different from his/her previous week's lesson. This class-to-class comparison is effective in improving classes. This paper proposes a method for automatically visualizing the process and content of classes. Although there are ways to visualize the contents of classes manually, these approaches involve considerable investments of time and money. Machine learning techniques can automate the visualization. Our method estimated content with an average accuracy of 72.4%. Through our visualization, we confirmed that individual instructors use time differently from others and use their own time differently from lesson to lesson.**

***Index Terms*—video analysis; active learning; class improvement; faculty development; scene recognition; visualization.**

I. Introduction

Many educational organizations have introduced active learning classes, which aim at increasing student participation. Researchers have demonstrated that active learning classes can achieve better learning effects than the conventional style classes do.

An active learning class demands more management skills from the instructor than a conventional lecture class does. However, the instructor rarely recognizes how his/her lessons are different from those of others and only has an intuitive grasp of how one of his/her lessons is different from his/her previous week's lesson. This class-to-class comparison is effective in improving classes.

Video-recorded lessons (called as video reflection [1-3] or video ethnography [4-6]) are efficient for review purposes. Metadata, annotations, and comments on each video provide analytical resources [2, 7-9].

This paper proposes a method for recognizing and visualizing the content of lesson videos. Machine learning estimates the content of videos automatically. In this paper, we assume 5 content categories: Group work, Student presentation, Lecture, Private work, and Movement. We do not deal with estimations of more semantic-level content, which are possible via manual techniques but require considerable time and money. We designed a simple feature vector consisting of the video and its sound. Obtaining the feature vector does not require any special equipment. To visualize the content of a lesson, we employed the technique of "timelines" [10]. The timelines approach shows the start and end of each activity and the corresponding transition. Multiple paralleled lines and horizontal linkages show the entire process of the lesson. For visualizing the content of multiple lessons, mainly to compare multiple lessons, we employ histograms arranged on a matrix. Based on the results of our visualization, we can discuss the tendencies of 79 lessons in 8 subjects over 4 months through numerous videos with long running times.

Scholars have proposed many methods and tips for drawing students attention and energizing classes in active learning settings, including instructional design tools [11] for both active learning and general classes. Common tips and methods include making a clear statement of the target of the day [12], an online voting system for increasing student participation [13, 14], and approaches to coordinating group members [15].

However, it is not easy for instructors to apply these new methods and tips in real classes due to the substantial burden of trial and error involved. In particular, most of these methods and tips are for individual, isolated situations or purposes; actually organizing an effective lesson by combining the methods remains a formidable challenge in most cases. Whereas questionnaire surveys aid in reviewing lessons, it is difficult to conduct questionnaire surveys many times, and the complete results often only emerge after the lessons are finished. To review lessons, instructors can capture video and thereby accelerate the PDCA (Plan-Do-Check-Act) cycle of improving the lessons. Other instructors might provide good suggestions, as well [1, 9]. Previous works conducted manual analyses of limited numbers of lessons, considering the effort and time requirements. Our content estimation method contributes to visualizing larger volumes of lessons. One of our reasons for developing the method was the fact that researchers have never applied the recent techniques of scientific visualization or information visualization toward efficient lesson reviews, as far as we know.

Video scene analysis has become a popular theme and inspired a variety of methods, most of which deal with relatively large motions like gestures and actions [16, 17]. Meanwhile, videos capturing lessons include small motions only. Surveillance cameras also produce this type video. Recent efforts in focused fine-grained recognition [18, 19] deal with this brand of video, as well.

The remainder of the paper proceeds as follows: Section II introduces fine-grained content recognition for lesson videos. Section III describes the visualization of estimated content. Section IV describes the experiments for evaluating the proposed techniques and analyzes the results. Section V discusses possibilities for future work and concludes the paper.

II. Lesson Process Recognition from Videos

As shown in Figure 1, we installed cameras and microphones in a large room with a maximum occupancy of 80 and arranged 80 movable chairs in the room. We color-coded the chairs with 8 colors to facilitate the formation of groups.

Fig. 1. The lesson video archiving system in a classroom. Two box cameras and a fish-eye camera are installed on the ceiling. The sound from the ceiling and handy microphones are also archived.

We also took measures to facilitate lesson archiving and the recognition of lesson content. We used two box cameras (SONY SNC-EB630) to produce videos at a maximum resolution of 1920x1080 (progressive). To reduce the cost of processing time and data size, we captured videos at a resolution of 960x540. Based on the observation of the video signals, we found that the two box cameras produced almost the same videos in terms of what we needed to classify the content into 5 types. We thus decided to employ the video captured by the frontal camera only.

On the ceiling, we installed a fish-eye camera (SONY SNC-HM662). However, we cannot assume that many classrooms would equip fish-eye cameras. To ensure that we could avoid limiting the use of our proposed method, we did not use the video captured by the ceiling camera but rather employed it for sound recording purposes only. We also archived the sound from handy microphones to provide the other channel of sound. We divaricated and sent the sound to speakers and a control terminal.

In total, we employed 1 channel of video and 2 channels of sound for content recognition. All the lessons covered in this paper were given in the classroom described above.

In our method, we extract the number of pixels with frame differential in each frame of the video. We denote the number as $v(t)$ for time t. When the color of a pixel changes over a threshold, the pixel has a high probability of representing a moving object. A large $v(t)$ indicates that the activity level is high at t. $v(t)$ is a cue of content recognition.

High-level face recognition [20] or human tracking [21, 22] may provide more detailed information on the content of a lesson. While low-level features, such as $v(t)$, are generally applicable in many kinds of classrooms and robust for the occlusion between students in a crowded classroom, it would be difficult to extract high-level features correctly in classrooms.

The two channels of sound $a_1(t)$ and $a_2(t)$ come from the ceiling camera and handy microphones, respectively. The difference between $a_1(t)$ and $a_2(t)$ provides an important cue for recognizing who is speaking at a given time t. In group work and presentation settings, $a_1(t)$ is expected to be larger than $a_2(t)$. Lectures, however, are expected to produce an inverse relationship. In private work settings, the classroom is relatively silent and noisy when people are moving around. It is possible to employ natural language analysis for analyzing lessons at a semantic level, but the sound from a ceiling microphone is too noisy for proper processing, and handy microphones likewise suffer from noise in general classrooms. To analyze sound correctly, one might register technical terms for each class — a time- and labor-intensive process. We employed the power of sound only to build a simple and robust system. We normalize the powers so that the top 1% is equivalent to 1 for estimating the content.

Figure 2 is overview of the content recognition approach. We employ 1 channel of video $v(t)$ and 2 channels of sound $a_1(t)$ and $a_2(t)$ for content recognition. We categorize lessons into slots via a Bag-of-Features-based approach [23-25]. We also denote three channels of signal as $C(t)$, which we can re-define with other signal channels.

In the learning phase, we prepare videos with content tags. We assign tags manually so that each time t corresponds to only 1 of the 5 categories (*Group work*, *Student presentation*, *Lecture*, *Private work*, and *Movement*). If the situation is applicable to none of the above, we assign *Movement* to the corresponding t. There is one tag for every second of video, but our method does not require users to tag every second. An effective way of tagging is to define when the content changes by jumping the video backward and forward via shortcut keys. For one lesson, it took us approximately 15 minutes to tag 90 minutes of video. As we could understand what the instructors were talking about from the audio, there was little confusion during the tagging procedure.

Five 90-minute videos provide 27,000 (= 60 sec. $\times$ 90 min. $\times$ 5) learning samples of $C(t)$. We divide the samples into k clusters via the k-means algorithm and empirically set k to 15. Researchers have shown that arranging only a few samples in several clusters results in a good k [26], which contributes to accurate estimation. Individual clusters indicate individual situations evident at certain moments in lessons. At a given moment, the video may have active visuals and relatively high-volume sound; another moment, on the other hand, might have substantial sound but minimal motion. As we define the content for specific time ranges, we cannot determine the exact content for every moment. Thus, we build a histogram consisting of the situation of a moment and its neighboring N moment. The dimensions of the histogram correspond to the clusters of $C(t)$, and the value of a dimension of the histogram corresponds to the sample number observed within $2N + 1$ seconds. We expect the histograms of two moments to be similar when the content of the moments is

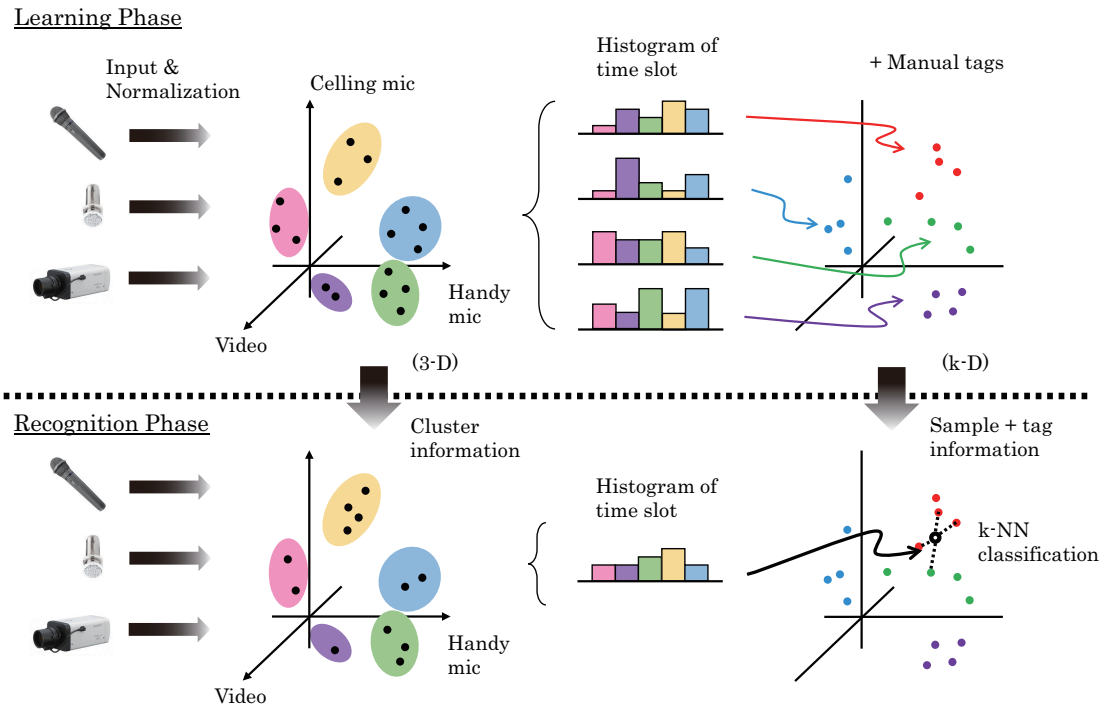


Fig. 2. Content recognition via the bag-of-features based approach for the video and sound in lessons.

comparable. Therefore, we plot the histogram-derived vector in k-dimensional space again. Every vector calculated from the learning data has a content tag, and the vectors describing the same content are expected to fall in the corresponding vicinity in k-dimensional space.

In the execution phase, we extract $C(t)$ without a manual tag from the video and the sound and plot $C(t)$ and its neighbors in 3-dimensional space. We then construct a histogram by counting the samples of k clusters and plot a $2N+1$ dimensional vector corresponding to the histogram in the space, where we have already arranged vectors with manual tags in the learning phase. k vectors with the smallest Euclidean distances from the new vector form k-nearest neighbors. We then estimate the results of voting for the tags of the k-nearest neighbors, the tag, or the content of the new vector. This is the "k-nearest neighbor" algorithm.

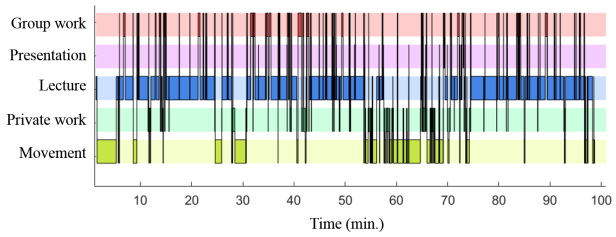
The above method generally produces comb-shaped results, as Figure 3(a) shows. Compared with the manual tags (Figure 3(b)), our approach results in very short time slots. Realistically, there should not be any 1-second activity in a given lesson. We thus introduce a voting algorithm again by assuming that the classes cannot switch between activities in such short times. Based on the tags of a second and its neighboring P seconds, we accept the highest-voted tag as the tag for the second. We empirically set the value of P to 15, which means that the content should switch within 31 (=$2P+1$) seconds. An excessively large P produces over-smoothed results, and an excessively small P creates comb shaped results. Figure 3(c) shows the results of applying the voting algorithm. The continuity of the activities is more similar to the actual manual tags than the original results were. The remaining problem of poor relationship between manual and automatic marking of the "movement" cannot be solved with the voting algorithm. It should be solved with the revision of learning method or training data.

The k-nearest neighbor algorithm takes a considerable amount of time when there are many samples in the learning phase. While a binary tree method [27] can accelerate the speed in some cases, it was not effective in our case as the number of samples is too large. We thus employ a simple random sampling that results in 1/30 of the samples remaining in the execution phase. Without the random sampling, it takes 24 hours to estimate the content of a 90-minute lesson. With the random sampling, however, the same 90-minute lesson takes only about 30 minutes to estimate. This enables us to apply the execution to a running lesson.

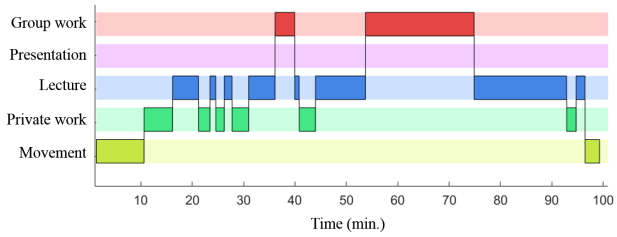
III. Visualizing Class Content

With the ability to recognize class conditions by the second, the user can (a) visualize how a given class developed (which activities the class went through) over the course of a single lesson and (b) visualize how the class changed from lesson to lesson.

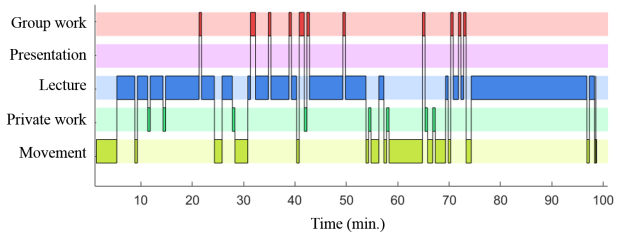
We used the Timelines approach [10], which visualizes changes in conditions over time, for our first objective (a). Figure 3 showed visualizations of class content. In addition to illustrating how much time a class spent on each activity and how long each instance of a given activity lasted, the visualizations also reveal the activities that came before and



(a) Results for individual moments



(b) Results of manual tagging



(c) Voting results of neighboring estimated tags

Fig. 3. Refinement of tags by neighboring voting.

after each instance. Putting the Timelines for 2 classes side by side allows the user to compare how the classes developed.

The Timelines approach is not particularly optimal for evaluating trends in large numbers of classes, however, as the method limits the number of items visible on a single screen. As Figure 4 shows, we first created a histogram showing the amount of time that each activity category consumed during a single class lesson. While the approach offers less renderable information than the Timelines method does, it enables users to see what the target classes focused on. We also designed our approach so that the user could display histograms for multiple classes in a matrix, making it possible to visualize the content in a list format that shows how each teacher allocates his or her class time. Our approach also gives the user insight into how a given instructor's class time distributions change on a lesson-to-lesson basis. We envision our method serving as a complement — not an exclusive alternative — to Timelines. Together, the two approaches give users an efficient, effective tool for analyzing class content. As our approach involves creating Timelines by obtaining second-by-second recognition results for each class, it takes very little processing time to generate matrix-based renderings of the histograms.

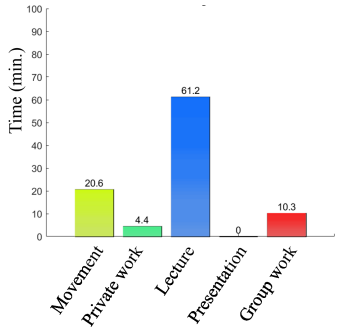


Fig. 4. A histogram of the class in Figure 3.

IV. Experimental Results

To evaluate our method, we recorded activity in a classroom with movable chairs (enabling a complete range of student placement configurations) for 4 months and visualized the development of the classes that met in the classroom. We studied 79 lessons in 8 different classes and looked at differences by instructor and lesson timing. Below is a summary of our findings.

A. Timelines-based visualization of class development

We manually tagged 5 lessons per subject to create a set of correct data. Using the jump function and other features of our video viewer, it took us approximately 15 minutes to tag 90-minute class.

Movement:
: Entering/leaving the classroom, moving to get class materials, and all other activity that does not fall into any of the following categories.

Private work:
: Tests and practice problems.

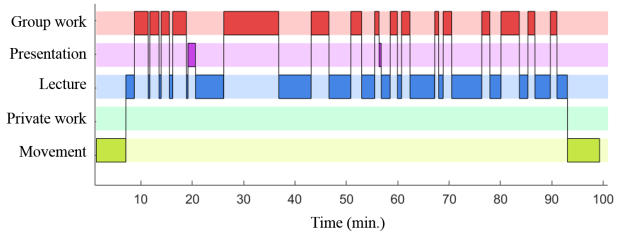
Lecture:
: Explanations and instructions from the instructor.

Student presentation:
: Explanations/answers from students.

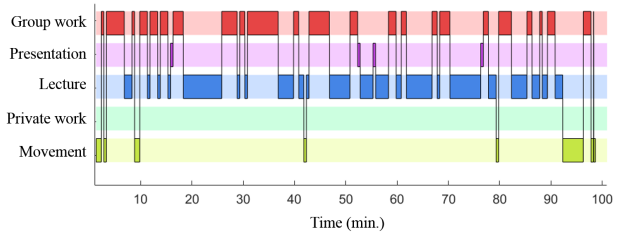
Group work:
: Work involving multiple students.

We conducted learning procedures on 4 classes and then performed Leave-one-out Cross Validation (LOOCV) [28] on 1 class. Examining the video on a second-by-second basis, we deemed each second where the manual tag and the automatic recognition matched to be a "correct" second and used the ratio of "correct" seconds to the total number of seconds in the video to calculate the corresponding "accuracy rate."

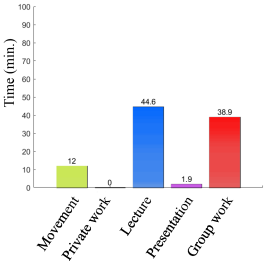
The average accuracy rate for the 5 classes was 72.4%, with a maximum accuracy of 86.8% and a minimum accuracy rate of 54.5%. The following section focuses on how successfully our estimation results captured overall trends. Figures 5, 6, and 7 show the results of automatic recognition for the 3 classes



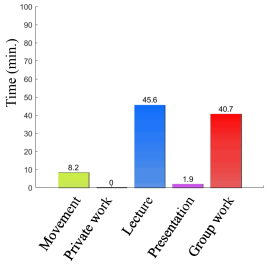
(a) Correct classification via manual tagging



(b) Classification via automatic estimation

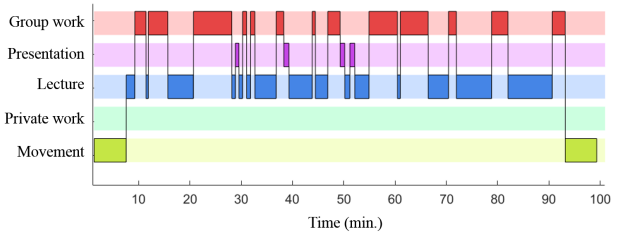


(c) Histogram of (a)

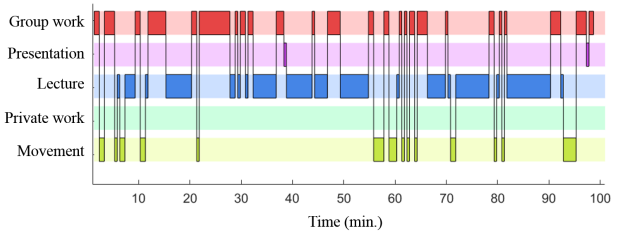


(d) Histogram of (b)

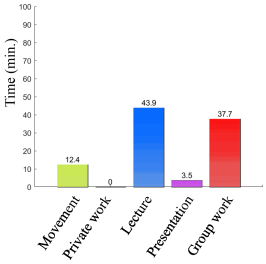
Fig. 5. Results of BoW-based automatic classification and Timelines-based visualization. (Accuracy: 76.0%)



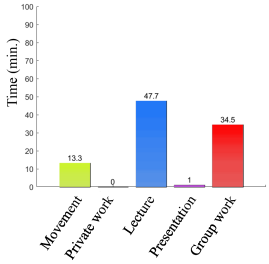
(a) Correct classification via manual tagging



(b) Classification via automatic estimation



(c) Histogram of (a)



(d) Histogram of (b)

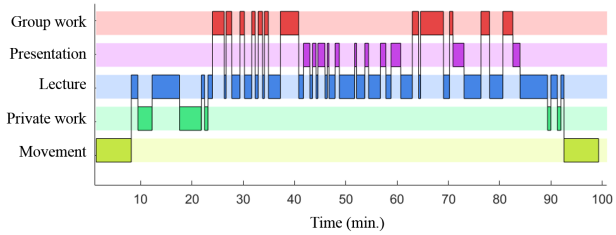
Fig. 6. Results of BoW-based automatic classification and Timelines-based visualization. (Accuracy: 70.3%)

with correct manual tags and the corresponding Timelines-based visualizations. In each Figure, (a) is the version with the correct manual tags, and (b) shows the results of the corresponding automatic recognition process. The (c) and (d) portions, meanwhile, are histograms of the activity times for (a) and (b), respectively.

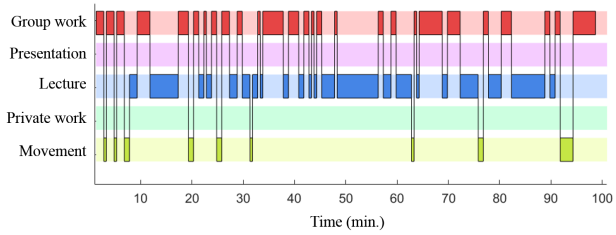
The class in Figure 5 represents a typical active learning lesson, switching back and forth between the Lecture and Group work categories at short intervals. The same trend is evident in the results of the automatic classification process, as well. While the manual tags indicate that there were substantial amounts of Movement time at the beginning and the end of the class, as students filtered in and out of the room, the automatic recognition process tended to mistake these periods of motion for Group work. The method appeared to have difficulties distinguishing the video and sound of active group discussions from the video and sound of movement. Adding specific settings for the "start of the class" and the "end of the class" would likely improve the overall recognition rate, considering that the process would be able to treat the corresponding time periods separately. The automatic recognition-based histograms capture the trends correctly, showing that the class spent roughly the same amounts of time on Lecture content and Group work content.

In Figure 6, however, one can see that the approach often misinterpreted Group work as Movement — an error that makes it impossible to express individual periods of Group work correctly via the Timelines method. To ensure that the method recognizes Movement and Group work more accurately, we will need to improve the feature quantities and classification methods involved.

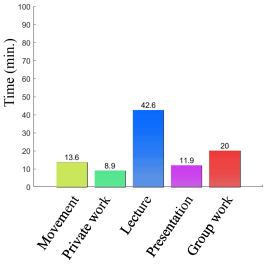
Figure 7 is an example of what an even lower accuracy rate looks like. As the Figure shows, our approach often put Private work periods into the Group work category and Student presentation periods into the Lecture category. These erroneous classifications resulted from insufficient learning sampling for Private work and Student presentations. With our system unable to guess correctly, the overall accuracy rate was relatively low. Boosting the accuracy rate would thus entail using a larger pool of learning data containing Private work and Student presentation periods. The histograms that our approach generates sometimes underestimate these periods, which could make it difficult for users to track the corresponding trends in the visualization results. The periods



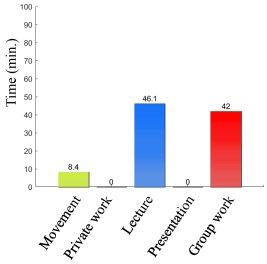
(a) Correct classification via manual tagging



(b) Classification via automatic estimation



(c) Histogram of (a)



(d) Histogram of (b)

Fig. 7. Results of BoW-based automatic classification and Timelines-based visualization. (Accuracy: 56.5%)

of Group work and Lecture activity also ended up longer than the original Movement period. Overall, the classification results could only provide a rough, basic approximation of how much time each activity type occupied.

B. Histogram matrix-based visualization of the content of multiple classes

Figure 8 is a visualization of how each of the 8 subjects developed over their 15-lesson programs. The x-axis represents the subjects, while the y-axis represents lessons 1 to 15. The histograms illustrate the time that each class spent on activity in the Movement (M), Private work (S), Lecture (L), Student presentation (P), and Group work (G) categories. For our analysis, each lesson lasted 98 minutes: 90 minutes of actual time plus 4 minutes before and after the class. The Mon 4 and Tue 3 courses were taught by the same instructor. The Wed 1 course was the only subject where all classes were taught by 3 instructors. Figures 5, 6, and 7 correspond to the 3rd, 6th, and 5th lessons of the Mon 4 subject course, respectively. We manually tagged the 1st, 2nd, 3rd, 5th, and 6th lessons of the Mon 4 subject course to create our base of learning data. The results for all of these lessons represent the automatic recognition results.

At the current stage, our method appears not to be accurate enough for certain lessons; the corresponding histogram matrices may thus lead to erroneous interpretations of class trends. Considering the imperfections in the method, users should treat the matrix-based histogram renderings as contextual aidesnot definitive visualizationsand validate the results by checking the source video for confirmation. Improving the overall accuracy rate is one of the key issues for future work.

The lessons with no histograms either took place outside our experiment classroom or were unfilmable due to scheduling or other factors. Many instructors used lessons 8 and 15 to administer written mid-term and final tests, respectively, which meant that several groups moved to other classrooms with fixed seating for those two weeks.

Optimized for active learning with several small whiteboards for facilitating group discussions and projectors capable of displaying content on all 4 walls, the classroom we used for our study is a frequent class setting for teachers who prefer the active learning approaches and employ various active learning techniques. Despite the environmental conditions, however, most of the classes spent a majority of their class time on instructor lectures.

The data shows that Group work occupied a significant proportion of the classes in the Tue 1 subject course. Lessons 7, 11, 12, and 13 devoted a particularly substantial amount of time to Group work. Lesson 6, however, spent little time on Group work and instead of consisted primarily of Student presentations.

Lessons 12 and 13 of the Wed 1 subject course had specifically large amounts of Movement time. A look at the source video reveals a different picture, though: lessons 12 and 13 actually included long stretches of Private work activity, with the instructor walking around the class to check on the students. The method misinterpreted this period of activity as movement. The recognition problems probably stemmed from issues with the learning data for the Movement category, which contained pre- and post-class periods when the classroom was completely empty. When our system encountered lessons 12 and 13, the generally minimal movement and low sound pressure likely led the system to recognize the conditions as similar to the silent, empty-room conditions of the pre- and post-class periods in the learning data. At the current stage, the accuracy rates for special situations lacking sufficient learning data tend to be low. Getting accurate readings thus currently requires the user to check both the histogram and the video itself. Tweaking the system to process the pre- and post-class lesson periods properly could also be one way of boosting overall recognition rates.

Although the Mon 4 and Tue 3 subject courses had the same instructor, the two courses exhibited different trends: the Mon 4 course emphasized Group work, while the Tue 3 course trended more toward the Lecture category. The differences in the two courses may have had something to do with the target students. Whereas the Mon 4 subject course consisted

Fig. 8. A histogram of automatically recognized class content in a matrix rendering; we manually tagged the classes outlined in yellow and used the tags as our learning data.

primarily of 2nd-year students, thereby allowing for a broader range of activities, the instructor probably spent more time on lectures in the Tue 3 course because most of the students in the class were in their 1st years of study. The 2 classes generally included a significant amount of Movement, and the instructor also often had students work in groups and submit *minute paper* [27].

While the Mon 3 and Fri 2 subject courses exhibited consistent time allocation trends from lesson to lesson, other courses allocated time differently every lesson. These transi-

tions suggest that instructors either modified their time usage as the class progressed or gradually worked to find the optimal class style through trial and error.

The data for lesson 14 reveals that all the classes followed a similar time management pattern as the students prepared for their respective final examinations in lesson 15.

V. Conclusion and Issues for Future Works

For this study, which focused on active learning classes, we used mechanical learning to automate the content classification process and then combined the Timelines approach and matrix-based histogram renderings to visualize the class content. Our content estimation method had an average accuracy rate of 72.4%. By visualizing class content, our method makes it possible for users to get a look at what goes into 90-minute classes and establish a context for examining trends across multiple classes.

Improving the overall accuracy rate is one of the key issues for future work. For about half of the classes we studied, we had the instructors wear watch-type sensors (EPSON Wristable GPS SF-810) to measure and record the instructors' heart rates and walking speeds. Although we did not incorporate the readings from the wearable technology into our analysis for the present paper, we believe that the data could aid in differentiating between Lecture and Student presentation periods — two activity categories that the system had trouble recognizing correctly. The reason why we could only obtain data for approximately 50% of the classes was that the instructors either forgot to wear the devices or intentionally chose not to wear them amid the hectic conditions that often characterize the start of a given class. Given the issues with the wearable devices, future research will thus also need to focus on finding ways of simplifying complicated tasks and improving accuracy rates via video and sound only.

Acknowledgment

This work was partially supported by JSPS Grants-in-Aid for Scientific Research KAKENHI (16K12784, 26282062), and Artificial Intelligence Research Promotion Foundation (25AI61-10).

References

[1] A. M. Pellegrino and B. L. Gerber, "Teacher Reflection through Video-Recording Analysis," Georgia Educational Researcher, vol. 9, p. 1, 2012.

[2] P. J. Rich and M. Hannafin, "Video Annotation Tools Technologies to Scaffold, Structure, and Transform Teacher Reflection," Journal of Teacher Education, vol. 60, pp. 52–67, 2009.

[3] T. Tripp and P. Rich, "Using Video to Analyze One's Own Teaching," British Journal of Educational Technology, vol. 43, pp. 678–704, 2012.

[4] R. C. Harris, S. Pinnegar, and A. Teemant, "The Case for Hypermedia Video Ethnographies: Designing a New Class of Case Studies that Challenge Teaching Practice," Journal of Technology and Teacher Education, vol. 13, p. 141, 2005.

[5] K. Krammer, N. Ratzka, E. Klieme, F. Lipowsky, C. Pauli, and K. Reusser, "Learning with Classroom Video: Conception and First Results of an Online Teacher-Training Program," ZDM, vol. 38, pp. 422–432, 2006.

[6] A. Piki, "Post-Implementation Evaluation of Collaborative Technology: a Case Study in Business Education," The Electronic Journal of Information Systems Evaluation, vol. 13, pp. 77–86, 2010.

[7] A. Fathi, M. F. Balcan, X. Ren, and J. M. Rehg, "Combining Self Training and Active Learning for Video Segmentation," in Eds. Jesse Hoey, Stephen McKenna and Emanuele Trucco, In Proceedings of the British Machine Vision Conference(BMVC), pp. 78.1–78.11, 2011.

[8] S. Chandra, "Lecture Video Capture for the Masses," ACM SIGCSE Bulletin, vol. 39, pp. 276–280, 2007.

[9] J. A. Paro, R. Nazareli, A. Gurjala, A. Berger, and G. K. Lee, "Video-based Self-Review: Comparing Google Glass and GoPro technologies," Ann Plast Surg, vol. 74, Suppl 1, pp. S71-4, 2015.

[10] R. N. Owen, R. M. Baecker, and B. Harrison, "Timelines, a Tool for the Gathering, Coding and Analysis of Usability Data," in Conference Companion on Human Factors in Computing Systems, pp. 7–8, 1994.

[11] B. G. Davis, Tools for Teaching: John Wiley & Sons, 2009.

[12] D. A. Bligh, What's the Use of Lectures?: Intellect books, 1998.

[13] S. J̀arvelä, P. Näykki, J. Laru, and T. Luokkanen, "Structuring and Regulating Collaborative Learning in Higher Education with Wireless Networks and Mobile Tools," Educational Technology & Society, vol. 10, pp. 71–79, 2007.

[14] K. Moss and M. Crowley, "Effective Learning in Science: The Use of Personal Response Systems with a Wide Range of Audiences," Computers & Education, vol. 56, pp. 36–43, 2011.

[15] C. C. Bonwell and J. A. Eison, Active Learning: Creating Excitement in the Classroom. ASHE-ERIC Higher Education Reports, 1991.

[16] I. Laptev, "On Space-time Interest Points," International Journal of Computer Vision, vol. 64, pp. 107–123, 2005.

[17] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8, 2008.

[18] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal Localization of Fine-grained Actions in Videos by Domain Transfer from Web Images," in ACM Multimedia, pp. 371–380, 2015.

[19] D. Hall and P. Perona, "Fine-grained Classification of Pedestrians in Video: Benchmark and State of the Art," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5482–5491, 2015.

[20] P. Viola and M. J. Jones, "Robust Real-time Face Detection," International Journal of Computer Vision, vol. 57, pp. 137–154, 2004.

[21] G. Shu, "Human Detection, Tracking and Segmentation in Surveillance Video," University of Central Florida Orlando, 2014.

[22] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," IEEE Computer Vision and Pattern Recognition (CVPR), pp. 886–893, 2005.

[23] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," IEEE International Conference on Computer Vision (ICCV), pp. 1470–1477, 2003.

[24] M. M. Ullah, S. N. Parizi, and I. Laptev, "Improving Bag-of-Features Action Recognition with Non-local Cues," BMVC, pp. 95.1–95.11, 2010.

[25] J. Baber, S. i. Satoh, N. Afzulpurkar, and C. Keatmanee, "Bag of Visual Words Model for Videos Segmentation into Scenes," International Conference on Internet Multimedia Computing and Service, pp. 191–194, 2013.

[26] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," European Conference on Computer Vision (ECCV), pp.490–503, 2006.

[27] J. L. Faust and D. R. Paulson, "Active Learning in the College Classroom," Journal on Excellence in College Teaching, vol. 9, pp. 3–24, 1998.

[28] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," IJCAI, pp. 1137–1145, 1995.