Ranking Lecture Slides for E-Book Preview Recommendation

CHRISTOPHER C.Y. YANG[†1] GÖKHAN AKÇAPINAR [†2]
BRENDAN FLANAGAN [†3] HIROAKI OGATA[†4]

Abstract: In this paper, we propose a machine learning class probabilities ranking method for ranking lecture slides from the original learning material automatically. The proposed method ranks all the lecture slides by the class probabilities retrieved from machine learning models. The top-ranked lecture slides are selected accordingly to form the recommendation of e-book preview before the starting of the lecture. The proposed method utilizes text processing and image processing techniques to extract image features and text features from lecture slide contents individually. Moreover, in accordance with the rapid developments and uses of e-book systems for the learning supports, we are able to record and collect students' reading events through the database of e-book system. The e-book usage features are therefore extracted from the previous slide-related reading events in this paper. We train and compare several well-known machine learning classification models using the extracted features and investigate the optimal model. In this paper, we evaluate and compare the ranking performance of each model under two different conditions which are previous e-book usage features exclusion and previous e-book usage features inclusion. Based on the evaluations, we discuss the performances of slide ranking under two different conditions. The statistical results suggested to us that the proposed method can be used for the automatic e-book preview recommendation potentially.

Keywords: Lecture slide ranking, e-book preview, class probability, machine learning.

1. Introduction

As the increasing amount of lecture slides were generated for supporting class activities in higher educational institutions in recent decades, lecture slides have been an effective and popular means to deliver information and knowledges throughout the lecture [17]. Ausubel [1] emphasized the importance of providing a preview of information to be learned in advance. Additionally, Beichner [3] reported that adequate preparation prior to lectures leads to improved student performances. Accordingly, using lecture slides to provide students with the preview learning contents in advance of a class are an important task for lecturers. However, students attention spans are often limited [15], resulting in the low completion rate of previewing learning contents in advance of a class. Moreover, it has been reported that most of the students in higher educational institution would prefer a summarized preview material rather than the given original slide content [15]. Therefore, recent educational data mining and machine learning techniques can help with this situation, collecting data and extracting features from various types of learning source. By using educational data mining and machine learning techniques, we are able to predict important lecture slides from the original full set of learning materials and recommend the set of lecture slides for students to preview in advance of the lecture. The characteristics of important slides/pages and associated content features have been mentioned and used in several articles [9,15], where an important slide/page is considered to be associated with the following characteristics:

1. Sufficient content to be worth browsing
2. Unique visual content
3. Keywords that appear frequently in a page
4. Keywords that rarely appear throughout the whole learning material
5. Similar to the title of lecture
6. Similar to other pages

Additionally, in accordance with the rapid developments and uses of e-book systems for the personalized learning supports [4], educators are nowadays able to collect students' reading events from the corresponding learning behaviors through e-book reading system. Regarding this prior condition, in this paper, we assume that the collected previous reading events that related to the lecture slides including how many annotations created on a slide or how much time students spent on browsing a slide during the past semesters, should be taken into account when considering recommending lecture slides for students as well. To mention more specifically, our assumption is that the more students engaged on browsing one lecture slide during the past semesters, the more chances lecturers would like to recommend the same lecture slide to students in the current semester. Therefore in this paper, we propose a novel method using text processing, image processing, and machine learning techniques for ranking lecture slides from the original learning material automatically. The proposed method aims to contribute on not only generating the recommendation of e-book preview but also keeping the essential information to be learned adequately in advance of a class.

2. Related Works

Automatic summarization tasks have been considered for several decades for generating short summaries in various of content sources. Most of the summarization methods have been focused on video summarization [6,13] and text/document summarization [5,8]. In the meantime, along with the rapid development of machine learning techniques in recent years, many of the automatic summarization methods have therefore turned to be established based on machine learning theory

†1 Hitachi Ltd.
†2 Kyoto University
†3 Nara Institute of Science and Technology

accordingly. In the domain of video summarization, several works have been proposed to automatically create short summaries of video shots by predicting shot importance through supervised learning. Shot importance was measured with a pre-trained topic-specific binary SVM classifier [12] or a SVM ranker [17]. In addition, a hierarchical model was trained to generate a video summary that contains objects of interests by using a small number of labels [7]. In the domain of machine learning based text/document summarization, a system has been proposed using Support Vector Regression (SVR) to automatically generate presentation slides to represent the summaries of scientific papers by predicting the regression score of each sentence in a supervised learning process [18].

Shimada et al. [15] applied text processing and image processing techniques to generate a summarization of lecture slides by scoring each slide in the original learning material for the enhanced students' preview, which is the most similar work to this paper. In their method, they mentioned that they have to ask the lecturers in advance to prior specify appropriate browsing time on each slide so they could be able to combine the browsing time and the extracted word importance and visual importance for the calculation of importance score and generated the summaries of lecture slides afterward. However, this kind of work seems to put additional burdens on the lecturers. In addition, this work did not take e-book users' previous reading events into account where they just considered learning content analysis for lecture slide summary generation.

In contrast to the previous works above, lecture slides are the target of the proposed method instead of considering text content or image content only. Besides, a supervised machine learning method is applied to lecture slide content analysis in this paper. As it has been reported that extracting textual features only might not be sufficiently informative when considering lecture slide content analysis [16], as well as to follow the associated characteristics above, the proposed method covers various types of information, extracting not only text features but also image features from slide contents. In addition, the aggregated e-book usage features from the slide-related previous reading events in the database of e-book system are extracted in this study as well. By training a supervised machine learning models and selecting top-ranked lecture slides based on the retrieved class probabilities, the recommendation of e-book preview is generated without asking lecturers to additionally specify appropriate browsing time before starting the ranking of lecture slides through machine learning models.

In order to make contribution of understanding the features and prediction model that can be used for ranking lecture slides through machine learning techniques, we aim to answer the following 2 research questions in the present study:

1. What features can be used for ranking lecture slides?
2. What is the optimal model for ranking lecture slides?

3. Machine Learning Class Probabilities

Ranking Method

3.1 Overview

The proposed method aims to rank lecture slides by class probabilities and accordingly select the top-ranked lecture slides to generate the recommendation of e-book preview in advance of the lecture. The proposed method provides a flexible range of lecture slides selection for the class instructors as the length of e-book preview recommendation can be generated differently based on the slide selection from the ranking.

Figure 1 shows an overview of the proposed method. Firstly, we collect different types of data from both lecture material slides uploaded to e-book system, and lecture slides associated previous e-book reading events recorded in the database of e-book system as the input data in the proposed method. Secondly, we extract text features, image features and the aggregated e-book usage features from slide contents and the slide-related previous reading events, respectively. We preprocess the extracted features by using Z-score normalization. We then train several well-known supervised machine learning models through the selected features. In the process of model training, we retrieve class probabilities predicted by the trained models and rank the input lecture slides accordingly. In this paper, the retrieved class probability can be considered as the chance of a slide to be classified as "recommended slide" according to the training data and gold-standard (recommendation from the lecturer) in the process of model training. Lastly, by selecting the top-ranked slides in a preference range from the ranking, the recommendation of e-book preview will be generated as the output in this method.

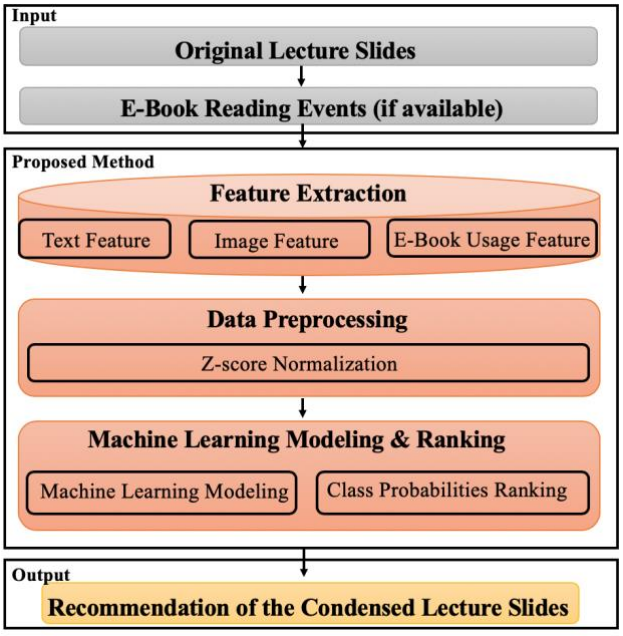

Figure 1. Overview of the proposed method.

3.2 Data Collection, Feature Extraction, and Data Preprocessing

In this paper, we use BookRoll system [10] as our data collection resource. BookRoll is a digital textbook reading system with plenty of functions in it such as annotation creation, annotation transfer across e-book revisions [20,21], internal learning contents searching, page jumping, etc.

Two types of data are collected from BookRoll as the input of the proposed method. The first type of data is text contents and image contents in original lecture slides that uploaded to BookRoll (79 lecture slides were collected in this paper) while the second type of data is the recorded previous e-book reading events that related to the same lecture slides (See Figure 2 and Figure 3 for demonstrations). The data collection process contains two results, one is lecture slide contents collection only if the previous e-book reading event collection is not available, while the another result is both lecture slide contents and the previous e-book reading events collection if available. We extract 3 categories of feature from the collected data based on the characteristics of important pages and our assumption. The first category is text features, the second category is image features, while the third category is e-book usage features.

In order to apply supervised machine learning model for the class probabilities prediction, we preprocess the extracted features by using Z-score normalization method.

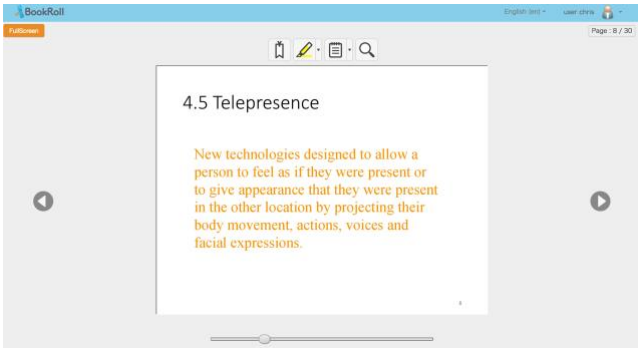

Figure 2. Example of lecture slide.

User ID	Operation Name	Operation Date	Memo Text	Marker Text	Contents ID	Page Number
xxxxx	OPEN	2017/11/22 6:01:18 AM	NULL	NULL	Content1	1
xxxxx	NEXT	2017/11/22 6:01:33 AM	NULL	NULL	Content1	1
xxxxx	NEXT	2017/11/22 6:03:02 AM	NULL	NULL	Content1	2
xxxxx	PREV	2017/11/22 6:03:24 AM	NULL	NULL	Content1	3
xxxxx	ADD MARKER	2017/11/22 6:03:28 AM	NULL	Test marker	Content1	2
xxxxx	ADD MEMO	2017/11/22 6:03:31 AM	Test memo	NULL	Content1	2
xxxxx	CLOSE	2017/11/22 6:11:36 AM	NULL	NULL	Content1	2

Figure 3. Example of reading events in BookRoll.

3.2.1 Text Feature Extraction

We applied text processing techniques including N-gram tokenizer, cosine similarity, and lemmatization to extract text features from the collected lecture slides. We generated a corpus to store text contents from lecture slides. After normalizing words in the generated corpus using stop-words removal, N-gram tokenizer and lemmatization method, we applied cosine similarity measure to calculate text similarities between each pairwise lecture slides, similarities to the title and keywords of the lecture, representing values of *Similarity to title*, *Similarity to keywords*, and *Page-Page cohesion*.

We counted the total number of characters and punctuations in a slide from the generated corpus to represent the values of *TotalChar* and *Punctuation*. We applied vector space modeling technique TFIDF weighting method to calculate the weight of each term on the slide and calculated the average of TFIDF value in each slide, representing the value of *AvgTFIDF*. By using the weights of the terms, features in the slide like sentences and tables are weighted. The details of TFIDF weighting methods have been described in several prior studies [14,19]. The extracted text features are given in Table 1.

Table 1. Description of text features (N=79)

Feature	Feature description
TotalChar	Total characters in slide
AvgTFIDF	Average of TFIDF values preprocessed by bigram tokenizer
Similarity to title	Cosine similarity to the title of lecture
Similarity to keywords	Cosine similarity to the keywords of lecture
Page-Page cohesion	Sum of cosine similarities to the rest slides
Punctuation	Total occurrence of punctuations in slide

3.2.2 Image Feature Extraction

We applied image processing techniques including background subtraction method based on background modeling strategy [22], and inter-frame difference method which both have been used for lecture slide summarization task to extract visual information as image features from the lecture slides [15]. According to the summarization method proposed in Shimada et al, the background subtraction technique extracts the foreground mask from each slide by subtracting background image pixels from each slide, followed by binarization processing. The slide content volume is then estimated by counting the total foreground pixels on the target slide, representing the value of *Background subtraction*. The inter-frame difference technique reveals changes between successive lecture slides. The subtracted image is also binarized to calculate a difference score, representing the value of *Background subtraction + Inter-frame difference* on the current slide.

Same as their method, in this paper, we performed the inter-frame difference calculation in both directions, between the current slide and both the previous slide and the next slide in the lecture material. The larger number of extracted pixels were then chose to represent the difference score on the current slide. This process supports the system on extracting slides that are significantly different from other neighboring slides. The extracted image features are given in Table 2.

Table 2. Description of image features (N=79)

Feature	Feature description
Background subtraction	Foreground pixels in slide
Background subtraction + Inter-frame difference	Absolute foreground pixel differences with previous slide and next slide (choose the higher value)

3.2.3 E-Book Usage Feature Extraction

We collected the previous reading events that related to the lecture slides from the database of BookRoll. We aggregated and extracted the collected reading events including how many

markers, memos, bookmarks, and reading events were created, representing the values of *Marker*, *Memo*, *Bookmark*, *TotalEvent*, *AvgEvent*. Moreover, how many students visited slides, how much time students spent on browsing learning contents in each slide are aggregated as well, representing the values of *UniqueVisit*, *TotalTime*, *AvgTime*. The extracted e-book usage features are given in Table 3.

Table 3. Description of e-book usage features (N=79)

Feature	Feature description
Marker	Total number of marker added in slide
Memo	Total number of memo added in slide
Bookmark	Total number of bookmark added in slide
UniqueVisit	Total students visit slide
TotalTime	Total browsing time in slide (t < 300 s)
AvgTime	Average browsing time in slide per student (t < 300 s)
TotalEvent	Total clicking events in slide
AvgEvent	Average clicking events in slide per student

3.3 Supervised Machine Learning Modeling and Lecture Slide Ranking

To generate the gold-standard of slides as the labels in binary classification task, we first asked the lecturer to determine whether a slide should be recommended to e-book users or not, and labeled each slide afterward. Due to the imbalanced proportion of "recommended slide" and "not recommended slide" (17 slides were labeled as "recommended slide" while 62 slides were labeled as "not recommended slide") in the given gold-standard, we applied resampling technique *Smote + Tomek* [2] to address the occurrence of bias in the process of model training.

In this paper, we train and compare several well-known machine learning classification models from the libraries in python [11] including *Neural Network*, *Gradient Boosting*, *Gaussian Naïve Bayes*, and *Logistic Regression* for the class probabilities prediction in the process of model training. We then rank the input slides by the predicted class probabilities.

3.4 Generation of E-Book Preview Recommendation

After the ranking process, the top-ranked slides are selected to form the recommendation for students to preview e-book slides in advance of class. The length of the e-book preview recommendation can be specified based on lecturers' preferences.

4. Evaluation

To evaluate the performance of the proposed method, in this paper, we first selected the top-ranked 17 slides from each model where these slides are the recommendations from the lecturer in the present learning material. After top-ranked slides selection, we use metrics including precision, recall, Area Under the Curve (AUC) and overall accuracy derived from the confusion matrix to evaluate the performances where greater values of these metrics indicate better performances of slide ranking. The results of ranking are validated with 6-folds cross-validation to ensure a generalized evaluation for the proposed method.

In order to test the performances of ranking models under different conditions, we conducted two experiments. For example, if we do not have related previous e-book reading events for feature extraction, e-book usage features will be excluded in the process of model training. On the other hand, if we are able to collect related previous reading events, e-book usage features will be included while training a ranking model.

5. Results

In this section, we give the statistical results of the ranking performances from each machine learning model under two different conditions.

5.1 E-Book Usage Feature Exclusion

Table 4 shows the statistical results of ranking performance when excluding e-book usage features in model training. As shown in Table 4, *Logistic Regression* is the optimal model for ranking lecture slides for e-book preview recommendation when excluding e-book usage features with precision 0.65, recall 0.65, accuracy 0.85 and Area Under the Curve 0.78.

Table 4. Statistical results of ranking performance when excluding e-book usage features in model training

Model	Precision	Recall	Accuracy	AUC
Neural Network	0.53	0.53	0.80	0.70
Gradient Boosting	0.59	0.59	0.82	0.74
Gaussian Naïve Bayes	0.41	0.41	0.75	0.63
Logistic Regression	**0.65**	**0.65**	**0.85**	**0.78**

5.2 E-Book Usage Feature Inclusion

Table 5 shows the statistical results of ranking performance when including e-book usage features in model training. As shown in Table 5, *Neural Network* and *Gradient Boosting* are the optimal models for ranking lecture slides for e-book preview recommendation when including e-book usage features with precision 0.59, recall 0.59, accuracy 0.82 and Area Under the Curve 0.74.

Table 5. Statistical results of ranking performance when including e-book usage features in model training

Model	Precision	Recall	Accuracy	AUC
Neural Network	**0.59**	**0.59**	**0.82**	**0.74**
Gradient Boosting	**0.59**	**0.59**	**0.82**	**0.74**
Gaussian Naïve Bayes	0.53	0.53	0.80	0.70
Logistic Regression	0.47	0.47	0.77	0.66

6. Discussions

According to the statistical results shown in Table 4 and Table 5, when excluding e-book usage features in the process of model training, the optimal ranking model for the proposed method is *Logistic Regression* with accuracy 0.85 and Area Under the Curve 0.78. On the other hand, when including e-book usage features in the process of model training, the optimal ranking models for the proposed method are *Neural Network* and *Gradient Boosting* with accuracy 0.82 and Area Under the Curve 0.74. In addition, we compared the performances under two different conditions shown in Table 4 and Table 5. The comparison results showed us that the overall performance of the ranking model would not be better when including e-book usage features, which is not our expected result. We hypothesize that the reason of this result could be the weak correlation between the training data of e-book usage that we used (one dataset, 79 samples) and the recommendation from the lecturers in this study. Therefore, the extracted e-book usage features could not strongly affect the recommendation result and we need to use more set of samples to investigate the deeper correlation between e-book usage features and the recommendation from the lecturers. Nevertheless, for the other 3 models *Neural Network*, *Gradient Boosting*, and *Gaussian Naïve Bayes*, the performances were better or kept as the same when including e-book usage features in model training compared to model itself.

7. Conclusions

In the paper, we proposed a machine learning class probabilities ranking method for the automatic recommendation of the condensed set of lecture slides. In order to answer research question 1, we extracted features that can be used for the proposed method by analyzing text and image contents from lecture slides and the related previous reading events from the database of the e-book system. We compared several classification models from the libraries of python. In order to answer research question 2, we evaluated the performances of slides ranking by using metrics such as overall accuracy and Area Under the Curve (AUC) under two conditions which are e-book usage feature exclusion and inclusion. The statistical results shown in Table 4 and Table 5 suggested to us that when excluding e-book usage features in the process of model training, the optimal model for ranking lecture slides was *Logistic Regression*. On the other hand, when including e-book usage features, the optimal models were *Neural Network* and *Gradient Boosting*. The proposed method in this paper can be used to automatically recommend e-book users several lecture slides in advance of the class for the adaptive e-book preview. Moreover, this paper showed different type of features that should be concentrated when considering the recommendation of e-book preview, including text features and image features extracted from learning slide contents themselves, and e-book usage feature from the previous usage of learning materials recorded in the database of e-book system.

To mention the limitation and future work in this study, we only used one set of sample which might not be enough for these kind of performance generalizations. For example, different lecturer may has his or her own preference of giving set of slides as e-book preview recommendation. In other words, using only one set of sample is difficult to generalize the performances of slide ranking as the input lecture slides come from many different lecturers actually. In the future we will look for more training samples from lectures in different research domains, to create an optimal lecturer model and rank different input lecture slides using corresponding training data based on the lecturer model.

Reference

[1] Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of educational psychology*, *51*(5), 267.

[2] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.

[3] Beichner, R. J. (1995). Improving the effectiveness of large-attendance lectures with animation-rich lecture notes. *AAPT Announcer*, *20*(917).

[4] Fletcher, G., Schaffhauser, D., & Levin, D. (2012). Out of Print: Reimagining the K-12 Textbook in a Digital Age. *State Educational Technology Directors Association.*

[5] Gupta, S., Nenkova, A., & Jurafsky, D. (2007, June). Measuring importance and query relevance in topic-focused multi-document summarization. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 193-196). Association for Computational Linguistics.

[6] Gygli, M., Grabner, H., & Van Gool, L. (2015). Video summarization by learning submodular mixtures of objectives. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3090-3098).

[7] Liu, D., Hua, G., & Chen, T. (2010). A hierarchical visual model for video object summarization. IEEE transactions on pattern analysis and machine intelligence, 32(12), 2178-2190.

[8] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

[9] Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002, November). Automatic text summarization using a machine learning approach. In *Brazilian Symposium on Artificial Intelligence* (pp. 205-215). Springer, Berlin, Heidelberg.

[10] Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In International Conference on Computer in Education (ICCE 2015) (pp. 401-406).

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

[12] Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014, September). Category-specific video summarization. In *European conference on computer vision* (pp. 540-555). Springer, Cham.

[13] Rajendra, S. P., & Keshaveni, N. (2014). A survey of automatic video summarization techniques. *International Journal of Electronics, Electrical and Computational System*, *2*(1).

[14] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

[15] Shimada, A., Okubo, F., Yin, C., & Ogata, H. (2018). Automatic Summarization of Lecture Slides for Enhanced Student Preview–Technical Report and User Study–. *IEEE Transactions on Learning Technologies*, *11*(2), 165-178.

[16] Shimada, A., Okubo, F., Yin, C. J., Oi, M., Kojima, K., Yamada, M., & Ogata, H. (2015, January). Analysis of preview behavior in E-book system. In *Paper presented at the 23rd international conference on computers in education (ICCE 2015), Hangzhou, China*.

[17] Sun, M., Farhadi, A., & Seitz, S. (2014, September). Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision* (pp. 787-802). Springer, Cham.

[18] Syamili, S., & Abraham, A. (2017, March). Presentation slides generation from scientific papers using support vector regression. In *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on* (pp. 286-291). IEEE.

[19] Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 13.

[20] Yang, C., Flanagan, B., Akcapinar, G., & Ogata, H. (2018). Maintaining reading experience continuity across e-book revisions. *Research and practice in technology enhanced learning*, *13*(1), 24.

[21] Yang, C. C.-Y., Flanagan, B., & Ogata, H. (2018). Connecting learning footprints across versions within e-book reader. In The 32nd annual conference of the Japanese society for artificial intelligence, 2018 (pp. 1-4).

[22] Zivkovic, Z., & Van Der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, *27*(7), 773-780.

Acknowledgments This work was partly supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innocation Program on AI and Big Data 18102059-0.