Constructing Knowledge Representation from Lecture Videos through Multimodal Analysis

Pak-Ming Fan and Ting-Chuen Pong

***Abstract*—E-learning has presented new opportunities for learning with the rapid development of information and communication technologies (ICTs). Learners are no longer restricted by the location and time to learn. Lecture video is one of the most commonly used learning materials on e-learning platforms. It presents knowledge in a lively manner and keeps the learners more attentive during the learning process. While organizing lecture videos in a sequential list seems to be a natural choice, it presents the problems of inefficiency in searching for domain concepts and the inability to show relationships between such concepts. In this work, the task of constructing a knowledge representation scheme for video corpuses is explored. The knowledge representation aims to achieve the goals of facilitating the searching of domain concepts and to extract the relationships between the concepts so as to identify effective learning strategies for the corpus. A framework using text recognition, speech recognition, multimodal analysis and clustering techniques is proposed for the construction of the knowledge representation. Two lecture video corpuses on the topics of general chemistry and geometry are acquired from the Khan Academy for demonstrating the feasibility of the proposed framework. Experimental results have shown that the framework can be used to achieve the intended goals in specific domains.**

***Index Terms*—E-Learning, knowledge representation, online education, learning strategies.**

I. Introduction

Traditional learning requires the learner to attend lessons at specific locations at fixed times. With the advancement in information and communication technologies (ICTs), a new learning paradigm, e-learning, has arisen. With this new paradigm, new opportunities in learning are made possible. Learners are no longer restricted by the time and location to learn. Moreover, ubiquitous learning is also attainable through mobile communication devices such as smartphones and tablets. Self-paced learning and blended/hybrid learning are also possible in this new paradigm.

E-learning has presented a new way of acquiring knowledge and has gained significant support from various educational institutions, ranging from elementary schools to collages. Massive open online course (MOOC) is one of the means for e-learning. It provides learning materials to learners for free through a web-based platform and has gained significant attention recently. Some of the media coverage including articles from the New York Times [1] and the cover story from the Time Magazine [2]. One of the most popular MOOC platforms is Coursera. Coursera is an organization founded by two Stanford University computer scientists. It announced the offering of more than 100 MOOCs from 33 partner universities in Fall 2012 and has attracted over 2 million registered users by the end of 2012. Other MOOC organizations such as edX, Udacity and Khan Academy have also gained publicity for providing free e-learning materials and benefiting learner around the world.

Lecture video is one of the most commonly used learning materials on e-learning platforms. It allows learning to be delivered in a lively manner, thus keeping the learners more attentive during the learning process. The gestures and speech of the instructor in the videos also help the learners to focus on the important contents of the course and therefore building a more structural understanding of the subject. Although the number of lecture videos varies across different subjects and different e-learning platforms, the lecture videos are usually represented in a simple manner. Many e-learning platforms represent the videos in form of a list, arranged in sequential order. Learners are assumed to have no previous experience on the subject and are expected to start watching according to the ordering sequence of the videos. After finishing the complete list, the learner is expected to have a good understanding on the subject.

Organizing the lecture videos in a sequential list appears to be a natural choice for representing the video corpus, but this representation presents some limitations. For example, when a learner wants to search for a particular concept in the video corpus, he/she will have to skim through the corpus to find the related videos. Moreover, it will be difficult for the learners to know what possible concept he/she can proceed to after studying the current concept. It will also be hard to find out the pre-requisite for learning a new concept. The inefficiency in searching for concepts in the lecture videos and the inability to show relationships between concepts suggest room for improvement on the current representation.

The aim of this research is to explore the possibility of constructing a knowledge representation on the video corpus with the goals to facilitate the searching of concepts in a lecture video corpus and to extract the relationships between the concepts so as to identify effective learning strategies for the corpus.

II. Related Word

Knowledge representation refers to the representation of

Manuscript received January 1, 2013; revised March 18, 2013. This work was supported in part by the Research Grants SSRI99/00.EG1, HIA01/02.EG04 and RGC 618208.

The authors are with the Hong Kong University of Science and Technology Computer Science and Engineering Department, Clear Water Bay, Kowloon, Hong Kong (e-mail: leofpm@ust.hk, tcpong@ust.hk).

knowledge by symbols so as to allow reasoning and infer new knowledge. According to [3], a knowledge representation should have the following five main properties: 1. Representing the subject completely; 2. Showing the basic entities, properties and the relationships between entities; 3. Guide the reasoning between the basic entities without needing to take action on it; 4. Allows efficient reasoning; 5. Facilitates the exchange of ideas between humans on the subject.

Knowledge representation on videos describes the contents inside the videos [4]. The entities in a video knowledge representation can be in form of low abstraction, e.g. objects appeared inside the video, or high abstraction, e.g. events happened in the video. Instead of concentrating on efficient reasoning between different entities, video knowledge representations focus on facilitating the search and retrieval of videos, such as answering the query of "find a shot of a dog chasing a ball". Reference [5] gives a broad overview on knowledge representation of videos, including entities of different levels of abstraction and different models on representing the entities.

Reference [5] presents a knowledge representation on American football videos. The trajectories of the America football are used as highly abstracted entities in the representation. Searching of the football video clips at a finer granularity is then allowed by using the trajectories as queries. Reference [6] shows a knowledge representation on a diverse set of videos. Entities of different levels of abstraction such as objects, shot transitions and actions are used in the representation. These entities are labeled by icons and are categorized. The searching process is then conducted by inputting one or more icons in different categories.

In this work, a framework is proposed to construct a knowledge representation on a lecture video corpus. Concept words extracted from the lecture video corpus are used as the basic entities in the knowledge representation. The lecture videos are clustered into subtopics which are then indexed by the concept words. Finally, the subtopics are linked to give a hierarchical representation. The searching of subtopics is achieved by inputting concept words as queries. Possible learning strategies can also be identified by observing the hierarchical structure of the representation.

III. Methodology

A. Data Source

Among a number of lecture video collections existing in the internet, lecture videos from the Khan Academy [7] are chosen to study the feasibility of the proposed framework based on the following reasons: 1. Popularity – Khan Academy's YouTube channel has over 150 million total views; 2. Diversity – Khan Academy contains lecture videos of a wide range of subjects; 3. Video Quality – Most Khan Academy videos are available in High Definition (720P). 4. Preprocessing Work – Khan Academy videos are created by using a digital painter software and therefore do not suffer from problems like occlusion or poor lighting conditions. Hence less preprocessing work is required.

B. Concept Words Extraction

Concept words are extracted by applying text recognition, speech recognition and multimodal analysis. The video titles, handwriting trajectories and the audio transcripts represent textual, visual and acoustic information respectively. This information is combined to produce a list of concept words which are then used in the knowledge representation. To be precise, handwriting trajectories are extracted from each lecture video by a frame differencing algorithm. Transcripts of each video are also recognized by using the "System.Speech library" in the Microsoft .NET framework [8]. The handwritten concept words are then extracted by using a library included in the Microsoft Windows XP Tablet PC Edition SDK [9], taking the trajectories and the transcripts as input. The handwritten concept words, together with the text extracted by analyzing the video titles are treated as an initial set of potential concept words. The transcripts are then used again to filter the list of potential concept words. Words that are absent from the transcripts are removed. The filtered list is the finalized list of concept words. The information of which concept words are first extracted from each of the videos is also recorded for indexing at a later stage.

C. Subtopics Formation

Subtopics are formed by grouping videos with high similarity score together. To calculate the similarity score, a vector space model [10] is used to represent the videos. In the vector space model, each video is represented as a vector of weights on the list of concept words extracted. A boolean weighting scheme is used to define the weights given to the elements in the vector (i.e. weight = 1 if concept word appears in the video; weight = 0 otherwise.) The similarity score between any two videos is then formulated by the cosine similarity between the weight vectors of the two videos. The score ranges from 0 to 1 and is calculated by taking the dot product between the normalized weight vectors of the two videos.

Given a sequence of videos, an iterative clustering algorithm based on the construction of a maximal spanning tree (MST) is used to form the subtopics. The MST connects all the videos together in an optimal manner according to the similarity scores between the videos. The algorithm contains the following steps:

1) Put each video into a single cluster
2) For each consecutive pair of clusters, mark the pair for grouping if a MST edge exists between the two clusters
3) Terminate if no clusters are marked for grouping, each remaining cluster is treated as a subtopic
4) Otherwise, combine each marked pair into a single cluster
5) Repeat step 2

However, the above algorithm suffers from the disadvantage that only the most similar videos are grouped. Videos with high similarity might not be grouped together. Therefore, two initialization steps are taken before the algorithm for better clustering:

1) Grouping Series of videos – videos that have similar titles and contain serial numbers are grouped to be a single cluster.
2) Grouping neighboring videos with certain number of

common concept words – two neighboring videos having number of common concept words greater than a threshold are grouped to be a single cluster. The threshold can be a factor of the maximum number of common concept words between any two videos.

Finally, each subtopic is indexed by the first extracted concept words from the videos within it to facilitate searching.

D. Constructing the Knowledge Representation

After forming the subtopics, a knowledge representation can be constructed so that possible learning strategies on the subtopics can be identified. Two different methods of constructing the knowledge representation are presented.

Method 1: Hierarchical clustering

Hierarchical clustering groups more similar subtopics in lower levels of the hierarchy and less similar subtopics in higher levels of the hierarchy. The number of common concept words between two subtopics defines the similarity between two subtopics and the result is visualized by a dendrogram. The clustering algorithm is as follows:

1) Initialize each subtopic to be a cluster
2) Mark all clusters unprocessed and having an empty cluster number
3) Initialize a threshold, k, to be a factor of the maximum number of common concept words between two clusters
4) For each pair of unprocessed clusters, if the number of common concept words > threshold, mark the pair with the same cluster number; a new cluster number is used only if both clusters have an empty cluster number
5) For each group of clusters marked with the same cluster number, form a new unprocessed cluster; mark all the clusters in the group processed
6) Decrement k by 1, if $k \geq 0$, repeat step 4; otherwise, terminate

The algorithm takes at most $O(n^2)$ time to process all the pairs of clusters in each level of the hierarchy. Thus the total running time of the algorithm is $O(kn^2)$, where n is the number of subtopics.

Method 2: MST-based construction

MST-based construction utilizes the maximal spanning tree on the videos created during the subtopics formation phase. Two subtopics are linked together if a MST edge exists across them. The construction begins from the first subtopic formed. A tree structure is used to visualize the results where each subtopic is regarded as a node in the structure. The construction algorithm is as follows:

1) Let V be the set of subtopic nodes, E be the set of tree edges, Q be an empty queue
2) Mark all nodes unprocessed
3) Put the first subtopic into Q
4) Repeat while Q is not empty
   a) Remove the first subtopic, u, from Q and mark u processed
   b) For each MST edge crossing u and another subtopic v, if v is not processed and v is not in Q, put v into Q and add edge (u, v) to E

Each node is input to the queue once and each edge is checked once, thus the algorithm runs at $O(|V| + |E|)$ time.

The two methods for constructing the knowledge representation differ in the learning strategies they produce. Hierarchical clustering produces a learning strategy on the entire lecture video corpus, giving a simple guideline on viewing all the subtopics. MST-based construction produces learning strategies base on a user selected subtopic. Suggestions on other subtopics for viewing will be given depending on the user input. Examples on the learning strategies produced by both methods will be given in the following section.

IV. Result and Discussion

Two lecture video corpuses on the topics of general chemistry and geometry are acquired from the Khan Academy for testing the feasibility of the proposed methodology. Table I shows a brief summarization on the two corpuses:

TABLE I: Lecture Video Corpuses

	General Chemistry	Geometry
Number of videos	104	136
Total Duration	~24 hours	~17 hours
Total Size	1.4 GB	0.8GB

A. Concept Words Extraction

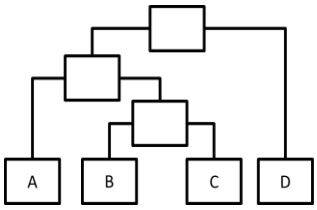
Speech recognition and handwriting recognition are two essential steps on extracting concept words. Speech recognition gives a recognition error rate of 15% with a concept word based evaluation approach. Handwriting recognition by using handwritten trajectories alone gives a low recognition rate of 18%. This recognition rate is later improved to 50% with the speech recognized transcripts used as a language model. By processing the video titles and the above recognition steps, a total of 203 and 117 concept words are extracted from the general chemistry corpus and the geometry corpus respectively. These concept words are then used for forming and indexing subtopics. A total of 27 and 45 subtopics are formed from the general chemistry corpus and the geometry corpus respectively.

B. Knowledge Representation Construction by Hierarchical Clustering

A learner can identify a learning strategy on the corpus by observing the levels of which the subtopics are linked together in the hierarchy. If the subtopics are linked in a lower level of the hierarchy, it indicates that they are more related and can be watched in a closer manner. If the subtopics are linked in a higher level of the hierarchy, it indicates that they are less related and can be watched more separately. By watching subtopics starting from the lower levels to the higher levels of the hierarchy, the learner will be able to view the subtopics according to their relatedness. Fig. 1 shows part of the hierarchy extracted from the general chemistry corpus. By following the above mentioned strategy, subtopic B and subtopic C are watched together first; and then followed by subtopic A and then finally subtopic D.

The complete representation on the general chemistry is shown in Fig. 2. Some of the closely related subtopics like *ideal gas* and *thermodynamics*; *balancing chemical*

equations and *empirical formulas* are indeed linked together in a lower level by the dendrogram. However, some less related subtopics in the geometry corpus are also linked in a lower level by the dendrogram. This is caused by concept words that appear significant number of times in most of the videos. Since the hierarchical clustering algorithm links subtopics based on the number of common concept words, the clustering result is affected greatly. It can also be observed from Fig. 2 that a significant number of subtopics are linked together in the lower levels to form a single large group while the remaining subtopics are linked to this large group in the higher levels of the hierarchy. Our investigation found that these remaining subtopics only contain a few videos and a lower number of concept words. Therefore, they can only be linked in a later stage of the algorithm where the threshold for linking is low. Similar observation is also obtained from the geometry corpus.



Subtopic A	Stoichiometry
Subtopic B	Atom, Elements and Electrons
Subtopic C	Periodic Table
Subtopic D	Spectrophotometry

Fig. 1. Part of the dendrogram extracted from the general chemistry corpus.

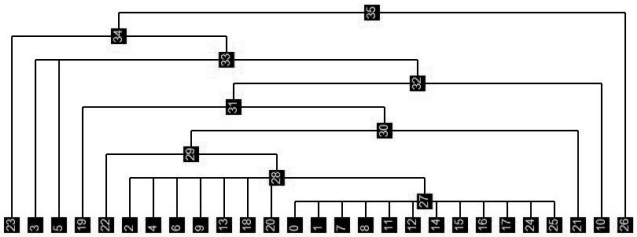


Fig. 2. Dendrogram produced by hierarchical clustering on the general chemistry corpus, subtopics represented square blocks.

C. Knowledge Representation Construction by Mst-Based Construction

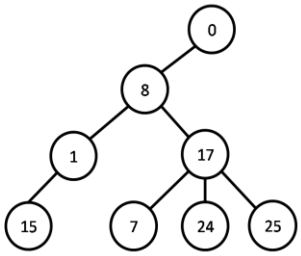
Two Learning strategies can be identified from the tree representation.

1) Breadth-first learning strategy

When the learner finished a subtopic, he/she can identify the possible subtopics for viewing by observing the direct neighbors in the tree. The direct neighbors represent all the subtopics that are most related to the current subtopic. By applying this learning strategy, learners can get a more specific understanding of the subtopic he/she has finished. Fig. 3 shows part of the tree representation extracted from the general chemistry corpus. As an example, when the learner finished subtopic 17 in Fig. 3, subtopics 7, 8, 24 and 25 are suggested for viewing according to the breadth-first strategy.

2) Depth-first learning strategy

The learner can also follow paths from the root node to a leaf node or from an intermediate node to a leaf node. This learning strategy suggests one of the most related subtopics repeatedly until no suggestions can be made further. By applying this strategy, the learner can get a more broad understanding of the subtopic he/she has finished. As an example, when the learner finished subtopic 0 in Fig. 3, subtopics 8, 1 and 15 are suggested for viewing according to the depth-first strategy.



Subtopic 8	States of matter
Subtopic 17	Thermodynamics
Subtopic 7	Ideal gas equation
Subtopic 24, 25	Thermodynamic calculation Examples
Subtopic 0	Atoms, electrons, electron configurations
Subtopic 8	States of matter
Subtopic 1	Periodic table, chemical bonds
Subtopic 15	Oxidation states

Fig. 3. Dendrogram produced by hierarchical clustering on the geometry corpus, subtopics represented in square blocks.

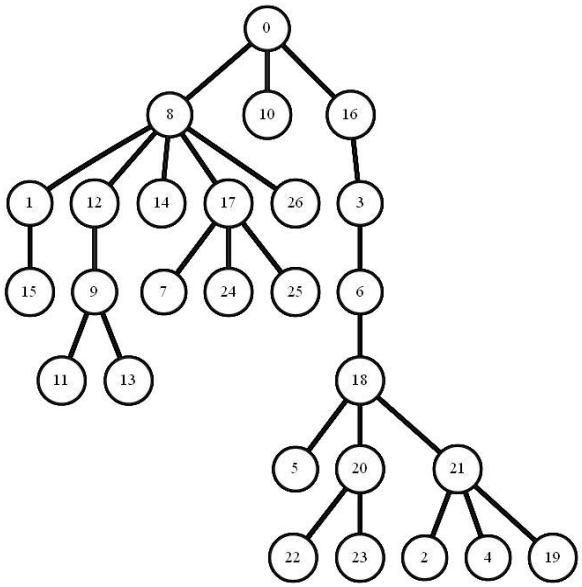


Fig. 4. Tree representation produced by MST-based construction on the general chemistry corpus, subtopics are represented by tree nodes.

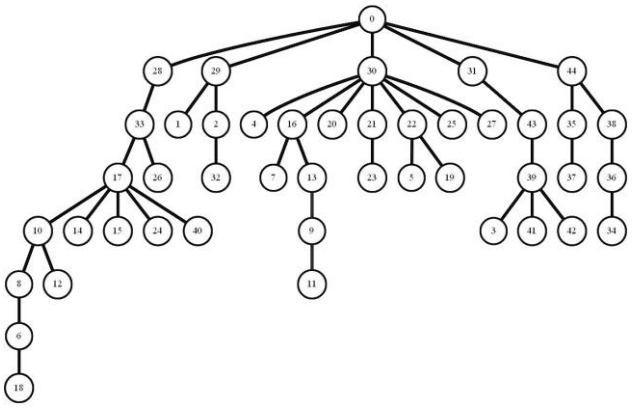


Fig. 5. Tree representation produced by MST-based construction on the geometry corpus, subtopics are represented by tree nodes.

The complete representations on the general chemistry and geometry corpus are shown in Fig. 4 and Fig. 5. Related subtopics are linked together as direct neighbors or along a path from the root to leaf nodes by the proposed algorithm in both corpuses. For example, some subtopics about stoichiometry are linked together as a subtree appearing on the right hand side of Fig. 4 whereas some subtopics on circle are linked as a subtree appearing on the right part of Fig. 5.

The proposed algorithm can also be applied on the video Corpus directly without forming the subtopics first. Fig. 6 shows the tree representation of the general chemistry corpus by applying the algorithm without subtopics formation. However, applying the algorithm directly will result in a tree of a larger size, which might not be an effective representation when the number of videos is large. On the other hand, if the MST is formed by considering the distance between the videos in the sequence instead of the similarity score, the resulting tree will just be a linked list on the videos, following the order of appearance in the sequence.

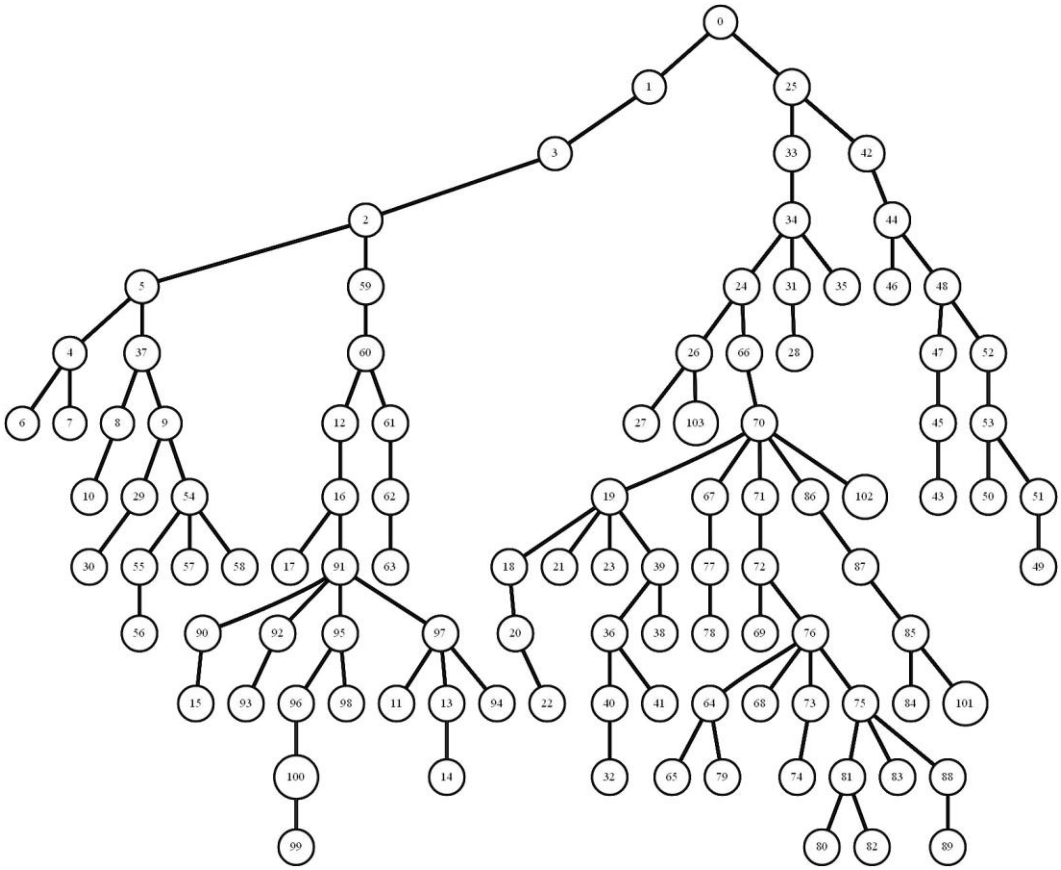


Fig. 6. Tree representation produced by MST-based construction on the general chemistry corpus without subtopics formation, videos are represented by tree nodes.

V. Conclusion

In this work, a framework for constructing a knowledge representation for lecture video corpuses is presented. The framework includes the three main phases of concept words extraction, subtopics formation and knowledge representation construction. While the representations constructed are able to aid learners in searching subtopics and identifying learning strategies at certain extent, possible improvements can be considered. For example, using the proportion of common concept words between two subtopics instead of just the number of common concept words could link subtopics more meaningfully in the hierarchical clustering algorithm. It is also possible to consider additional aspects, such as the existing concept words and new concept words inside a video, so as to extract possible directional relationships between the videos. These directional relationships can then be used to construct a directed graph representation where each node is a video or a subtopic, and the directed edges between the nodes represent the suggested order for viewing the videos or subtopics.

It is also worth mentioning that although the proposed framework is tested on Khan Academy lecture videos, applying it on lecture videos from other sources is also possible with slight modification in the concept words extraction phase.

References

[1] T. Lewin. (July 2012). Universities Reshaping Education on the Web. *The New York Times*. [Online]. Available: http://www.nytimes.com/2012/07/17/education/consortium-of-colleges-takes-online-education-to-new-level.html.

[2] H. McCraken. (October 2012). MOOC Brigade: What I Learned From Learning Online. *Time Magazine*. [Online]. Available: http://nation.time.com/2012/10/22/mooc-brigade-what-i-learned-from-learning-online/#ixzz2Iwg4zSYY.

[3] R. Davis, H. Shrobe, and P. Szolovits, "What is a Knowledge Representation?" *AI Magazine*, vol. 14, no. 1, pp. 17-33, 1993.

[4] M. Davis, "Knowledge representation for video," in *Proc. Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Washington, USA, 1994, pp. 120-127.

[5] W. Al-Khatib, Y. F. Day, A. Ghafoor, and P. B. Berra, "Semantic modeling and knowledge representation in multimedia databases," *IEEE Trans. on Knowledge and Data Engineering*, vol. 11, no.1, pp. 64-80, Jan. 1999.

[6] M. Davis, "An iconic visual language for video representation," in *Readings in Human-Computer Interaction: Toward the Year 2000*, R. M. Baecker, J. Grudin, W. A. S. Buxton, S. Greenberg, Ed., San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 854-866.

[7] Khan Academy. [Online]. Available: https://www.khanacademy.org.

[8] Microsoft. NET framework System.Speech. [Online]. Available: http://msdn.microsoft.com/en-us/library/gg145021(v=vs.100).aspx.

[9] Microsoft Windows XP Tablet PC Edition SDK. [Online]. Available: http://www.microsoft.com/en-us/download/details.aspx?id=20039.
[10] C. D. Manning, P. Raghavan, and H. Schtze, "Scoring, term weighting & the vector space model," *Introduction to Information Retrieval*, New York, NY, USA: Cambridge University Press, ch. 6, 2008.

Pak-Ming Fan was born in Hong Kong, China. He obtained a Bachelor degree in computer engineering and computer science in 2010 and a Master of Philosophy degree in computer science and engineering in 2012 both from the Hong Kong University of Science and Technology (HKUST). He is currently working as a research assistant in education technology also in HKUST.

Ting-Chuen Pong received his PhD in Computer Science from Virginia Polytechnic Institute and State University, USA in 1984. He joined the University of Minnesota - Minneapolis, USA as an Assistant Professor of Computer Science in 1984 and was promoted to Associate Professor in 1990. In 1991, he joined the Hong Kong University of Science & Technology (HKUST), where he is currently a Professor of Computer Science and Engineering. He was an Associate Vice-President for Academic Affairs at HKUST from 2002 to 2010, an Associate Dean of Engineering from 1999 to 2002 and Director of the Sino Software Research Institute from 1995 to 2000. He also served as an Academic Research Adviser for the Hong Kong Research Grants Council (RGC) from 2010 to 2012. Professor Pong is a recipient of the HKUST Excellence in Teaching Innovation Award in 2001.Professor Pong's research interests include computer vision, image processing, multimedia computer, and IT in Education. He is a recipient of the Annual Pattern Recognition Society Award in 1990 and Honorable Mention Award in 1986. He is a registered auditor of the Quality Assurance Council of the University Grants Committee of Hong Kong since 2008.