TÌM KIẾM VIDEO BÀI GIẢNG DẠNG SLIDE DỰA VÀO NỘI DUNG

Lê Văn Hào[1], Lê Thị Hồng Hà[2], Trịnh Thị Anh Loan[3]

TÓM TẮT

Trong những năm gần đây, giáo dục trực tuyến E-learning, thư viện bài giảng số hay cổng thông tin bài giảng đang trở nên ngày càng phổ biến hơn. Khi số lượng video bài giảng đang tăng trưởng nhanh chóng thì các phương pháp tìm kiếm hiệu quả những video bài giảng này vẫn đang còn là nhiệm vụ thử thách. Các văn bản hiển thị trong một video bài giảng có mối quan hệ chặt chẽ với nội dung bài giảng, cung cấp nguồn dữ liệu có giá trị cho việc lập chỉ mục và tìm kiếm nội dung. Trong bài báo này, chúng tôi trình bày một cách tiếp cận để tự động lấy được các văn bản nội dung, từ đó tiến hành lập chỉ mục và cho phép tìm kiếm bài giảng dựa trên các từ khóa liên quan đến nội dung của video.

Từ khóa: *Tìm kiếm video dựa trên nội dung, nhận dạng kí tự quang học, trùng lặp văn bản, sửa lỗi chính tả, lập chỉ mục tài liệu.*

1. ĐẶT VẤN ĐỀ

Cùng với sự phát triển của công nghệ thông tin, số lượng video bài giảng, diễn thuyết… phục vụ học tập cho mọi lứa tuổi đang được tải lên và chia sẻ trên internet nhanh chóng. Đối với lượng video đang tăng trưởng từng ngày, cơ chế tổ chức lưu trữ phục vụ cho việc tra cứu, tìm kiếm là rất quan trọng và là nhiệm vụ thách thức.

Giáo dục trực tuyến hay E-Learning không còn là khái niệm mới lạ mà đang phát triển mạnh mẽ. Nhu cầu tìm kiếm của người dùng càng yêu cầu khắt khe hơn: cả về độ chính xác và thời gian tìm kiếm. Tuy nhiên, các chức năng tìm kiếm bài giảng của các hệ thống hiện tại chỉ cho phép người dùng tìm kiếm với tên bài giảng, tên học phần, hoặc tên giảng viên... Các chức năng này thường cho kết quả có độ chính xác không cao, và các kết quả trả về có nhiều nội dung không liên quan đến mục đích tìm kiếm thực sự của người dùng.

Những công cụ tìm kiếm phổ biến hiện nay là những hệ thống tìm kiếm dựa trên "từ khóa", và tìm kiếm trên dữ liệu văn bản. Chính vì thế, nếu video không có bất kỳ siêu dữ liệu (metadata) ví dụ như ngày, tác giả, từ khóa, hoặc mô tả thì không thể tìm kiếm được bằng cách sử dụng các công cụ tìm kiếm hiện tại. Siêu dữ liệu thường được thêm bằng tay, quá trình này sẽ rất tốn thời gian. Hơn nữa, ngay cả khi một đoạn video có thể được tìm thấy bằng siêu dữ liệu của nó, công cụ tìm kiếm thông thường không có khả năng tìm kiếm một đoạn bài giảng, khung hình cụ thể trong video mà người dùng quan tâm.

Trong bài báo này, chúng tôi hướng tới tìm hiểu và nghiên cứu một hệ thống tìm kiếm các video bài giảng, thuyết trình, trình diễn bằng silde dưới dạng video. Cho phép tìm thấy những video dựa vào văn bản xuất hiện trong đó. Giải pháp này cũng cho phép người dùng tìm kiếm các video không cần có siêu dữ liệu.

[1,2,3] *Giảng viên khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Hồng Đức*

2. NỘI DUNG

2.1. Phương pháp tiếp cận

2.1.1. Kiến trúc của hệ thống tìm kiếm video dựa trên nội dung

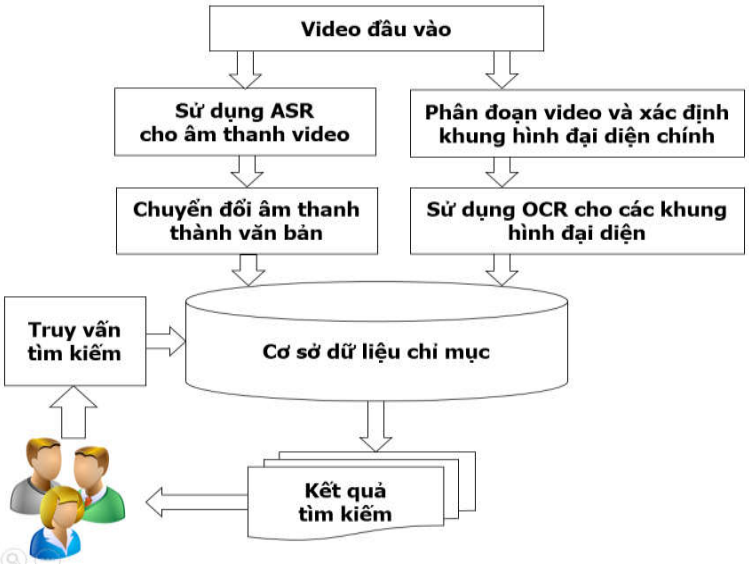
Một hệ thống tìm kiếm thông thường gồm hai thành phần, thành phần lập chỉ mục và thành phần xử lý truy vấn tìm kiếm (hình 1). Bài toán xây dựng một hệ thống tìm kiếm video được chia thành hai bài toán con được miêu tả như sau:

Bài toán 1: Xử lý video đầu vào, trích chọn văn bản đại diện cho video

Đầu vào: Tập video đầu vào.
Đầu ra: Văn bản đại diện cho video đầu vào.

Bài toán 2: Lập chỉ mục và xử lý truy vấn tìm kiếm của người dùng

Đầu vào: Truy vấn từ người dùng.
Đầu ra: Danh sách xếp hạng các video có liên quan đến truy vấn.



Hình 1. Kiến trúc chung của hệ thống tìm kiếm video dựa trên nội dung

Trong bài toán thứ nhất, vì đặc tính của video là có cả hình ảnh và âm thanh nên sẽ có hai cách phương pháp tiếp cận chính để trích xuất văn bản từ video:

Phân tách video thành các khung hình để lựa chọn khung hình đại diện chính, sau đó sử dụng công nghệ nhận dạng kí tự quang học (Optical Character Recognition - OCR) để trích xuất văn bản từ các khung hình đó.

Sử dụng công nghệ nhận dạng giọng nói tự động (Automatic Speech Recognition - ASR), để chuyển đổi phần âm thanh của video thành văn bản.

2.1.2. Các nghiên cứu về tìm kiếm video bài giảng dựa trên nội dung

Liška et al và cộng sự đã đề xuất một hệ thống tự động cho việc lập chỉ mục video bài giảng [4]. Họ sử dụng toàn bộ khung hình phân đoạn từ video và sử dụng công cụ OCR để

trích xuất văn bản trên tập khung hình đó. Văn bản sau khi thu thập được tiến hành lập chỉ mục và cho phép tìm kiếm. Giải pháp này hiệu quả kém do không loại bỏ các tệp văn bản trùng lặp. Thời gian xử lý video mất nhiều thời gian do số lượng lớn các khung hình.

Hunter et al đề xuất một hệ thống lập chỉ mục cho các bài thuyết trình đa phương tiện [5]. Đầu tiên, mọi người sẽ phải chuẩn bị một tệp tin thuyết trình định dạng PDF và gửi lại sau khi đã trình bày. Sau đó tệp tin sẽ được đồng bộ với video thuyết trình. Công việc OCR sẽ được thực hiện trên tệp tin PDF mà không cần quan tâm đến video thuyết trình.

Yang et al sử dụng công cụ nhận dạng giọng nói tự động ASR để trích xuất nội dung video thành văn bản [5]. Các kết quả cho thấy độ chính xác của nhận dạng giọng nói thấp hơn rất nhiều so với công nghệ OCR.

Lienhart et al đề xuất một phương pháp phát hiện văn bản trong video và hình ảnh [4]. Họ xây dựng một mạng noron nhiều tầng để huấn luyện phát hiện văn bản. Thuật toán của họ xử lý với tất cả các khung hình phân đoạn được và cách tiếp cận này kém hiệu quả về thời gian xử lý.

2.1.3. Phương pháp của nhóm tác giả

Dựa vào các phương pháp tiếp cận nghiên cứu đã nêu trong phần trước, công cụ tìm kiếm video mà chúng tôi mong muốn xây dựng được hình thành từ cách giải quyết các bài toán cụ thể sau (hình 2):

Phân đoạn video.

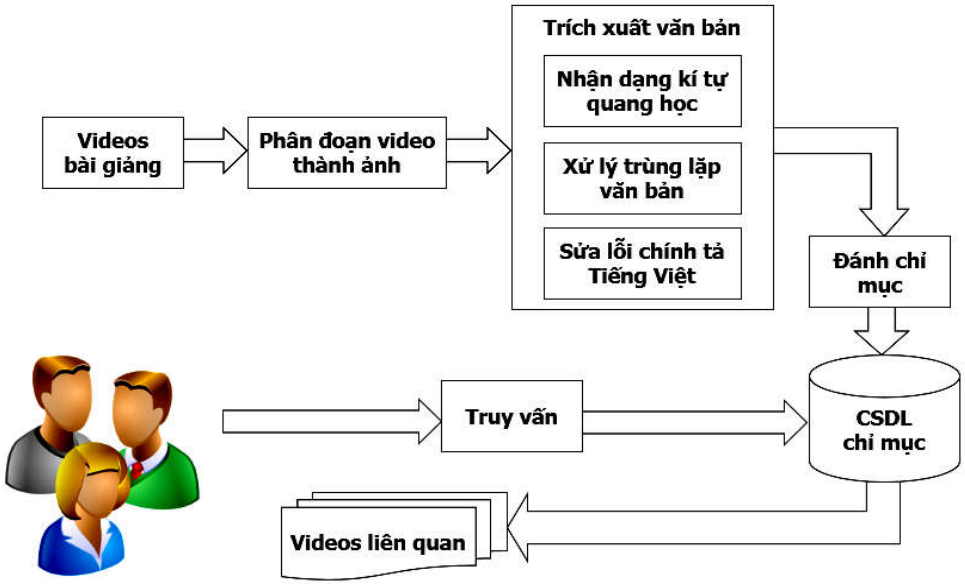
Trích xuất văn bản đại diện:

Nhận dạng kí tự quang học.

Xử lý trùng lặp văn bản.

Sửa lỗi chính tả văn bản.

Đánh chỉ mục và tìm kiếm.



Hình 2. Kiến trúc hệ thống tìm kiếm video bài giảng của nhóm tác giả

2.2. Kỹ thuật tiến hành

2.2.1. Phân đoạn video

Về mặt bản chất thì video mà chúng ta thấy trên tivi, máy tính, điện thoại… được cấu thành từ những ảnh tĩnh. Những ảnh này sau đó được sắp xếp liên tiếp nhau và cùng trình diễn trong một đơn vị thời gian đủ nhỏ để làm cho mắt của chúng ta cảm nhận rằng các đối tượng này đang chuyển động. Thông thường thì các video được quay ở khoảng 24-30 hình mỗi giây.

Mỗi hình này được gọi là một frame. Số frame trên một giây được đo bằng một số nguyên được kí hiệu FPS. Một video đơn giản được hiểu là tổng số khung hình được lưu trữ cùng nhau và trình chiếu theo một thứ tự, do vậy một video thông thường có khoảng vài trăm đến vài trăm nghìn khung hình.

Có nhiều công cụ hỗ trợ phân đoạn video, nhưng chúng tôi sử dụng FFMpeg[4] bởi: FFMpeg là một thư viện nguồn mở có rất nhiều tiện ích cho việc xử lý video. Tính năng nổi bật nhất là khả năng encode/decode nhiều video định dạng khác nhau, giúp chuyển đổi qua lại nhiều định dạng video. Ngoài ra, cũng có thể dùng FFMpeg để chia cắt một đoạn video, chụp lại các frame và xuất ra dạng hình ảnh.

2.2.2. Trích xuất văn bản đại diện

Trong bài toán trích xuất văn bản, để nâng cao hiệu quả và thay vì phải xử lý toàn bộ khung hình khi phân đoạn video. Chúng tôi chia bài toán thành ba vấn đề nhỏ hơn đó là:

Bài toán nhận dạng kí tự quang học để trích xuất văn bản từ video;

Bài toán xử lý trùng lặp văn bản để thu được tệp văn bản đại diện cho video;

Bài toán sửa lỗi chính tả tiếng Việt.

2.2.2.1. Nhận dạng kí tự quang học

Nhận dạng kí tự quang học là công việc đầu tiên trong quá trình trích xuất văn bản. Nhóm tác giả sử dụng Tesseract-OCR[5] để thực hiện trích xuất nội dung văn bản từ ảnh. Tesseract là một công cụ nhận diện kí tự quang học mã nguồn mở và hiện nay được phát triển bởi Google [3]. Tesseract-OCR có các ưu điểm: công cụ miễn phí, hỗ trợ nhiều hệ điều hành (Windows, Linux, Mac…), hỗ trợ trích xuất đồng loạt nhiều tệp tin cùng lúc, hỗ trợ trên 100 ngôn ngữ khác nhau, một trong những công cụ mã nguồn mở OCR chính xác nhất hiện nay [3]. Bảng 1 cho kết quả sau khi chúng tôi thực hiện nhận dạng kí tự quang học.

Bảng 1. Kết quả thực hiện Tesseract-OCR đối với tập khung hình thu được

STT	Số lượng	Kích thước tập kết quả (KB)	Độ chính xác (%)	Độ hồi tưởng (%)	Độ F1 (%)
1	382	136,3	71,2	81,8	76,13
2	398	100,5	71,1	82,0	76,16
3	187	33,7	76,4	67,0	71,39

[4] Phần mềm mã nguồn mở cung cấp thư viện xử lý video: http://ffmpeg.org/

[5] Phần mềm nguồn mở nhận dạng ký tự quang học: http://github.com/tesseract-orc

4	1707	529,1	66,4	76,2	70,96
5	155	45,0	77,5	66,3	71,46
Trung bình			72,52	74,66	73,22

2.2.2.2. Xử lý trùng lặp văn bản

Với đặc thù là các văn bản được trích xuất từ các khung hình video bài giảng liên tiếp theo thời gian. Chính vì thế tập hợp văn bản thu được tồn tại cả hai loại đó là trùng lặp và gần trùng lặp văn bản.

Theo các nghiên cứu ở tài liệu tham khảo [1] [2] có nhiều phương pháp tiếp cận để giải quyết vấn đề tìm các văn bản trùng lặp như:

Bag of words: So sánh các từ và tần số của những từ đó trên một văn bản với những văn bản khác.

Shingling: Cải thiện hơn so với Bag of words, phương pháp này sẽ tiếp cận bằng cách so sánh các cụm từ "shingle". Phương pháp này quan tâm đến ngữ cảnh của các từ (thứ tự của các từ).

Hashing: Các cụm từ sẽ được băm thành các con số và sau đó so sánh để tìm ra sự trùng lặp.

MinHash, SimHash: Cải tiến của phương pháp Hashing, giúp sắp xếp hợp lý quá trình lưu trữ nội dung được băm.

Dựa trên các kết quả nghiên cứu đã nêu thì phương pháp shingling cho kết quả độ chính xác cao và phù hợp với kiểu dữ liệu đầu vào như tập dữ liệu của tác giả. Chính vì thế, nhóm tác giả lựa chọn và cài đặt thuật toán phát hiện trùng lặp văn bản dựa vào kỹ thuật Shingling của Broder và cộng sự. Kết quả thể hiện ở bảng 2.

Bảng 2. Kết quả thực hiện NDD với kỹ thuật Shingling

STT	Tập đầu vào	Số văn bản đại diện thu được	Số slide thực tế	Số văn bản đại diện đúng	Độ chính xác (%)	Độ hồi tưởng (%)	Độ F1 (%)
1	382	14	22	12	85,7	54,5	66,63
2	398	24	25	22	91,6	88,0	89,76
3	187	42	35	34	80,1	97,1	87,78
4	1707	14	18	13	92,8	72,2	81,21
5	155	21	24	18	85,7	75,0	79,99
Trung bình					87,18	77,36	81,07

2.2.2.3. Sửa lỗi chính tả văn bản

Đối với vấn đề phát hiện lỗi chính tả thì thường có hai phương pháp tiếp cận chính đó là kỹ thuật tra cứu dùng từ điển và kỹ thuật phân tích N-gram [7], [9]. Nhóm tác giả lựa chọn cả hai phương pháp để nhằm cải thiện chất lượng sửa lỗi chính tả. Các bước phát hiện và sửa lỗi chính tả văn bản được thực hiện như sau:

Bước 1: Đây là bước đầu tiên trong quá trình phát hiện và sửa lỗi chính tả. Dữ liệu đầu vào sau khi được nạp cần được loại bỏ một số kí tự dư thừa (không có ý nghĩa trong từ) như các khoảng trắng, các dấu chấm, hoặc các kí tự đặc biệt…

Bước 2: Phát hiện lỗi chính tả: Có nhiều cách, tiêu chí để phân loại nhưng trong khuôn khổ chương trình phát hiện lỗi chính tả ở mức từ thì lỗi chính tả được chia làm hai loại là lỗi non-word và lỗi real-word:

Lỗi non-word là lỗi tạo ra từ sai, từ đó hoàn toàn không có trong từ điển từ vựng hoặc một số từ điển tên riêng, từ điển viết tắt, từ điển vay mượn... Đây là loại lỗi dễ phát hiện.

Lỗi real-word là lỗi chính tả mà từ đó có trong từ điển nhưng sử dụng từ sai. Nếu không dựa vào ngữ cảnh xung quanh thì không thể xác định được đó có phải là lỗi chính tả hay không. Đây là loại lỗi khó phát hiện và xử lý.

Bước 3: Dựa vào từng loại lỗi để lựa chọn từ thay thế cho từ bị lỗi.

Bảng 3. Kết quả quá trình phát hiện lỗi chính tả dùng Aspell kết hợp Bi-gram

STT	Tập đầu vào (số từ)	Tổng số lỗi thực tế	Số lỗi phát hiện được	Số lỗi phát hiện đúng	Độ chính xác (%)	Độ hồi tưởng (%)	Độ F1 (%)
1	946	77	71	66	92,9	85,7	89,15
2	1365	121	112	96	85,7	79,3	82,38
3	2482	43	33	18	54,54	41,8	47,33
4	786	96	91	85	93,4	88,54	90,91
5	1520	31	26	22	84,6	70,9	77,15
Trung bình					82,23	73,25	77,38

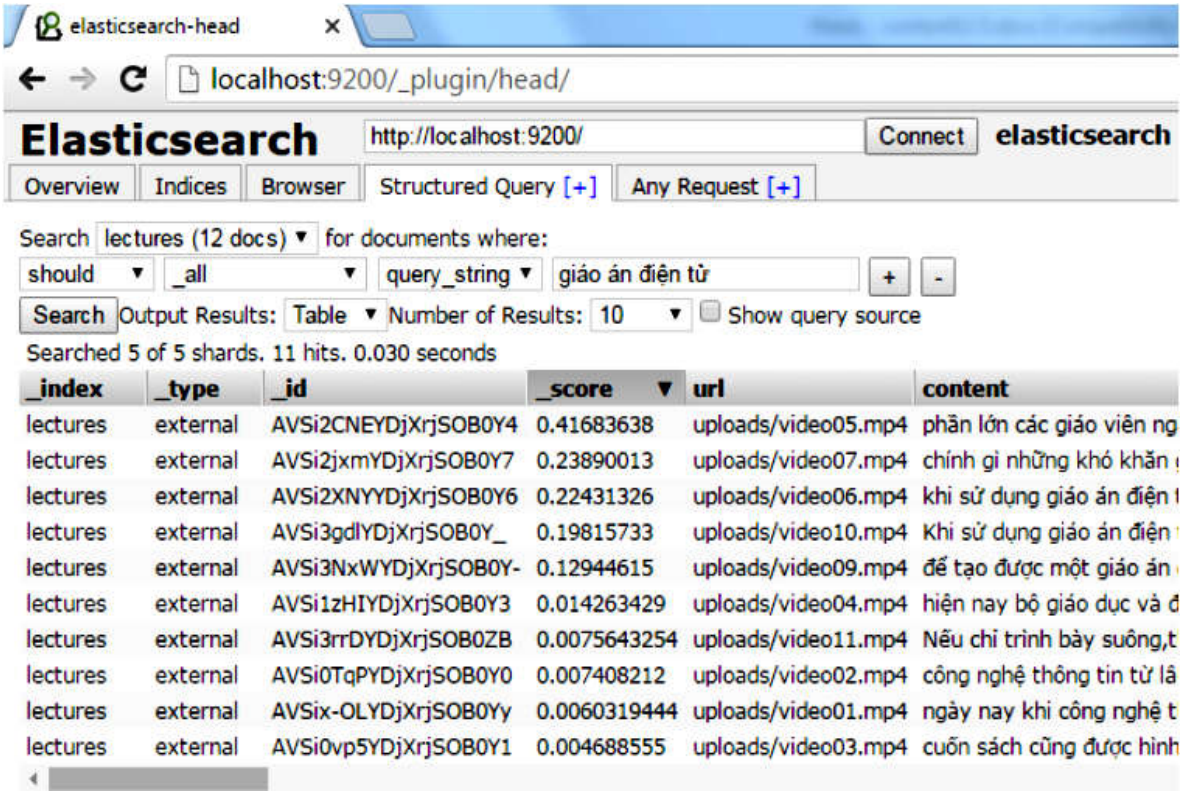
Bảng 4. Kết quả quá trình sửa lỗi chính tả

STT	Số lỗi phát hiện	Số lỗi sữa	Số lỗi sửa đúng	Độ chính xác (%)	Độ hồi tưởng (%)	Độ F1 (%)
1	71	69	49	71,0	69,0	69,99
2	112	102	62	65,8	55,4	57,97
3	33	16	9	56,3	27,3	36,77
4	91	84	43	51,2	50,5	49,17
5	26	28	18	64,3	69,2	66,66
Trung bình				60,72	53,64	56,11

2.2.3. Lập chỉ mục và tìm kiếm

Các phương pháp lập chỉ mục đóng vai trò quan trọng trong việc xây dựng một hệ thống tìm kiếm thông tin hiệu quả. Lập chỉ mục tài liệu là công việc sắp xếp tài liệu nhằm đáp ứng nhanh chóng yêu cầu tìm kiếm thông tin của người sử dụng.

Elasticsearch[6] là một máy chủ tìm kiếm dựa trên Lucence. Hiện nay, đánh giá của DB-Engines thì Elasticsearch là công cụ tìm kiếm doanh nghiệp phổ biến nhất, tiếp theo là Apache Solr, cũng dựa trên Lucene. Elasticsearch nhiều ưu điểm như: không cần cấu hình phức tạp, hỗ trợ thêm, xoá, sửa chỉ mục thông qua các phương thức HTTP như GET, POST, DELETE và PUT, hỗ trợ tham số dưới dạng JSON thay vì chỉ là GET params, cài đặt và sử dụng dễ dàng mà không cần cài thêm bất cứ ứng dụng nào khác, tìm kiếm gần như thời gian thực (real-time). Hình 3, là minh họa cho một truy vấn tìm kiếm video có chứa từ "giáo án điện tử".



Hình 3. Ví dụ minh họa tìm kiếm với Elasticsearch

3. KẾT LUẬN

Trong bài báo này, chúng tôi hướng tới mục đích tìm hiểu và nghiên cứu phương pháp để xây dựng một hệ thống tra cứu video dựa trên nội dung. Video nhóm tác giả quan tâm là các video bài giảng dạng silde. Nội dung của truy vấn sẽ là các từ hoặc các cụm từ có liên quan đến nội dung văn bản bên trong các video bài giảng.

Bài báo đã trình bày về các khái niệm liên quan đến công cụ tìm kiếm. Các phương pháp tiếp cận, kỹ thuật áp dụng để giải quyết các bài toán về xây dựng công cụ tìm kiếm video. Ứng dụng các phương pháp, kỹ thuật để thực nghiệm xây dựng một hệ thống tìm kiếm video bài giảng dựa trên nội dung.

[6] *Công cụ tìm kiếm dựa trên phần mềm Lucence: http://www.elastic.co/*

TÀI LIỆU THAM KHẢO

[1] Andrei Z. Broder (2000), *Identifying and Filtering Near-Duplicate Documents*, 11th Annual Symposium on Combinatorial Pattern Matching, Springer-Verlag London, pp.1-10.

[2] Bassma S. Alsulami (2012), *Near Duplicate Document Detection Survey*, International Journal of Computer Science & Communication Networks, pp. 147-151.

[3] Chirag Patel, Atul Patel, Dharmendra Patel (2012), *Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study*, International Journal of Computer Applications, Vol. 55, No.10, pp. 50-56.

[4] Haojin Yang, Maria Siebert, Patrick Lühne, Harald Sack, Christoph Meinel (2011), *Automatic Lecture Video Indexing Using Video OCR Technology*, 2011 IEEE International Symposium on, pp. 111-116.

[5] Haojin Yang (2011), *Lecture Video Indexing and Analysis Using Video OCR Technology,* 7th International Conference IEEE Dijon France, pp. 54-61.

[6] Nguyen Thi Xuan Huong, Tran Thai Dang, Nguyen The Tung, Le Anh Cuong (2015), *Using Large N-gram for Vietnamese Spell Checking*, Advances in Intelligent Systems and Computing, pp. 617-627.

[7] Pratip Samanta, Bidyut B. Chaudhuri (2013), *A simple real-word error detection and correction using local word bigram and trigram*, Association for Computational Linguistics and Chinese Language Processing, pp. 211-220.

[8] Radu Gheorghe, Matthew Lee Hinman, Roy Russo (2016), *Elasticsearch in Action*, Manning Publications Co, Shelter Island.

[9] Youssef Bassil, Mohammad Alwani (2012), *Context-sensitive Spelling Correction Using Google Web 1T 5-Gram Information*, Computer and Information Science, Vol. 5, No. 3, pp. 37-48.

SEARCHING FOR LECTURE SLIDE VIDEO BASED ON CONTENT

Le Van Hao, Le Thi Hong Ha, Trinh Thi Anh Loan

ABSTRACT

For the last years, digital lecture libraries and lecture video portals have become more and more popular. However, finding efficient methods for indexing multimedia still remains a challenging task. Since the text displays in a lecture video is closely related to the lecture content, it provides a valuable source for indexing and retrieving lecture contents. In this paper, we present an approach for automatic lecture video indexing based on video

OCR technology. To archieve this aim, we have studied the solutions how to search video based on the content, including frame separation from a video; text recognition from images; spelling correction; text indexing and searching. Addittionally, we also prove the accuracy of the modules by evaluation.

Keywords: *Content based video retrieval, optical character recognition, near-duplicate detection, spelling correction, indexing document.*