Navigating and Discovering Educational Materials through Visual Similarity Search

Suzanne Little, Rebecca Ferguson, Stefan Rüger
Knowledge Media Institute
The Open University, UK
{s.little|r.m.ferguson|s.rueger}@open.ac.uk

Abstract: We describe the development and implementation of visual multimedia similarity search within a platform for social exchange of educational experiences and material to provide services for finding related media. The SocialLearn project develops tools to support the building and exploration of personal learning networks. With the ever-increasing volumes of educational resources being made available, it is a challenge to find new material and forge appropriate learning pathways. Visual search can help when it is difficult to describe your interests in words ("search terms") or when you want to browse for inspiration without a specific result in mind. In this paper we present the usage scenarios for visual search within education and describe the design and implementation of visual similarity search within the SocialLearn platform. The outcomes from this work are not only directly useful for the SocialLearn project but also for others who are interested in the challenges of using multimedia for education.

Introduction

Educational materials commonly include items beyond simple text documents. The Open University (OU), in particular, has a very rich archive of multimedia educational resources to offer. These include videos, photographs, screenshots, illustrations, diagrams, slideshow-based presentations, bundled educational archives and web pages. Traditional text-based search creates an index by using formal metadata or text descriptions of items, occasionally automatically generated but more often manually added. To find material of interest, a learner supplies a keyword or phrase that they think best describes what they are looking for. Alternatively if the learner is familiar with an institution and its metadata model, they may supply a course code, faculty or subject heading.

The problem with traditional text-based search is it often places the onus on the learner to have sufficient knowledge of what they are looking for, or the specific search system they are using, to be able to create a successful query. More frequently material is discovered through recommendations from educators or peers or perhaps through serendipitous browsing. One facet of the SocialLearn project is to provide a framework for learning with others – peers, educators, informal learners – and enable educational material to be shared, recommended or placed within a defined learning pathway.

Visual Similarity Search (VSS) is a type of Content-Based Multimedia Search (CBMS) that uses the low-level features from images or video (e.g., colour, texture, shape) to find material that is visually related. These features summarise a particular characteristic of the media's pixels that can then be used to index the media independently of any other available metadata, tags or descriptions. Calculating a mathematical difference between the values of these descriptors produces a ranked list of the "most similar" media items based on one or more query images.

Visual similarity search can be used for different purposes depending on configuration and implementation choices. Three common purposes are near-duplicate detection, known object identification and general search. Near-duplicate detection focusses on finding reuse of specific media often with slight modifications such as resizing, cropping, changes in colour etc. Known object identification uses many of the same techniques as near-duplicate detection but is intended to identify an object from a known set (e.g., book/cd cover, tourism site, museum artefact) where the search operates across a database of example images to match the object. General search is more difficult to quantify and includes both vague ("find holiday photos") and specific ("find illustrations of photosynthesis") search goals. Key outcomes for the search results are their accuracy, coverage, diversity and "serendipity". The importance of these measures varies with the intended purpose of the search.

These applications of VSS can be used to support navigation of educational materials in a variety of ways.

For example, finding reuse of material in different contexts with different supporting materials, discovering the source of a screenshot or finding items that share visual features and may provide new ways of understanding a concept. The SocialLearn+Media project was born out of the desire to provide new methods for educators and learners to explore resources and find multimedia material that may be of interest. Existing research into content-based multimedia search was adapted to implement a number of web services providing functionality to index and query a range of multimedia educational resources. These web services are then used by the SocialLearn platform to provide novel methods of suggesting, browsing or finding educational media.

In this paper we describe the SocialLearn project and give a brief overview of the topic of visual similarity search. We present usage scenarios for SocialLearn+Media and discuss how visual similarity searching can be used to augment traditional indexes and provide new ways of navigating educational resources. The technical considerations, architecture and implementation of the SocialLearn+Media web services are described and the results from technical performance evaluations are given. Finally we discuss the future possibilities of this work and how it may be used to help connect searchers with new types of resources.

Related Work

SocialLearn[1] (Buckingham Shum & Ferguson, 2010) is a platform for online open learning that aims to give people control of what and how they learn, and with whom, by harnessing the power of the web and social networking. The project runs at The Open University and utilises the wealth of open educational resources (OER) and distance learning experience available there. The motivation behind the project was the realisation that while OER may improve the quality of material available to online learners, it also greatly increases the quantity. The result is a data deluge leaving learners to struggle to identify useful material, understand how to build connections between concepts or even to recognise solutions when they find them. The challenge of successful education and training in these circumstances is explored in detail by many others (e.g., the collection edited by (Deakin Crick, 2009)).

The conception of learning underpinning SocialLearn is summarised by (Seely Brown, J. and Adler, 2008), being… “based on the premise that our understanding of content is socially constructed through conversations about that content and through grounded interactions, especially with others, around problems or actions”. In the early days of the SocialLearn project, (Weller, 2008) identified six broad principles of SocialLearn, connecting it with the underpinnings and origins of The Open University: *Openness, Flexibility, Disruptive, Perpetual beta, Democracy* and *Pedagogy*. This flexibility and willingness to provide novel methods for learners to build their understanding of the OER materials available to them opened the possibility of integrating the latest research into content-based multimedia search into the SocialLearn platform.

Multimedia information retrieval and content-based multimedia search are areas of broad and active research with a wide variety of applications in entertainment, business, science and education. Commercial media search engines and databases (e.g., (Subrahmanian & Jajodia, 1996)) have long been used to manage the ever-increasing quantities of digital media. Good general overviews of the research trends and technical challenges of content-based search can be found in (Datta et al., 2008; Lew et al., 2006; Rüger, 2010).

In the multimedia and education field, earlier work by (Neumüller, 2002) acknowledges the special challenges when trying to discover knowledge in multimedia in the aptly named “Because I seek an image, not a book” discussing the applications of semiotics in hypermedia browsing. The notion of gaining understanding through exploratory search (for all types of documents) is also discussed by (Marchionini, 2006) – particularly relevant are the *learn* and *investigate* facets of exploratory search. This is a key concept that we try to implement in this work. The outcome from visual similarity search is the ability to construct an automatic hypermedia network available to be explored by the user in the hope that they will discover new pathways to learning. Recent work by (Ah-Pine et al., 2011) continues the theme of using media for understanding. The reflections on user evaluations and experience of their proposed system are very relevant to the work we discuss here.

Finding Learning Materials

In this section we present three usage scenarios around SocialLearn+Media. These serve to describe the initial motivation of the project, the needs of its users and how visual similarity search may be used in this context.

[1] SocialLearn, http://sociallearn.net, Last accessed: 11 April 2011

Scenario 1: Exploring Learner

The user is a learner, involved in a following a formal learning pathway prescribed by an educator. The SocialLearn platform allows them to exchange suggestions or hold discussions with their peers. While viewing educational material the user can call the “Suggest Related Media” function from the SocialLearn toolbar to query the indexed media collection and get visually related objects that may lead them to other resources, explanations or source material. For example, a screenshot in a set of slides is used to query and find the original video containing the slide giving the learner more context for the concept.

Scenario 2: Developing Educator

The user is a lecturer developing a learning pathway for a course to be distributed to remote students. While developing new materials from their notes and slides, they want to find alternative illustrations and video material to supplement their explanations. The SocialLearn platform can help the user to navigate and find available media in new ways. For example, based on some of their old slides they find and are able to link to a newer presentation by a colleague that has a better illustration of the point they want to make.

Scenario 3: Curious Explorer

The user is an informal learner who discovers some open educational resources and wishes to find related information. They are interested in browsing the SocialLearn platform and available resources, idly following connections that catch their eye and learning more about one or more topics and the relationship between them. For example, an initial starting point of a video of a presentation about solar energy leads to a discussion group on alternative energy sources through media searches that first find the slides from the talk and then connect the example screenshots to further video linked to the discussion group.

Visual Similarity Search and Social Learning

Architecture

The integration of visual similarity search functionality with the SocialLearn platform was designed around a web services based architecture. This has the advantage of isolating the specific implementation of VSS from the requirements of SocialLearn enabling the services to be hosted independently and future improvements to be implemented without significant disruptions to the main platform. The architecture is illustrated in Figure 1.

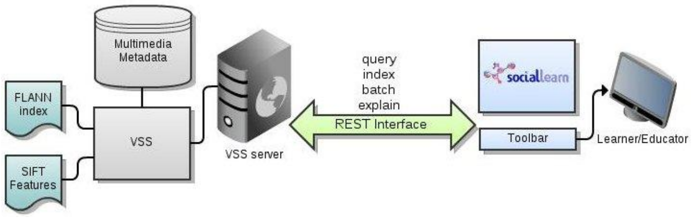


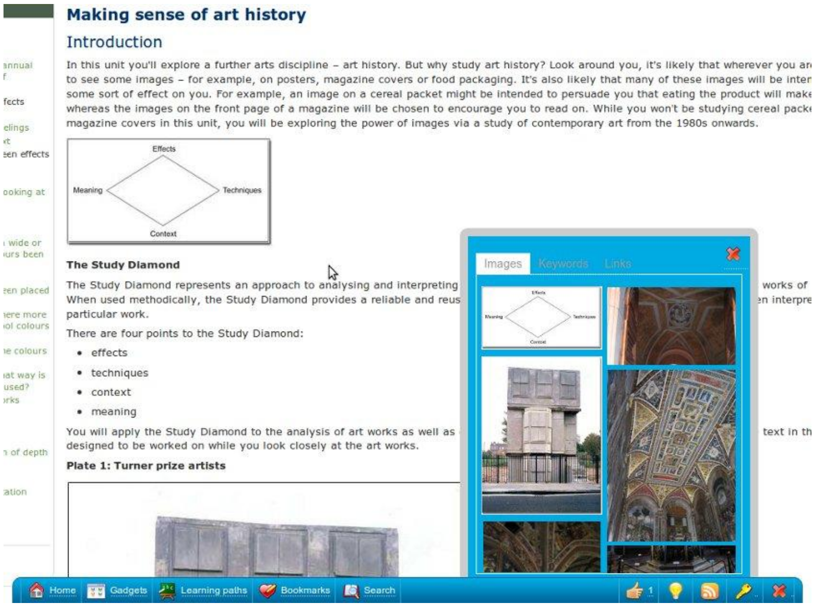
Figure 1: SocialLearn+Media Architecture

Two main services are provided – *query* and *index* – and two supplementary services – *batch* and *explain* (experimental). Table 1 gives the specifications of the web services. The services are implemented as REST services using the GET directive and served through the simple Python web.py framework.

Command	Input	Returns (JSON encoded)
query	url=<url of query image>	thumbnail, title, timestamp/slidenumber, confidence value, url to containing object (e.g., slideshow, video, webpage)
index	url=<url of media to add>&metadata=<json dictionary>	GUID of media in index
batch	-	success/fail of batch indexing process
explain	image1=<url to image>&image2=<url or index GUID>	image showing the matching keypoints

Table 1: SocialLearn+Media Web Service Specifications

The principal service used is *query* which is called from the SocialLearn platform to find indexed media related to something the user is viewing. This is currently implemented in the SocialLearn toolbar and Figure 2 shows a screenshot of the tool bar in use. The toolbar is able to extract media objects from a web page and provide thumbnails linking to query with the chosen image. Results are displayed in a popup window linking to the complete learning resource – i.e., slides, learning pathway, video etc. This keeps the new media items connected with the original resource.



Making sense of art history

Introduction

In this unit you'll explore a further arts discipline – art history. But why study art history? Look around you, it's likely that wherever you ar
to see some images – for example, on posters, magazine covers or food packaging. It's also likely that many of these images will be inter
some sort of effect on you. For example, an image on a cereal packet might be intended to persuade you that eating the product will make
whereas the images on the front page of a magazine will be chosen to encourage you to read on. While you won't be studying cereal pack
magazine covers in this unit, you will be exploring the power of images via a study of contemporary art from the 1980s onwards.

The Study Diamond

The Study Diamond represents an approach to analysing and interpreting works of
When used methodically, the Study Diamond provides a reliable and reus en interpre
particular work.

There are four points to the Study Diamond:

- effects
- techniques
- context
- meaning

You will apply the Study Diamond to the analysis of art works as well as text in th
designed to be worked on while you look closely at the art works.

Plate 1: Turner prize artists

Figure 2: The SocialLearn Toolbar using VSS

A wide variety of media can be indexed by the services either as one off calls to *index* or through the *batch* indexing process that adds a directory of media objects. Currently supported types are images, video, slides (as PDF), Open University (OU) course archives (ZIP files containing PDFs, images and videos) and Open University course videos hosted as part of the online podcast archive and indexed in the data.open.ac.uk portal. The common object for indexing is an image and all media types are processed to produce a set of images and any available metadata. The final image sets are all processed to prepare the visual search index described in the next section.

Videos are analysed using a shot-boundary detection tool (our own implementation, hive (Pickering & Rüger, 2003)) that extracts keyframe images for each shot and calculates video related metadata such as format, duration, shotrate etc. PDF files are either converted into a sequence of JPG images or, if they contain mostly text, they are processed to extract image objects that are then filtered using a colour histogram threshold to exclude text-heavy diagrams or incorrectly formatted objects. Text-heavy diagrams have very little visual distinction and without applying OCR will not provide any useful results. OU course archives are separated into the component media objects and indexed with the same course code. Course related videos are found by using the course code to query a SPARQL endpoint developed as part of the LUCERO project[2] and the open/linked data initiative. This returns a list

[2] LUCERO project, http://lucero-project.info, Last accessed: 11 April 2011.

of video urls that are downloaded and indexed. Implementation of indexing for other media types including PPT, Word documents and diagrams is ongoing.

The *explain* service is an experimental attempt to provide some reason behind the results given for the search. As described in the next section, matches are calculated based on the similarity of certain points in the images. By showing the matching points between the query and result images the user has some information about why a result is a similar image. See Figure 5 for an example. The confidence value provided by the query service fulfils a similar purpose and is also based on a rough calculation of the similarity of matching points. Optimising the ability of the visual similarity search services to provide explanations and more transparency for the results is continuing work.

Visual Search Implementation

Developing and implementing visual similarity search is a series of decisions and tradeoffs. The key process of VSS is illustrated in Figure 3.

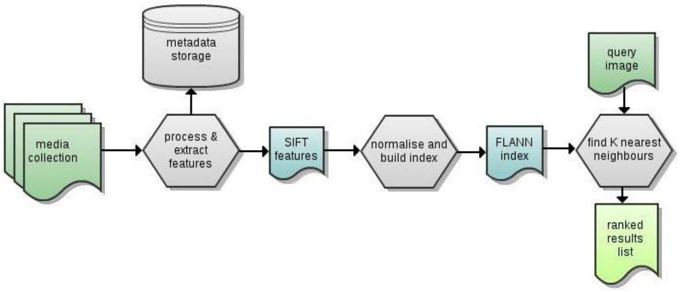


Figure 3: Visual Similarity Search Process

The selection of the optimum feature descriptions for similarity search is dependent on the types of media (images, illustrations, video, audio), the general content (holiday snaps, formal object portraits (e.g., for archives/museums), geographical/aerial survey shots, microscope images, security camera footage) and the purpose of the search (to find duplicates, identify specific objects/people, find general similarities). The choice of feature descriptor also impacts upon issues such as responsiveness, transparency (i.e., why is this a match?) and scalability. Features can be global – calculated on the whole of the image – or local – focussed on particular points. For the implementation of search used in SocialLearn+Media we chose to use local features called SIFT (Scale Invariant Feature Transforms)[3] proposed by (Lowe, 2004). SIFT identifies 'keypoints' in a grayscale image and calculates a description of the area around this point (vector of 128 real numbers) that is invariant to changes in scale and orientation. Figure 4 shows an example image with marked SIFT keypoints.

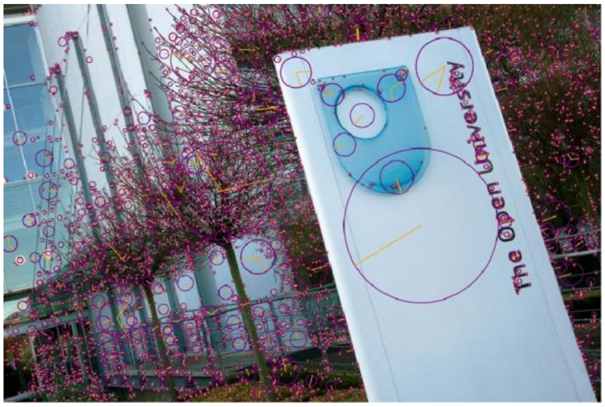


Figure 4: Example Image with SIFT keypoints

[3] The SIFT method is restricted for commercial use. We use it, as permitted, for research purposes only.

SIFT or similar features are commonly used in both near-duplicate detection and object identification tasks. SIFT keypoints can be compared to find matches using the description vector that indicate similarities between images. A match is calculated based on the mathematical distance (e.g., manhattan, euclidean distance) between the numerical vector. A ranked list of the most visually similar images from the indexed collection is produced by counting the sufficiently matching keypoints with the query image over a pre-set threshold value. Figure 5 shows two similar images and the matching keypoints between them.

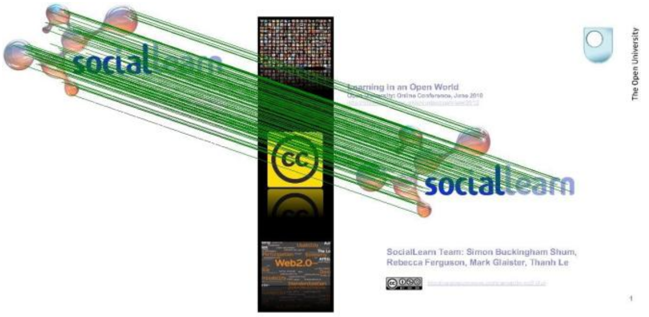

socialLearn
Web2.0
Learning in an Open World
Open University Online Conference, June 2010
The Open University
socialLearn
SocialLearn Team: Simon Buckingham Shum,
Rebecca Ferguson, Mark Glaister, Thanh Le
1

Figure 5: Matching SIFT keypoints

Exhaustive calculations of the distances for every keypoint and every image are computationally expensive. Average images may have been 500 and 2000 keypoints depending on their size and complexity. Therefore techniques such as approximate nearest neighbour search and index quantisation are used to improve the efficiency of the implementation.

Searches are conducted on “collections”. A collection includes the set of media and metadata, the normalised feature descriptors (stored in hdf5 format) and the approximate nearest neighbour index file. This has two advantages. Firstly the search performance is maximised with the reduction of potential noise or confusing data. It is also possible for users to set preferences for their main areas of interest or to direct the searches based on other user profile information.

Secondly the collections can be stored and searched independently of each other. They can even be stored on completely separate servers, managed independently and the search results combined at query time to be presented to the user. This has great advantages for the scalability of the system, distributing the storage and query load across multiple servers.

Evaluation

Evaluating a system for navigating educational materials in a platform designed for social learning is challenging. There are two facets of performance that must be considered. Firstly the information retrieval (IR) ability of the underlying visual similarity search – traditional metrics such as precision, recall etc. Secondly the user perception of the system's usefulness and capability. Standard datasets can be used to evaluate the IR performance while user testing and pilot studies are required for the second stage of evaluation.

As can be seen from the usage scenarios presented in this paper, there are a few different requirements for the visual similarity search in this context. Perhaps the most straightforward is detecting reuse of media objects – often in new contexts. For example, a slide is reused, a diagram is taken for a paper and shown in a video or some screenshots from a video are used in a slide. This is an example of near-duplicate detection where the goal is to find the reuse of the specific media item. The standard approach for evaluating the performance of near-duplicate detection is to run a series of transformations on a set of images (e.g., stretch, crop, change colour) and test the ability of the search system to find the transformed copies using the original image as the query. This follows the examples of (Ke et al., 2004; Meng et al., 2003). Table 2 gives the results of using 100 photos randomly selected from the Corel photo library and processed using ImageMagick[4] to apply 40 transforms as defined by (Ke et al., 2004). A further 100 random images were also added to the index to add potential noise to the dataset and increase the difficulty.

[4] Script to apply the transformations is available from http://kmi.open.ac.uk/people/suzanne/transform.sh

Collection	P@1	P@(N)
Transformed Corel	1.00	0.77(40)
ukbench	1.00	0.62(4)

Table 2: Initial Evaluation Results

A similar, but more challenging, type of retrieval is to find different images of the same object, e.g., a book cover, the entrance to The Open University. Standard datasets containing multiple images of the same object are used to evaluate the performance. Table 2 contains the results from using the ukbench dataset (Nister & Stewenius, 2006) which contains 10200 photographs of 2550 objects (sets of 4 photos per object). By querying using one example from each object the aim is to find the matching photo plus the other 3 photos of the object in the top 4 results.

The other type of performance to be evaluated is more nebulous in definition. This is the serendipitous match that users will just "know when they see it". User evaluations and pilot studies are needed to test this. These are not easy to set up as managing user's expectations about what type of media is available in the system and how it might be used will form a key part of the process. Future work will include targeted studies with users who are interested in particular domains. Media for that domain can then be included in the index. Observation, result logging and user interviews will be used for these evaluations.

Discussion: Using VSS for learning

In this section we discuss the performance of the VSS as judged by the evaluations and consider some of the limitations and challenges for using VSS in education.

The IR based evaluations have demonstrated good performance according to traditional metrics although there is still room for refinement. Errors in the transformed dataset are generally limited to the really extreme cases with precision generally remaining at 1.00 up to position 37. This is to be expected with a local descriptor that is invariant to changes in colour, scale and orientation. The VSS can usually identify three out of the four available images for each object. This is a more challenging dataset and variations in the configuration options can have significant effects. More fine tuning is possible using this set although we are conscious of avoiding any over fitting of the VSS system when tested on an artificial dataset. Questions, therefore, still remain about the user perception of the performance and how it can be used within SocialLearn.

One critical limitation with using visual similarity search is that visual similarity does not always equal semantic similarity. That is, media that looks the same is not necessarily depicting the same concept or object. In the application of VSS to SocialLearn this effect is slightly mitigated by the use of near-duplicate and object detection based approaches and the integration as a recommendation service to supplement user navigation in conjunction with learning pathways. Hence the service seeks to add value by using all available information to help the user. The use of confidence values to act as a threshold to results is also important as some images will have no good matches. The explain option has also received good (informal) feedback as it helps the user to understand why particular results are suggested.

Performance is also dependent on the indexing of good visual materials. Fortunately the OU has a considerable collection of multimedia resources from over 40 years of distance education. Grouping media into separately indexed collections also helps to improve the performance. It also gives a guide to the user of the types of media that can be expected in the results list.

Finally a significant issue is that of scalability of the system. Nearest neighbour search, even with the very fast approximation methods used, is still computationally expensive. Large images give lots of keypoints, each with 128 numbers to describe it. This can produce a very large set of data very quickly. Two approaches are being used in our implementation to manage this problem. First, as discussed previously, the ability to divide the index into collections that can be queried independently and then fusing the results to return to the user. Second reducing the amount of information needed for each image.

Reducing the size of the descriptors for each indexed image can be accomplished in two ways. Either reduce the number of keypoints by altering the sensitivity of the extraction process or downscaling the image or compress the size of the descriptor by summarising or reducing the 128 numbers. Initial experiments have indicated that using downscaled images for indexing (e.g., maximum size of 256) produces much faster performance with no loss of precision and often an increase. We posit that this is due to the reduction in noisy outlier keypoints and better summarisation of the images. Alternative local keypoint methods (Juan & Gwun, 2010) such as SURF (Bay et al., 2006), PCA-SIFT or Colour-based SIFT (van de Sande et al., 2009) are also strong options to reduce the descriptor size and potentially improve performance.

Future Work and Conclusions

The visual similarity search web services are currently deployed in the test installation of the SocialLearn platform. Quantities of educational resources are being indexed using the batch adding feature and large video collections are being acquired to add. Further evaluation of the IR performance of the underlying VSS is ongoing and new configurations are being considered to improve the performance, flexibility and scalability. The key next stage is the assessment of the performance through user evaluations and pilot studies.

A number of extensions to the functionality of the media search component are also being investigated including incorporating modules to analyse and index audio, perform speech-to-text transcription and OCR on images with text. Text output from content-based analysis may also be indexed in traditional methods and merged with the main SocialLearn search systems. Of particular interest are technical drawings and diagrams that may perform better with specialised features to assess the semantic similarity of the diagrams rather than only their visual similarity. As these modules are added and refined the ability of the SocialLearn system to recommend content to users can be augmented by combining content-based, social network, personal metadata and semantic text analysis to give better or more varied suggestions backed up by explanations showing how the connections were made.

This work has demonstrated a method for enhancing social and computer-supported cooperative learning with more effective integration of multimedia learning materials. The use of a web services based architecture and structuring of the visual similarity search index into distributable collections has enabled the VSS component to be implemented independently of the existing SocialLearn platform and be plugged in where needed. The performance of the VSS according to traditional IR metrics is good with promising future research options to refine and improve the search. The integration of content-based media search with the traditional search techniques and social networks provides exciting new ways to navigate learning pathways and explore online educational materials.

Acknowledgements

The contributions of members of the SocialLearn team at The Open University – Darius Augaitis, Thang Dang, Thanh Le – in developing the toolbar for integration with the SocialLearn platform is gratefully acknowledged.

References

Ah-Pine, J., Renders, J. M., & Viaud, M. L. (2011). A continuum between browsing and query-based search for user-centered multimedia information access. *Adaptive Multimedia Retrieval. Understanding Media and Adapting to the User,* pp111-123.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded up robust features. Computer Vision–ECCV 2006, pp404–417.

Buckingham Shum, S., & Ferguson, R. (2010). Towards a social learning space for open educational resources. OpenED2010: Seventh Annual Open Education Conference.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv*, 40(2), pp1-60.

Deakin Crick, R. (2009). Pedagogical challenges for personalisation: Integrating the personal with the public through context-driven enquiry. *Curriculum Journal, Special Issue*, 20(3), pp185-306.

Juan, L., & Gwun, O. (2010). A comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing (IJIP)*, 3(5), pp143-152.

Ke, Y., Sukthankar, R., & Huston, L. (2004). Efficient near-duplicate detection and sub-image retrieval. ACM Multimedia

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: state of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1), pp1-19.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), pp91-110.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4), pp41-46.

Meng, Y., Chang, E., & Li, B. (2003). Enhancing DPF for Near-replica Image Recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Neumüller, M. (2002). Because I seek an image, not a book. *Hypermedia: Openness, Structural Awareness, and Adaptivity*, pp167-170.

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161-2168.

Pickering, M., & Rüger, S. (2003). Evaluation of key-frame based retrieval techniques for video. *Computer Vision and Image Understanding*, 92(1), pp217-235.

Rüger, S. (2010). Multimedia information retrieval. Morgan & Claypool Publishers.

Sande, K. van de, Gevers, T., & Snoek, C. (2009). Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp1-33.

Seely Brown, J. and Adler, R. P. (2008). Minds on Fire: Open Education, the Long Tail, and Learning 2.0. *EDUCAUSE Review*, 43(1), pp16-32.

Subrahmanian, V. S., & Jajodia, S. (1996). Multimedia database systems. Springer.

Weller, M. (2008). The SocialLearn Project. Blog post: http://bit.ly/1dVwQ6. Last accessed: 4th May 2011.