On Machine Learning and Knowledge organization in Multimedia Information Retrieval

Andrew MacFarlane, Sondess Missaoui and Sylwia Frankowska-Takhari[1]
[1]Centre for HCI Design
Department of Computer Science
City, University of London
London EC1V 0HB

andym@city.ac.uk
sondess.missaoui@city.ac.uk
sylwia.frankowska.1@city.ac.uk

Abstract

Recent technological developments have increased the use of machine learning to solve many problems, including many in information retrieval (IR). Deployment of machine-learning techniques is widespread in text search, notability web search engines (Dai et al., 2011). Multimedia information retrieval as a problem however still represents a significant challenge to machine learning as a technological solution, but some problems in IR can still be addressed by using appropriate AI techniques. In this paper we review the technological developments, and provide a perspective on the use of machine-learning techniques in conjunction with knowledge organisation techniques to address multimedia IR needs. We take the perspective from the MacFarlane (2016) position paper, that there are some problems in multimedia IR that AI and machine learning cannot currently solve. The semantic gap in multimedia IR (Enser, 2008) remains a significant problem in the field, and solutions to them are many years off. However, there are occasions where the new technological developments allow the use of knowledge organisation and machine learning in multimedia search systems and services. Specifically we argue that the improvement of detection of some classes of low level features in images (Karpathy and Li, 2015), music (Byrd and Crawford, 2002) and video (Hu et al., 2011) can be used in conjunction with knowledge organisation to tag or label multimedia content for better retrieval performance. We advocate the use of supervised learning techniques. We provide an overview of the use of knowledge organisation schemes in machine learning, and make recommendations to information professionals on the use of this technology with knowledge organisation techniques to solve multimedia IR problems.

1. Introduction

AI techniques, in particular machine learning, have become a significant technology in information retrieval software and services. Machine learning is defined as a method that learns from data with minimal input from humans. A key example is search engines (Dai et al., 2011) which use learning to rank algorithms to keep results presentation up to date given the inherent dynamism of the web. The web changes constantly both in terms of content and user requests, the data being documents, queries and click throughs etc. For text retrieval the machine-learning infrastructure is an essential part of the provision of a service that meets user needs, but the same could not be said of multimedia information retrieval where many challenges are still evident. By multimedia retrieval we mean search for non-text objects such as images, pieces of music or videos (moving images). Because of the semantic gap (Enser, 2008), the features of these objects can be hard to identify and index, which leads to a separation of techniques in terms of concept-based retrieval and content-based retrieval (with text we have terms that represent both). In MacFarlane (2016) it was argued that human involvement is necessary in many circumstances to identify concepts recognisable to humans – the example being a picture of a politician in an election. Whilst the politician can be easily identified, the election is a more nebulous concept that is difficult to extract from an image, without context. Low-level features of objects are often difficult if not impossible to match with concepts, and this problem is likely to be one that persists for a significant length of time. Knowledge organisation methods are essential to ensure that these conceptual features are captured and recorded in multimedia software and services.

However, recent advances in machine-learning methods such as machine vision algorithms (Karpathy and Li, 2015) have provided the functionality to identify specific objects in images, giving multimedia IR designers and implementers the ability to address the semantic gap to some extent. It is argued that in conjunction with knowledge organisation, machine learning can be used to provide better and more relevant results to users for a given set of information needs that require the identification of specific objects. This paper puts forward an argument for a supervised learning approach in multimedia search, where a knowledge organisation scheme is used as a rich source of information to augment the objects identified by any machine-learning algorithm. This is to provide an enhanced index of objects, allowing more effective search for those objects by the user. In this paper, we address the technological changes which have led to the potential for improvements in multimedia search, and argue that knowledge organisation can used with a supervised learning technique. We then review the landscape of multimedia search and show some possibilities for using knowledge organisation and machine learning to improve results for users in some types of information needs. Features in various types of multimedia objects are reviewed, and we provide some advice on how to use these feature and machine learning in conjunction with knowledge organisation in multimedia IR systems and services. We provide some ideas for the way forward, together with the practical implications for knowledge organisation practitioners.

The contribution of the paper is a process that uses knowledge organization and machine learning to create a database of objects for the purposes of multimedia information retrieval. The proposed process uses both high level and low level features identified for a multimedia object and the creation of an index within a database for the purpose of retrieval.

2. Technological developments in machine learning

What are the key developments which have led to improvements in technology, and which have significant implications for the use of knowledge organisation in multimedia search? In recent years, Deep Learning has become much more prominent in machine-learning circles (Pouyanfar et al., 2018), for a wide range of

different applications such as speech processing and machine vision (Deng and Yu, 2014). As one might expect, there is a wide range of definitions of Deep Learning, depending on the context, but the most appropriate in this context is a "class of machine-learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification" (Deng and Yu, 2014). Whilst the underlying technology for deep learning (Artificial Neural Networks) has been around for many years (McCulloch and Pitts, 1943), it is only recently that the widespread use of the techniques has become widespread and available in open frameworks such as TensorFlow (Abadi et al., 2016). Over the years the AI community has developed a strong body of knowledge in the use of the techniques, but a key turning point has been the availability of Graphical Processing Units (GPUs), which are specialist chips that are able to significantly increase the processing of arithmetical operations (Singer, 2013). They are particularly useful for image processing, but have become very useful generally for other types of applications such as neural networks that require significant processing of numbers. A benchmarking experiment conducted by Cullinan et al (2013) showed significant advantages for the GPU over CPU's (Central Processing Units) in terms of raw processing. The raw processing power from GPUs has proved to be the catalyst for a massive increase in Deep Learning algorithms in areas such as machine vision to detect features in images. This includes features such as the detection of neuronal membranes (Ciresan et at., 2012), breast cancer (Ciresan et al., 2013) and handwritten Chinese character recognition (Ciresan and Meier, 2015). Whilst these unsupervised examples of feature detection are relevant, it is argued that knowledge organisation can be used to augment feature extract to improve multimedia retrieval performance.

3. Machine learning and knowledge organisation

As feature extraction from various media has improved in recent years, what are the implications of the use of knowledge organisation techniques? Knowledge organisation in its many forms (thesauri, taxonomies, ontologies) are human-generated schemes which provide a rich source of evidence to describe features of objects that are of interest – in this case multimedia objects such as images, music and video. The key to understanding the contribution knowledge organisation can make in multimedia search is to consider the types of learning: unsupervised, semi-supervised and supervised (Russell and Norvig, 2016). These are classed by their access to labelled or categorised data. Unsupervised learning (Russell and Norvig, 2016, p.694) is where algorithms work without any labelled data, for example with clustering objects together based on the features extracted from them. This does not apply to our context, where we examine the use of knowledge organisation techniques to the problem. Semi-supervised learning (Russell and Norvig, 2016, p.695) does provide some access to labelled data, and it is possible to use this technique in some contexts where a limited number of multimedia objects have been manually classified by a practitioner. Supervised learning (Russell and Norvig, 2016, 695) requires access to data that is completely labelled, and is appropriate here – where we consider a large number of multimedia objects have been classified by a practitioner. Using supervised learning techniques we can either match features detected by both the machine-learning algorithm and the practitioner (exact match case) or estimate the probability of features matching from both sources (best match case). We consider both examples later on the paper in section 6. In this paper, we focus on the user of knowledge organisation and supervised learning in multimedia search, in the context of large amounts of data that have been labelled by practitioners.

4. Machine learning and multimedia information retrieval

There are limits to the use of machine learning/AI techniques to the application of multimedia information retrieval (MacFarlane, 2016). However, with the new advances in technology laid out in section 2 above and the ability of machine-learning algorithms to detect objects in media e.g. images (Karpathy and Li, 2015), there is scope to improve multimedia search results using knowledge organisation. In (MacFarlane, 2016) we argue that media of various kinds (e.g. images, music) require cultural knowledge which can often be only expressed tacitly, and thus require human input. The advantage of knowledge organisation schemes is that they provide this knowledge which is hard for machine-learning algorithms to detect, and can therefore be used with features extracted from multimedia objects to augment the indexing of that object.

The key to understanding the application of knowledge organisation and machine learning to multimedia information retrieval problems is to consider different types of information needs in particular domains. One particular domain that provides useful examples is the creative domain, where various media are required on a daily basis e.g. video, music (Inskip et al., 2012) and images (Konkova et al., 2016) for advertising campaigns, or images for online news stories (Frankowska-Takhari et al., 2017). A specific example of information needs is

the use of briefs in the advertising world which provide an overview of the media required and some specification of the criteria for the object to be suitable for that particular campaign. Analysis of these briefs has demonstrated that there are some aspects that can be easily detected by machine-learning algorithms, whilst others are too abstract for current techniques to work. For example, in music, Inskip et al. (2012), found that mood was a significant criterion for relevance in music briefs, which would be hard for an algorithm to detect. However, knowledge organisation schemes with human input can help to resolve the need. Inskip et al. (2012) also found that music features such as structure are also important, which machine-learning algorithms can clearly be applied to. In terms of images, Konkova et al. (2016) found three categories of facets in image briefs including syntactic features such as 'colour' and 'texture' as well as high level general and conceptual image features such as 'glamorous' and 'natural'. These aesthetic features are an open problem in the field (Datta et al., 2008). As with music, there is a clear distinction as to which image facets can be detected using machine learning algorithms.

Machine-learning algorithms are very often used to detect features in a variety of different applications (Datta et al., 2008). The full range of algorithms can be found in Datta et al. (2008), Pouyanfar et al. (2018) and Murthy and Koolagudi (2018), but what problems are the algorithms applied to in the context of multimedia IR? Key problems which are addressed in many applications are classification, object detection and annotation. Examples include images where super-human performance has been recorded in the 2015 large scale visual recognition challenge (ILSVRC15) using deep learning methods (Poyyanfar et al., 2018), which has come about due to much improved object recognition (improving the ability to detect objects improves classification techniques). This has also led to techniques that can automatically annotate and tag images, including online services such as Imagga (https://imagga.com/). In music, techniques to apply classification and temporal annotation have been developed at low level (e.g. timbre), mid level (e.g. pitch and rhythm) and high level (e.g. artist and genre) in many music applications (Srinivasa et al., 2018). In video (which is moving images together with sound), problems addressed include event detection by locating scene changes and segmentation of the object into stories e.g. scenes and threads in a TV programme or film (Lew et al., 2006). A quick review of the literature shows that machine learning has been applied to many problems in multimedia successfully, but there are many issues to which the technique cannot be addressed (see above). The key therefore to augmenting any application that uses knowledge organisation as its core with machine learning is to identify the features with which the technique can be used. The features that have been used successfully in the field are the ones that are known to bare fruit given the empirical evidence available. It is to these that we turn to next.

5. Features in multimedia information retrieval

Features are aspects of an object that can be used for multimedia search purposes. The key to the application of search on multimedia objects is to identify these features and provide an index for them, allowing for applications such as direct search and classification or categorisation. In this section we review the features for images, music and video and provide an overview of what machine learning can identify and what is appropriate for knowledge organisation techniques, and when both can be combined. Our emphasis is on combining the features from both sources to improve multimedia search applications and services.

5.1 Image features

There is a wide variety of schemes that identify image attributes such as semantic (e.g. Panofsky/Shatford), syntactic and non-visual (Westman 2009, pp.65-66). Non-visual attributes (such as the metadata e.g. bibliographic data) can be useful (Konkova et al., 2016), but is not the concern here. Semantic information for an image will require human input to establish the 'aboutness' of a given object, through generic schemes such as the Thesauri for Graphic Materials (Library of Congress, N.D.b) and specific schemes such as Iconclass (http://www.iconclass.nl/) which is focused on art images. Syntactic attributes can either be primitive visual elements such as colour, texture, hue and shape, or compositional e.g. the relationship between shapes, motion, orientation, perspective, focal point (Westman, 2009, p.65). It is these syntactic attributes to which machine learning can be applied.

Specific application areas have particular needs. For example, the concept of 'Copyspace' is important in advertising, which is a clear space to insert text (Konkova et al., 2016). Further, studies from the user-centred tradition advocate that human image users in specific domains have specific image needs. Such studies aim to uncover the needs of users and identify which aspects of user needs can be used to facilitate automation of image-based tasks. For example, Frankowska-Takhari et al. (2017) investigated the needs of image users in

online journalism. Initially their findings were similar to those from earlier studies e.g., Markkula and Sormunen (2000) and Westman and Oittinen (2006), and showed that users' descriptions of their image needs were often limited to their conceptual needs, and search queries tend to relate to concepts, while information about users' needs on the perceptual level was limited to descriptions of visual effects required in images. As suggested in Machin and Polzer (2015), it was necessary to reach beyond these descriptions, to identify the concrete visual features that engendered the required effects. Frankowska-Takhari et al. (2017) applied the visual social semiotics framework (Kress and van Leeuwen, 2006) to analyse images used in online journalism. They identified a set of 11 recurring visual features that engender the visual effect required in images used for illustrating news headline content. These included: a strong single focal point to draw readers' attention, the use of specific palette of colours depending on the tone of the news story, a photographic shot from waist-up including head and shoulders and close-up on the face, and a preference for a large object/person in the frame. Most of the identified features are detectable by currently available systems that make use of advanced computer vision. They could be implemented, for example, as multi-feature filters for image retrieval. Such a system. firmly rooted in the image users' needs, could be a step towards automating image retrieval with a purpose to support a specific group of image users carrying out specific illustration tasks.

5.2 Music features

Downie (2002) identifies seven facets of music information that can be considered as features to learn for a retrieval system, which can be further classified into low-level, mid-level and high-level features (Murthy and Koolagudi, 2018). We merged these two schemes together as they provide a useful overall classification of features in which machine learning can be applied and where knowledge organisation schemes are appropriate, as well as identifying the key features. The features are not mutually exclusive (Downie, 2002), and low-level features are used to build mid level features, which in turn can be used to extract high level features (Murthy and Koolagudi, 2018). Low level features are defined as the fundamental property of sound, mid level features the fundamental properties of music and high level features the human perceptual interpretation of the mid level features.

The low-level features are timbre and tempo. Timbre is defined as an attribute related to the tone that differs in the instrument being played (e.g. trumpet vs piano). It is the sound, tone quality and colour that make up the voice quality of a musical note (Murthy and Koolagudi, 2018, p.7). Tempo is defined as the duration between two musical events (e.g. two notes). Timbre and tempo are strongly connected through frames, a short time segment of 10-100ms. These low-level features can fail to capture much information from a given song in their own right (Murthy and Koolagudi, 2018) and mid-level features are required to build up a picture of music which can be used for an application. These mid-level musical features are pitch, rhythm, harmony and melody – note that in our scheme these features are still low level. Pitch is frequency of sound, the oscillations per second. Differences between two pitches are defined as being the interval between them. Harmony is detected when two or more pitches sound at the same time to create polyphonic sound, which is determined by the interval. Rhythm is defined by an occurring or recurring pattern in the music e.g. the beat. Rhythm and pitch determine a further important feature of music namely melody, which is a succession of musical notes. Murthy and Koolagudi (2018) do not classify this feature, but it is clearly a mid-level feature as it strongly related to other mid-level features, but cannot be regard as a high-level feature. It is these mid-level features to which machine learning can be applied.

There is more ambiguity in terms of high-level features, and some can be detected through learning mid-level features, but others require human input. In some, both machine learning and knowledge organisation can be used. High-level features include editing, text, bibliography (Downie, 2002) and artist, genre, instrument and emotion (Murthy and Koolagudi, 2018). Editing is defined as performance instructions of a piece of music such as fingering, articulation etc. Knowledge organisation schemes such as the Library of Congress performance terms for music (Library of Congress, 2013c, 2014d), focused largely on western classical music, are appropriate. Text relates to any lyrics associated with a musical piece and can be handled via normal text retrieval techniques. It may be appropriate to use this feature to augment machine-learning algorithms (in conjunction with natural-language processing techniques). Bibliography refers to the metadata of the piece, which is determined by human entry of aspects such as composer, performer etc. Appropriate metadata standards in the field are appropriate here, and as with text can be used to augment machine-learning algorithms. Bibliography can determine the artist, genre, emotion and instrument features (depending on the metadata scheme used), but machine learning has been used to identify those high-level features from mid-level features extracted from a musical piece e.g. to classify it by the given feature ((Murthy and Koolagudi,

2018). The Genre feature can also be augmented with knowledge organisation schemes such as the Library of Congress music/genre headings (2013a, 2013b).

5.3 Video features

Video is multimedia in the complete sense, as it consists of moving images in sequence with audio. Image features identified in 5.1 above can be used here, and as we have extra evidence (e.g. a serious of images) we have more evidence to improve the detection of objects in the media being indexed. A practical example of the features that can be identified are outdoor and indoor shots, people and landscapes/cityscapes (Smeaton and Over, 2002). There are many features from audio that can be extracted via machine learning including speech to text (where text retrieval techniques can be used) and music (see 5.2 above). Whilst we can build on these features, there are unique features of video that can be used to classify or segment video objects. Video can be split up into scenes and threads (Lew et al., 2006), for example in a news programme where different news stories are presented to the viewer. The TRECVID track at the TREC (Text Retrieval Conference) investigated this in the shot boundary detection task (Smeaton and Over, 2002) by detecting different categories e.g. cut (short finishes, one starts right after), dissolve (one shot fades out while new one fades in), fadeout/in (one shot fades out, then the new one fades in) plus other categories which don't fit into these precise boundaries. Detecting shot boundary allows the detection of higher-level features such as events, embodied in LSCOM (LSCOM, 2011), the large scale concept ontology for multimedia (Naphade et al., 2006). This is a knowledge organisation scheme built via the empirical work carried out by the multimedia community, with TRECVID being particularly notable. Examples include people crying (007), maps (204) and people associated with commercial activities (711). These features can be augmented with other knowledge organisation schemes such as the Library of Congress (N.D.a) scheme for assigning genre/form terms to films and video.

5.4 Summary of features

In this section we have identified two classes of features, one to which machine learning can be applied and one to which it cannot. The low-level features such as colour and hue in images, pitch and tempo in music and shot boundaries in video are ones that can be extracted using machine-learning techniques, whilst high-level features such as 'aboutness' require the use of human intervention via the application of knowledge organisation schemes. Next we consider the use of these different classes of features in conjunction with each other to improve multimedia information retrieval services.

6. Using Machine Learning and Knowledge Organisation to enhance Multimedia Information Retrieval

We propose a process by which the features for a multimedia object are identified (both high level and low level) to create a database of objects for the purposes of retrieval. We assume access to digital objects (analogue objects are not considered here). We identify five steps in this process (see figure 1). In step 1 we identify the corpus and knowledge organisation scheme for the given corpus, which is split into two separate sub-steps: applying the knowledge organisation scheme to the high-level corpus objects (1a) and using machine learning to identify the low-level object features (1b). In step 2, we combine both high and low-level object features to provide a comprehensive set of features for multimedia, which is richer for retrieval purposes (step 3). From step 3 we have the information to create the application of our choice, either a classification or categorisation system, or to support multimedia search functionality (step 4). A further step is considered (step 5), given two scenarios – either a new set of features is identified (by a change in the knowledge organisation scheme or improved feature detection using machine learning) or a new set of objects is received and needs to be indexed. We discuss each of these steps below, highlighting the input and output data for each step.

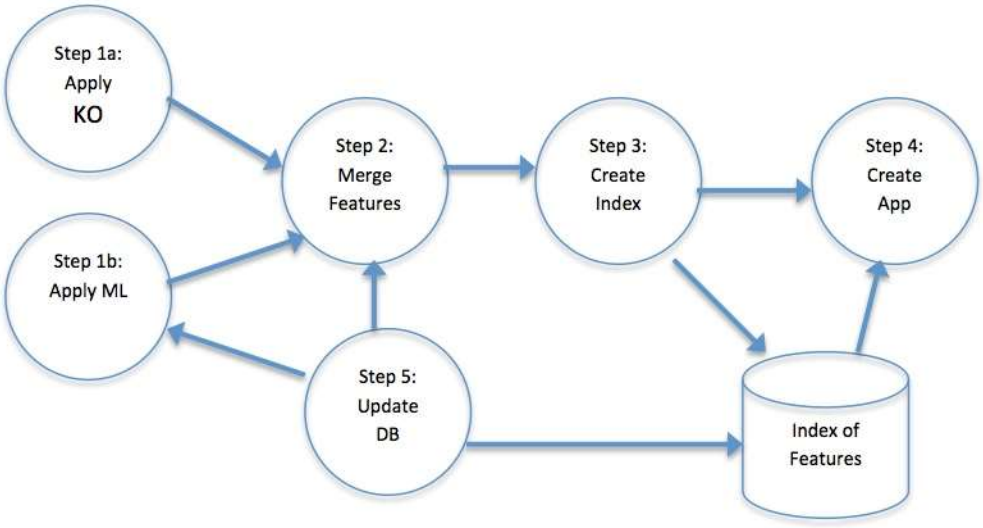

Figure 1 – Process using Knowledge Organisation and Machine learning to index Multimedia

6.1 Step 1A: Apply knowledge organization scheme to corpus

Input Data	**Output Data**
1. Corpus 2. Knowledge organisation scheme	Object features (high level)

Table 1 – Data required for Step 1a

The decision the information professional needs to make is to choose a relevant knowledge organisation scheme for the corpus they are managing. This will either be a standard scheme (examples are cited in section 5 above), or a specialist in-house scheme derived by the organisation that requires access to the multimedia. Collection size is a concern here – unless there are significant human resources, manually cataloguing multimedia objects using the knowledge organisation scheme might not be practical. In this case any metadata associated with the object can be used, with knowledge organisation applied to the metadata to identify relevant features for the database. In other cases, the corpus will already have been indexed (perhaps over many years) and high-level features for each object will be readily available. If the media contains speech (if the corpus is either audio, or video which contains audio), machine learning can be used to detect text to which the knowledge organisation can be applied. Whilst the word error rates might be high, the main bulk of concepts for the objects will be detected. This text might itself be indexed as part of the multimedia search service.

6.2 Step 1B: Apply machine-learning technique to corpus

Input Data	**Output Data**
Corpus	Object features (low level)

Table 2 – Data required for Step 1b

The next step for any information professional is to identify the low level features using machine learning. This may require the assistance of technical staff with AI expertise, but the information professional should be aware of the process used to generate these features. A key decision is to identify training and test objects from the corpus, or a subset of the corpus. The training set is used to detect the features from the corpus, whilst the test set is used to validate the features detected. Getting this right is key, as poor decisions can lead to over-fitting of features, reducing their utility for retrieval purposes. In general the standard way to split the corpus into training and test collections is two thirds for training and one third for testing at least. The training

set should always be much larger than the test set. A further step is to split a corpus into a number of segments (say k), and spilt each of these k segments applying the machine-learning algorithms to each of these segments, by treating each k segment as a test set with other segments as the training set. This can be repeated with all of the segments and the results merged to create a set of features that is more robust. This is known as cross-validation.

The type and size of corpus is a consideration. The professional should consider appropriate features identified in section 5 for their corpus, and the training and test sets should not be to large (in some cases corpora with many millions of objects and large feature sets may be difficult to manage as machine learning is computationally intensive). It should be noted that in order to get an unbiased estimate of how well an algorithm is doing, it was common practice to take all the data and split it according to a 70/30% ratio (i.e., 70/30 train test splits explained above). These ratios were perfectly applied when dealing with small datasets. However, in the Big Data and Deep Learning era, where, the data could exceed millions of instances, the test sets have been becoming a much smaller percentage of the total. For example, if you have a million examples in the dataset, a ratio of 1% of 1 million (99% train, 1% test) will be enough in order to evaluate your machine-learning algorithm and give a good estimate of how well it is performing. This scheme is manageable for large datasets. However, any sample chosen must also be representative, otherwise the features will not be valid. At the end of this step, the low-level object features will be identified.

6.3 Step 2: Merge features for multimedia objects

Input Data	**Output Data**
1. Object features (high level) 2. Object features (low level)	Object features (combined)

Table 3 – Data required for Step 2

The data produced in step 1 from both sub-steps needs to be merged together to create a comprehensive set of features for each object in the multimedia corpus. It is this comprehensive set of features which provides the enhancement required for better multimedia retrieval. Getting the merge process correctly configured therefore is critical, and there are two cases to consider, one straightforward and one that requires a little more thought. The simpler case is the exact match case. In this case we have the same feature identified in both sources (e.g. text extracted from images), and can use that evidence to estimate the probability of the features utility. In most cases, the features will be distinct and the information professional will need to think about which features to record from both sources. They may think it appropriate to record all features, but this may have drawbacks (features may not be useful for search). One way to get around this is to use machine learning to see which low and high-level features correlate with each other, and choose the best set of features – this is the best match approach. This would work by applying a further step of machine learning (as outlined in step 1B above), in which an appropriate sample would be used to generate a set of features for indexing. The advice given in section 6.2 would apply in the best match case. At the end of this step a full set of features appropriate for search will be identified.

6.4 Step 3: Create index of features (database of objects)

Input Data	**Output Data**
Object features (combined)	Database of Objects (Index)

Table 4 – Data required for Step 3

Once a full set of features has been identified, an index of objects using those features can be generated. This can be either an inverted list or a relational or object relational database, depending on the context. The information professional could consult a technical person to assist with this. Examples of software available include Elasticsearch (https://www.elastic.co/), MongoDb (https://www.mongodb.com/), Neo4j (https://neo4j.com/), MySQL (https://www.mysql.com/) and PostgresQL (https://www.postgresql.org/)

6.5 Step 4: Create application or service with combined features

Input Data	Output Data
Database of Objects (Index)	Object classification or categorisation

Table 5 – Data required for Step 4

Once the database has been created, the application or service to meet user needs can be produced. For retrieval purposes. This may just mean writing an appropriate front end given users needs, together with a back end that matches user-defined features identified at the front end. However, if categorisation or classification were required, a further round of machine learning would be appropriate. This would be taking the machine-learning process overviewed in step 1b above, but applying the algorithm to the combined feature set. An example can be found in Fan et al (2007), who combined WordNet and ontology data to support a surgery education application.

6.6 Step 5: Update database of objects with new information

Input Data	Output Data
1. New Objects 2. New Features	1. Updated Database 2. Updated Features and Database

Table 6 – Data required for Step 5

New information is generated all the time, and an information professional cannot assume that the corpus they manage will remain static. There are two scenarios to consider – one where new multimedia objects are received and need to be considered and one were new features are available. The first of these is easy to deal with as features can be assigned (high-level features in the knowledge organisation scheme, low-level features extracted by an algorithm) and the object recorded in the database. The second is not so straight forward and requires a restart of the process – either because new elements have been added to the knowledge organisation scheme or because machine-learning algorithms have been improved to provide a clearer picture of a feature already identified, or to identify new features. This will be an expensive and time-consuming process, so the information professional may wish to test the ideas on a sub-set of the corpus before restarting the whole process again.

7. Conclusion

In this paper we put forward some practical advice for information professionals who curate multimedia digital collections, and who are charged with supporting search services to those collections. We believe that information professionals should treat machine learning and/or AI techniques as an opportunity rather than a threat, and should seriously think about using technology to improve the multimedia services they manage. Information professionals should be wary of the hype that surrounds machine learning/AI that has all too often been exaggerated in terms of impact, leading to AI winters. However, the process we describe in section 6, we believe, gives the information professional an opportunity to seize the initiative and build on their domain knowledge gained in working on images, music and video. We urge the community to consider this when considering access to multimedia digital objects for their users.

8. Acknowledgements

Many thanks to Sven Bale for his advice and clarification of features in music.

9. References

Abadi, Martin, et al. 2016. “Tensorflow: a system for large-scale machine learning”. In *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation* (OSDI 18) eds. Andrea Arpaci-Dusseau and Geoff Voelker. 16: 265-283.

Byrd, Donald and Crawford, Tim. 2002. “Problems of music information retrieval in the real world”. *Information Processing and Management*. 38: 249-272.

Ciresan, Dan C., Giusti, Alessandro, Gambardella, Luca M. and Schmidhuber, Jurgen. 2012. “Deep neural networks segment neuronal membranes in electron microscopy images”. In *Advances in neural information processing systems*, eds. Leon Bottou and Chris Burges. Pp.2843-2851.

Ciresan, Dan C., Giusti, Alessandro, Gambardella, Luca M. and Schmidhuber, Jurgen. 2013. “Mitosis detection in breast cancer histology images with deep neural networks”. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, eds. Terry Peters, Lawrence H. Staib, Sean Zhou, Caroline Essert, Pew-Thian Yap and Ali Khan. Springer, Berlin, Heidelberg, pp.411-418.

Ciresan, Dan C. and Meier, Ueli. 2015. July. “Multi-column deep neural networks for offline handwritten Chinese character classification”. In *International Joint Conference on Neural Networks* (IJCNN 2015), ed. Yoonsuck Choe. IEEE. pp.1-6.

Cullinan, Christopher, Wyant, Christopher and Frattesi, Timothy. 2013. *Computing performance benchmarks among cpu, gpu, and fpga*. URL: www. wpi. edu/Pubs/E-project/Available/E-project-030212-123508/unrestricted/Benchmarking_Final.pdf.

Dai, Na, Shokouhi, Milda and Davison, Brian D. 2011. “Learning to rank for freshness and relevance.” In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, eds. Richardo Baeza-Yates, Tat-Seng Chua and W. Bruce Croft. ACM. pp.95-104.

Datta, Ritendra. Joshi, Dhiraj. Li, Jia and Wang, James Z. 2008. “Image retrieval: Ideas, influences, and trends of the new age”. *ACM Computing Surveys*, 40

Deng, Li. and Yu, Dong. 2014. “Deep learning: methods and applications”*. Foundations and Trends in Signal Processing.* 7(3–4): 197-387. http://dx.doi.org/10.1561/2000000039.

Downie, J. Stephen. 2003. “Music information retrieval”. *Annual review of information science and technology*, 37: 295-340.

Enser, Peter G.B. 2008. “The evolution of visual information retrieval.” *Journal of Information Science*, 34: 531-546.

Fan, Jianping, Luo, Hangzai, Gao, Yuli and Jain, Ramesh. 2007. “Incorporating concept ontology for hierarchical video classification, annotation, and visualization”. *IEEE Transactions on Multimedia*, 9: 939-957.

Frankowska-Takhari, Sylwia, MacFarlane, Andrew, Göker, Ayse and Stumpf, Simone. 2017. “Selecting and tailoring of images for visual impact in online journalism”. *Information Research*, 22: 1.

Hu, Weiming, Xie, Nianhua, Li, Li, Zeng, Xianglin and Maybank, Stephen. 2011. “A survey on visual content-based video indexing and retrieval”. *IEEE Transactions on Systems, Man, and Cybernetics*, Part C (Applications and Reviews), 41: 797-819.

Inskip, Charlie, Macfarlane, Andrew and Rafferty, Pauline. 2012. “Towards the disintermediation of creative music search: analysing queries to determine important facets.” *International Journal on Digital Libraries*, 12: 137-147.

Karpathy, Andrej and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of CVPR 2015, Boston, 7-12 June 2015*, eds. Kristen Grauman, Eric Learned-Miller, Antonio Torralba, and Andrew Zisserman. https://www.cv-foundation.org/openaccess/CVPR2015.py

Konkova, Elena, MacFarlane, Andrew and Göker, Ayse. 2016. “Analysing creative image search information needs”. *Knowledge Organisation* 43:1. 14-21.

Kress, A.G. and Leeuwen, T. van. 2006. *Reading images: the grammar of visual design*. London, UK: Routledge.

Lew, Micheal S, Sebe, Nicu, Djeraba, Chabane and Jain, Ramesh. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2: 1-19.

Library of Congress. 2013a. *Genre/Form Terms for Musical Works and Medium of Performance Thesaurus*. https://www.loc.gov/catdir/cpso/genremusic.html.

Library of Congress. 2013b. *Genre/form terms agreed on by the Library of Congress and the Music Library Association as in scope for Library of Congress Genre/form terms for Library and archival materials (LCGFT).* http://www.loc.gov/catdir/cpso/lcmlalist.pdf.

Library of Congress. 2013c. *Introduction to Library of Congress Medium of Performance Thesaurus for Music*. http://www.loc.gov/aba/publications/FreeLCSH/mptintro.pdf.

Library of Congress. 2013d. *Performance terms: Medium*. http://www.loc.gov/aba/publications/FreeLCSH/MEDIUM.pdf.

Library of Congress. N.D.a. *Library of congress genre/forms for films video.* http://www.loc.gov/aba/publications/FreeLCSH/GENRE.pdf.
Library of Congress. N.D.b. *Thesaurus for Graphical Materials (TGM*). http://www.loc.gov/pictures/collection/tgm/.
LSCOM. 2011. *Large Scale Concept Ontology for Multimedia*. http://www.ee.columbia.edu/ln/dvmm/lscom/
MacFarlane, Andrew. 2016. Knowledge Organisation and its role in Multimedia Information Retrieval. *Knowledge Organisation*. 43(3): 180-183.
Machin, D. and Polzer, L. 2015. *Visual journalism*. London: Palgrave Macmillan.
Markkula, M. and Sormunen, E. 2000. "End-user searching challenges indexing practices in the digital newspaper photo archive". *Information Retrieval*. 1: 259-285.
McCulloch, Warren S and Pitts, Walter. 1943. "A logical calculus of the ideas immanent in nervous activity". *The Bulletin of Mathematical Biophysics*, 5: 115-133.
Murthy, Y.V. Srinivasa. and Koolagudi, Shashidhar G., 2018. "Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review." *ACM Computing Surveys*. 51. Article No. 45.
Naphade, Milind, et al. 2006. "Large-scale concept ontology for multimedia". *IEEE multimedia*, 13: 86-91.
Russell, Stuart J. and Norvig, Peter. 2016. *Artificial intelligence: a modern approach*. Malaysia: Pearson Education Limited.
Pouyanfar, Samira, et al. 2018. "A Survey on Deep Learning: Algorithms, Techniques, and Applications". *ACM Computing Surveys*. 51: Article No. 92.
Smeaton, Alan F. and Over, Paul., 2003. "The TREC-2002 video track report". *Journal of Research*, 251.
Singer, Graham. 2013. "The history of the Modern Graphics Processor". *TechSpot*. 27 March.
Westman, Stina. 2009. "Image users' needs and searching behaviour". In *Information Retrieval in the 21st Century*, eds. Ayse Goker, and John Davies. John Wiley & Sons, 63-83.
Westman, S. and Oittinen, P. 2006. "Image retrieval by end-users and intermediaries in a journalistic work context". In *Proceedings of the 1st International Conference on Information Interaction in Context*, IIiX 2006, Copenhagen, Denmark, October 18-20, 2006. Pp.102 – 110.

Author Information

Andrew MacFarlane is a Reader in information retrieval in the Centre for HCI Design, Department of Computer Science at City, University of London. He got his PhD in information science from the same institution. His research interests currently focus on a number of areas including image retrieval, disabilities and information retrieval (dyslexia in particular) and AI techniques for information retrieval and filtering. He was principle investigator for the PhotoBrief project, which focused on meta-data an images and is current involved in the DMNIR project, which is investigating information verification tools for journalists.

Sondess Missaoui is a Postdoctoral Research Fellow in Information Retrieval in the Centre for Human-Computer Interaction Design, City University of London. She graduated in Computer Science at the University of IHEC Carthage (Tunisia) and she obtained her Ph.D. in Computer Sciences at the University of Milano-Bicocca, Department of Informatics, Systems, and Communication (Italy). Her research interests are Recommender Systems, Information Retrieval, Mobile Information Retrieval, Context-Awareness, User Profiling. Currently, she is working on a research project (DMINR) that aims to create a digital tool for researching and verifying stories. She focus on a number of areas including Aggregated search, Natural Language Processing, and Deep Learning.

Sylwia Frankowska-Takhari holds MA in Linguistics and Information Science from University of Poznan, Poland (2001) and MSc in Human-Centred Systems from City, University of London (2011). She completed her PhD under the supervision of Dr Andrew MacFarlane and Dr Simone Stumpf, at the Centre for HCI Design, Department of Computer Science at City, University of London. Her key research interest are information behaviour and image retrieval. Sylwia's PhD work investigates the information behaviour and image needs of professionals working in creative industries with a particular focus on how images are selected, used and tailored in online journalism.