On Machine Learning and Knowledge Organization in Multimedia Information Retrieval†

Andrew MacFarlane*, Sondess Missaoui**, and Sylwia Frankowska-Takhari***

Centre for HCI Design, Department of Computer Science, City, University of London, UK
*<andym@city.ac.uk>, **<Sondess.Missaoui@city.ac.uk>, ***<Sylwia.Frankowska.1@city.ac.uk>

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 the principle investigator for the PhotoBrief project, which focused on meta-data and 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 PhD 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, and user profiling. Currently, she is working on a research project (DMINR) that aims to create a digital tool for researching and verifying stories. She focuses on a number of areas including aggregated search, natural language processing, and deep learning.

Sylwia Frankowska-Takhari holds a MA in linguistics and information science from the University of Poznan, Poland (2001) and a 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 interests 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.

MacFarlane, Andrew, Sondess Missaoui and Sylwia Frankowska-Takhari. 2020. "On Machine Learning and Knowledge Organization in Multimedia Information Retrieval." *Knowledge Organization* 47(1): 45-55. 44 references. DOI:10.5771/0943-7444-2020-1-45.

Abstract: Recent technological developments have increased the use of machine learning to solve many problems, including many in information retrieval. Multimedia information retrieval as a problem represents a significant challenge to machine learning as a technological solution, but some problems can still be addressed by using appropriate AI techniques. We review the technological developments and provide a perspective on the use of machine learning in conjunction with knowledge organization to address multimedia IR needs. The semantic gap in multimedia IR remains a significant problem in the field, and solutions to them are many years off. However, new technological developments allow the use of knowledge organization 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 music and video can be used in conjunction with knowledge organization to tag or label multimedia content for better retrieval performance. We provide an overview of the use of knowledge organization schemes in machine learning and make recommendations to information professionals on the use of this technology with knowledge organization techniques to solve multimedia IR problems. We introduce a five-step process model that extracts features from multimedia objects (Step 1) from both knowledge organization (Step 1a) and machine learning (Step 1b), merging them together (Step 2) to create an index of those multimedia objects (Step 3). We also overview further steps in creating an application to utilize the multimedia objects (Step 4) and maintaining and updating the database of features on those objects (Step 5).

Received: 27 August 2019; Revised: 14 October 2019, 8 November 2019; Accepted: 15 November 2019

Keywords: features, machine learning, knowledge organization, multimedia

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

1.0 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 uses 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, and there is a large body of research for this domain going back many years (Smiraglia and Cai 2017). The same however, could not be said of multimedia information retrieval where many challenges are still evident, although technological developments are beginning to change the situation. By multimedia retrieval we mean search for non-text objects such as images, pieces of music (Byrd and Crawford 2002) or videos/moving images (Hu et al. 2011). 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 recognizable to humans—the example being a picture of a politician in an election. Whilst the politician can be easily identified (the "ofness" of the image), the election is a more nebulous concept that is difficult to extract from an image, without context (the "aboutness" of the image). 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 organization methods are essential to ensure that these conceptual features are captured and recorded in multimedia software and services.

In this paper, we address the technological changes that have led to the potential for improvements in multimedia search and argue that knowledge organization can be used together with a supervised learning technique. We then review the landscape of multimedia search and show some possibilities for using knowledge organization 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 features and machine learning in conjunction with knowledge organization in multimedia IR systems and services. We provide some ideas for the way forward together with the practical implications for knowledge organization practitioners. The contribution of the paper is a process model that uses knowledge organization schemes and machine learning algorithms to create a database of objects for the purposes of multimedia information retrieval. The proposed process model 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.

2.0 Machine learning and technological developments for machine learning

What are the key developments that have led to improvements in technology, and which have significant implications for the use of knowledge organization 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 you would 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 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 2019). 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 the deployment of 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 at. 2013) and handwritten Chinese character recognition (Ciresan and Meier 2015).

Such advances in machine learning methods, including 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 organization, machine learning can be used to provide better and more relevant results to users for a given set of needs that require the identification of specific objects to resolve or address that information need. This paper puts forward an argument for a supervised learning approach in multimedia search, where a knowledge organization 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. We review the overall approach we advocate when using machine learning in conjunction with knowledge organization next.

3.0 Machine learning and knowledge organization

As feature extraction from various media has improved in recent years through developments overviewed in Section 2, what are the implications for the use of knowledge organization techniques? Knowledge organization 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 organization 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, 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 organization techniques to the problem. Semi-supervised learning (Russell and Norvig 2016, 695) does have some access to some 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. We can either match features detected by both the machine learning algorithm and the practitioner (exact match case) or estimate the probability of a features matching from both sources using supervised learning techniques (best match case). We consider both examples later on the paper in Section 6. In this paper, we focus on the user of knowledge organization and supervised learning in multimedia search, in the context of large amounts of data that have been labelled by practitioners. The scope of our work is in the use of non-symbolic AI methods (such as neural networks), rather than symbolic methods deployed in prior work when knowledge organization has been used with machine learning, e.g., in expert systems (Lopez-Suarez and Kamel 1994).

4.0 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, 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), have provided scope to improve multimedia search results using knowledge organization . In MacFarlane (2016), we argue that media of various kinds (e.g., images, music) requires cultural knowledge that can often be only expressed tacitly and require human input. The advantage of knowledge organization schemes is that they provide this knowledge that 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 organization 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 is required on a daily basis, e.g. video, music (Inskip et al. 2012) and images (Konkova et al. 2016), for advertising campaigns, 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 organization 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 feature 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 (Datta et al. 2008) in a variety of different applications. 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 that are addressed in many applications are classification, object detection and annotation. Examples include images where superhuman performance has been recorded in the 2015 large scale visual recognition challenge (ILSVRC'15) 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 (Murthy and Koolagudi 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 multi-media 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 organization as its core with machine learning, is to identify the features to which the technique can be used. It is these features that have been used successfully in the field that are known to bear fruit given the empirical evidence available. It is to these that we turn to next.

5.0 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 organization 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 for modelling image retrieval. These include semantic (e.g., Panofsky/Shatford), syntactic and non-visual attributes (Westman 2009, 65-66). While non-visual attributes (such as the meta-data, e.g. bibliographic data) can be useful (Konkova et al. 2016), this is not the concern here, as we focus on the semantic and syntactic features. One of the earliest frameworks is Panofsky's theory (Panofsky, 1962) that describes three levels of meaning in a work of art: pre-iconographical, iconographical and iconological. Shatford (1986) extended this model and proposed that semantic information in an image may be analysed on the level of generic and specific elements present in the image (the "ofness" of the image), and on the level of the abstract themes present in the image (the "aboutness" of the image). While describing the "ofness" involves decoding and naming of the objects in the image, interpreting the "aboutness" from the image, especially, an image rich in symbolic meaning (e.g., a work of art), requires previous personal, cultural knowledge and experience from the viewers. Therefore, semantic information for an image will require human input to establish the "aboutness" of a given object. Currently, this can be done 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/) that is focused on art images. While most existing frameworks stem from the Panofsky/Shatford matrix (Shatford 1986), the more recent models (e.g., Eakins et al. 2004; Hollink et al. 2004; Jaimes and Chang 2000) allow the distinction between the semantics and syntax of images. Syntactic attributes can either be primitive visual elements such as colour, texture, hue and shape, or compositional, e.g., relationship between shapes, motion, orientation, perspective, focal point (Westman 2009, 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), 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 eleven recurring visual features that engender the visual effect required in images used for illustrating news headline content (see Table 1). 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 to 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 merge these two schemes together as they provide a useful overall classification of features in which machine learning can be applied and where knowledge organization 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.

Feature	Visual image features
1	The specific (identifiable) person/people related to the topic depicted in the image
2	The person/people depicted in the foreground
3	Shot from waist up
4	Face visible: frontal or profile shot
5	Gaze: direct or side gaze
6	The depicted person is "large" in the frame
7	Positioned centrally or to the right within the frame
8	Colour image
9	Colour intensity: saturated or soft colours used
10	Blurry or monotone background
11	The person's face in focus (sharp)

Table 1. Image features recurring in news headline images. Source: Frankowska-Takhari et al. (2017).

The low-level features are timbre and tempo. Timbe 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, 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 that 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 regarded 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 organization 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 organization schemes such as the Library of Congress performance terms for music (Library of Congress 2013c; 2013d) 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 meta-data of the piece, which is de-

termined by human entry of aspects such as composer, performer, etc. Appropriate meta-data standards in the field are applied 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 meta-data 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 organization 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 series 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 (sort finishes, one starts right after), dissolve (one shot fades out while new one fades in), fadeout/in (one shot fades out then 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 (http://www.ee.columbia.edu/ln/dvmm/lscom/), the large-scale concept ontology for multimedia (Naphade et al. 2006). This is a knowledge organization 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 organization 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 which 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 organization schemes. Next, we consider the use of these different classes of features in conjunction with each other to improve multimedia information retrieval services.

6.0 Using machine learning and knowledge organization to enhance multimedia information retrieval

We propose a process model 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 model (see Figure 1). In Step 1, we identify the corpus and knowledge organization scheme for the given corpus, which is split into two separate sub-steps: applying the knowledge organization 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 categorization 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 organization 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.

6.1 Step 1a: apply knowledge organization scheme to corpus

Input Data	Output Data
1. Corpus 2. Knowledge Organization Scheme	Object features (high-level)

Table 2. Data required for Step 1a.

The information professional needs to choose a relevant knowledge organization scheme for the corpus they are

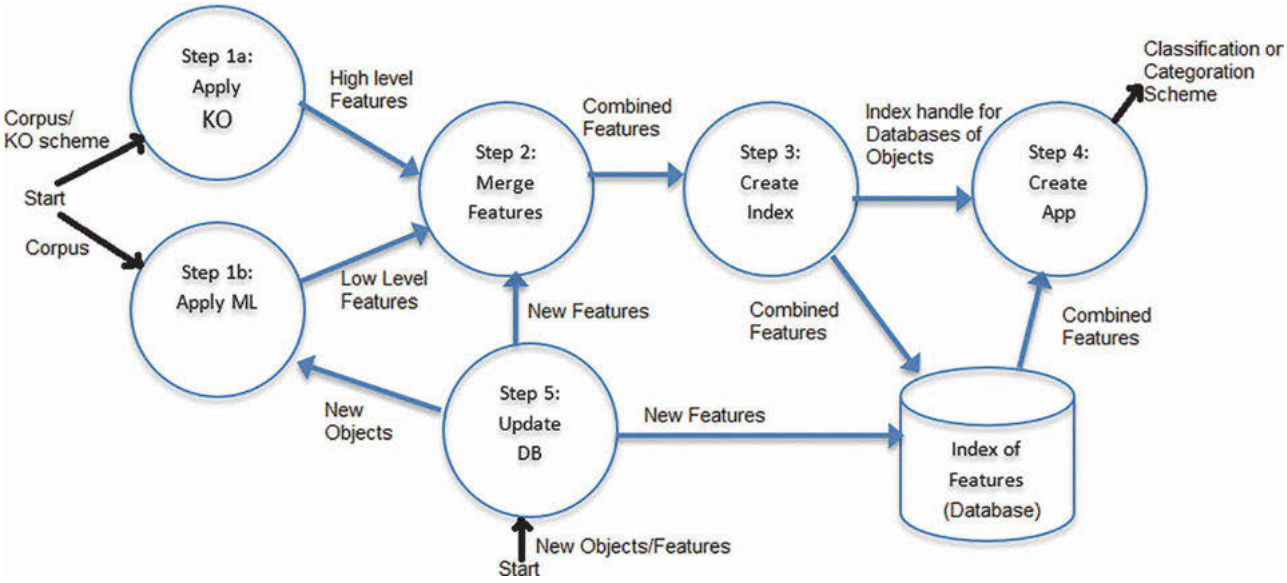

Figure 1. Process using knowledge organization and machine learning to index multimedia.

managing. This will either be a standard scheme (examples are cited in Section 5 above), or a specialist in house scheme derived by the organization 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 organization scheme might not be practical. In this case any meta-data associated with the object can be used, with knowledge or ganization applied to the meta-data 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 that contains audio), machine learning can be used to detect text, on which the knowledge organization 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.

An example to illustrate this is from the advertising domain. Konkova et al. (2016) provide a list of facets for images in which knowledge organization elements can be placed. Examples of this are image style and conceptual features, which are very subjective and require human input. Image style could include glamour, whether it is natural or manipulated (using photoshop), amateur or professionally taken photo. Conceptual features could include positive busy images of bustling street life, innocence/guilt, freedom/slavery, beauty/ugliness, etc (the "aboutness"). General semantics of what is in the image could also be detected, e.g., beautiful images of clouds on the planet Jupiter, family walking together on a beach, etc.

6.2 Step 1b: apply machine learning technique to corpus

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

Table 3. 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 too large (in some cases corpuses with many millions of objects and large features 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 good your algorithm is doing, it was common practice to take all your 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 one million or so (99% train, 1% test) will be enough in order to evaluate your machine learning algorithm and give you a good estimate of how well it's doing. 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.

An example to illustrate this is from the advertising domain. Konkova et al. (2016) identifies a list of facets ripe for the application of machine learning. Composition of the image can be detected such as shooting distance (close up, panoramic view of a landscape), angle (shot taken from the left of a subject), object location (lamp on a desk) or focus (sharp, blurred). Light is a related facet where the time of day can be detected (shadows), type of light (natural, artificial) and by location (outside or inside shot). Specific semantics including particular entities/places/people can be detected, e.g., a human hand holding an archaeological artifact, a shot of St Peters Basilica in Rome, etc.

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 4. 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 that provides the enhancement required for better multimedia retrieval. Getting the merge process correctly configured is, therefore, 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, split into conjoint and disjoint sub-cases. In conjoint sub-case, we have the same feature identified in both inputs (e.g., text extracted from images may match a term in the knowledge organization scheme) and record that feature in the index. In most cases, the features will be distinct (the disjoint sub-case—a feature is identified either by the knowledge organization scheme OR by the machine learning algorithm) and the information professional will need to think about which features to record. 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 of the low and high-level features correlate with each other in the input dataset and choose the best set of features—this is the best match approach. In this, either all inputs from both sources or from the disjoint sub-case could be used. 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, a full set of features appropriate for search will be identified. There are many different contexts to consider, and the information professional will need to be clear about the particular implications for their given situation.

Taking the example given from the advertising domain above (Konkova et al. 2016), this would appear to be disjoint and the features about any given image object can be merged together quickly and easily. The facets and their qualities are really quite different and distinct, and it is clear which process will create the appropriate image description for that facet. It should be noted that improvements in machine learning may address the general semantics facet, which may need reviewing by the image indexer.

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

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

Table 5. 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
Link to Database of Objects (Index)	Object classification or categorisation

Table 6. 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 7. 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 organization scheme, low-level features extracted by an algorithm) and the object recorded in the database. The second is not so straight forward and it requires a restart of the process—either because new elements have been added to the knowledge organization 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.

7.0 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 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 overhyped in terms of impact, leading to AI winters. However, the process model 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.

References

Abadi, Martin. Paul Barham, Jianmin Chen, Zhifeng Chen andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vigay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu and Xiaoqiang Zheng. 2016. "Tensorflow: A System for Large-Scale Machine Learning." In *Proceedings of OSDI '16: 12th USENIX Symposium on Operating Systems Design and Implementation November 2–4, 2016 Savannah, GA, USA,* ed. Andrea Arpaci-Dusseau and Geoff Voelker. 16. Carlsbad, CA: Usenix, 265-83.

Byrd, Donald and Tim Crawford. 2002. "Problems of Music Information Retrieval in the Real World." *Information Processing & Management* 38: 249-72.

Cireşan, Dan C., Alessandro Giusti, Luca M. Gambardella and Jurgen Schmidhuber. 2012. "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images." In *Advances in Neural Information Processing Systems 25: 25th Annual Conference on Neural Information Processing Systems 2011, December 12-15, 2011, Granada, Spain*, ed. Fernando Pereira, Chris Burges, Leon Bottou and Kilian Weinberger. La Jolla, CA: Neural Information Processing Systems; Red Hook, NY: Printed from e-media with permission by Curran Associates, 2843-51.

Cireşan, Dan C. Alessandro Giusti, Luca M. Gambardella and Jurgen Schmidhuber. 2013. "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, ed. Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot and Nassir Navab. 16. Lecture Notes in Computer Science 8150. Berlin: Springer-Verlag, 411-8.

Cireşan, Dan C. and Ueli Meier. 2015."Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification." In *International Joint Conference on Neural Networks,* ed. Yoonsuck Choe. New York: IEEE. doi:10.1109/IJCNN.2015.7280516

Cullinan, Christopher, Christopher Wyant and Timothy Frattesi. 2019. "Computing Performance Benchmarks among CPU, GPU and FPGA." Accessed November 8. http://www.wpi.edu/Pubs/E-project/Available/E-project-030212-123508/unrestricted/Benchmarking_Final.pdf

Dai, Na., Milda Shokouhi and Brian D. Davison. 2011. "Learning to Rank for Freshness and Relevance." In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ed. Richardo Baeza-Yates, Tat-Seng Chua and W. Bruce Croft. 34. New York: ACM, 95-104. doi:10.1145/2009916.2009933

Datta, Ritendra, Dhiraj Joshi, Jia Li and James Z. Wang. 2008. "Image Retrieval: Ideas, Influences and Trends of the New Age." *ACM Computing Surveys* 40, no. 2: article no. 5. doi:10.1145/1348246.1348248

Deng, Li and Dong Yu. 2014. *Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing.* Delft: Now.

Downie, J. Stephen. 2003. "Music Information Retrieval." *Annual Review of Information Science and Technology* 37: 295-340.

Eakins, John P., Pam Briggs and Bryan Burford B. 2004. "Image Retrieval Interfaces: A User Perspective." In: *Proceedings of the 3rd International Conference on Image and Video Retrieval.* ed. Peter Enser, Yiannis Kompatsiaris, Noel E. O'Connor, Alan F. Smeaton and Arnold W. M. Smeulders. Lecture Notes in Computer Science 3115. Berlin: Springer, 628-37.

Enser, Peter G.B. 2008. "The Evolution of Visual Information Retrieval." *Journal of Information Science*, 34: 531-46.

Fan, Jianping, Hangzai Luo, Yuli Gao and Ramesh Jain. 2007. "Incorporating Concept Ontology for Hierarchical Video Classification, Annotation and Visualization." *IEEE Transactions on Multimedia* 9: 939-57.

Frankowska-Takhari, Sylwia. Andrew MacFarlane, Ayse Göker and Simone Stumpf. 2017. "Selecting and Tailoring of Images for Visual Impact in Online Journalism." *Information Research* 22, no. 1. http://informationr.net/ir/22-1/colis/colis1619.html

Hollink, Laura, Guss Schreiber, Bob J. Wielinga and Marcel Worring. 2004. "Classification of User Image Descriptions." *International Journal of Human Computer Studies* 61: 601-26.

Hu, Weiming, Nianhua Xie, Li Li, Xianglin Zeng and Stephen Maybank. 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. Andrew Macfarlane and Pauline Rafferty. 2012. "Towards the Disintermediation of Creative Music Search: Analysing Queries to Determine Important Facets." *International Journal on Digital Libraries* 12: 137-47.

Karpathy Andrej and Li Fei-Fei. 2015. "Deep Visual-Semantic Alignments for Generating Image Descriptions." In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*, ed. Kristen Grauman, Eric Learned-Miller, Antonio Torralba and Andrew Zisserman. Boston: IEEE, 3128-37. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.pdf

Konkova, Elena. Andrew MacFarlane and Ayse Göker. 2016. "Analysing Creative Image Search Information Needs." *Knowledge Organization* 43: 14-21.

Jaimes, Alejandro and Shih-Fu Chang. 2000. "A Conceptual Framework for Indexing Visual Information at Multiple Levels." In *Proceedings Volume 3964 Electronic Imaging | 22-28 January 2000 Internet Imaging*, ed. Giordano B. Beretta and Raimondo Schettini. 396. doi:10.1117/12.373443

Kress, Gunter and T. van Leeuwen. 2006. *Reading Images: The Grammar of Visual Design*. London: Routledge.

Lew, Michael S., Nicu Sebe. Chabane Djeraba and Ramesh Jain. 2006. "Content-Based Multimedia Information Retrieval: State of the Art and Challenges." *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)* 2: 1-19. doi:10.1145/1126004.1126005

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. 2019a "Library of Congress Genre/Forms for Films Video." Accessed November 8. http://www.loc.gov/aba/publications/FreeLCSH/GENRE.pdf

Library of Congress. 2019b "Thesaurus for Graphical Materials (TGM)." Accessed November 8. http://www.loc.gov/pictures/collection/tgm/

Lopez-Suarez, Alex. and Mohammed S. Kamel. 1994. "Dykor: a Method for Generating the Content of Explanations in Knowledge Systems." *Knowledge-Based Systems* 7: 177-88.

MacFarlane, Andrew. 2016. "Knowledge Organization and its Role in Multimedia Information Retrieval." *Knowledge Organization* 43: 180-3.

Machin, David and Lydia Polzer. 2015. *Visual Journalism. Journalism: Reflections and Practice*. London: Palgrave.

Markkula, Marjo and Eero Sormunen. 2000. "End-User Searching Challenges Indexing Practices in the Digital Newspaper Photo Archive." *Information Retrieval* 1: 259-85.

McCulloch, Warren S. and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115-33.

Murthy, Y.V. Srinivasa and Shashidhar G. Koolagudi, 2018. "Content-Based Music Information Retrieval (CB-MIR) and its Applications toward the Music Industry: A Review." *ACM Computing Surveys* 51, no. 3: article no. 45. doi:10.1145/3177849

Naphade, Milind. John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu. Lyndon Kennedy, Alexander Hauptmann and Jon Curtis. 2006. "Large-Scale Concept Ontology for Multimedia." *IEEE multimedia* 13: 86-91.

Russell, Stuart J. and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. 3rd ed. [UK]: Pearson.

Panofsky, Erwin. 1962. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. New York: Harper & Row.

Pouyanfar, Samira, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen and S.S. Iyengar,. 2018. "A Survey on Deep Learning: Algorithms, Techniques and Applications." *ACM Computing Surveys* 51, no. 5. doi: 10.1145/3234150

Shatford, Sara. 1986. "Analyzing the Subject of a Picture: A Theoretical Approach." *Cataloging and Classification Quarterly* 6: 39-62.

Smeaton, Alan F. and Paul Over. 2003. The TREC-2002 Video Track Report. *The Eleventh Text Retrieval Conference (TREC 2002)*, ed. Ellen Voorhees. NIST Special Publication SP 500-251. Gaithersburg, MD: NIST. https://trec.nist.gov/pubs/trec11/t11_proceedings.html

Smiraglia, Richard P. and Xin Cai. 2017. "Tracking the Evolution of Clustering, Machine Learning, Automatic Indexing and Automatic Classification in Knowledge Organization." *Knowledge Organization* 44: 215-33.

Singer, Graham. 2019. "The History of the Modern Graphics Processor." *TechSpot* (blog), November 21. https://www.techspot.com/article/650-history-of-the-gpu/

Westman, Stina. 2009. "Image Users' Needs and Searching Behaviour." In *Information Retrieval in the 21st Century*, ed. Ayse Goker and John Davies. Chichester: John Wiley & Sons, 63-83.

Westman, Stina and Pirkko Oittinen. 2006. "Image Retrieval by End-Users and Intermediaries in a Journalistic Work Context." *Proceedings of the 1st IIiX Symposium on Information Interaction in Context*, ed. Ian Ruthven. 1, New York: ACM, 102-10.