Application of Meta-learning Principles in Multimedia Indexing

Petr Pulc[1] and Martin Holeňa[2]

[1] Faculty of Information Technology, Czech Technical University,
Prague, Czech republic
petr.pulc@fit.cvut.cz,
[2] Institute of Computer Science, Academy of Sciences,
Prague, Czech republic

Abstract. Databases of video content traditionally rely on annotations and meta-data imported by a person, usually the uploader. This is supposedly due to a lack of an universal approach to the automated multimedia content annotation. As it may be hard or impossible to find a single classifier for all encountered combinations of different modalities or even a network of the classifiers, current interest of our research is to use meta-learning for multiple stages of the multimedia content classification. With this, we hope to handle correctly all modalities involved including their overlaps. Successively, the extracted classes will be used to build the index and later used for searching and discovery in the multimedia.

Keywords: multimedia, index, database, meta-learning, classification

1 Introduction

Most of the platforms, that store multimedia content, use some form of textual annotations for easy and quick indexing and searching. In the last few years, image and music databases have also enabled users to query by examples [21]. Also, thanks to methods that are able to describe individual objects in the image [5, 7] and possibly also actions, not annotated images can be found by a text query as well.

However this does not hold for video, which still mostly relies on title and description filled by the person who uploads it. Although there have been attempts to recognize activities of people [16, 17] and a lot of other high-level features, they are tightly fixed to specific conditions and therefore not of much use on typical hand-held camera footage, for example.

Data modalities as well as currently extracted high-level features will be presented in Section 2 of this paper. Information that we propose to be stored in the future index will be discussed in the Section 3.

1.1 Video Processing

To gather most information for further processing, the easiest solution would be to use all methods for individual modalities we have at our disposal. Analyse

the sound, moving picture and possibly also closed captions, if the multimedia includes them. After that, simply combine the outputs and present to the user or store to index.

As we have tested in our previous work, this can work well if all of the outputs create data with a homogeneous meaning. For example, in a case of lecture recording, automated speech recognition on audio signal returns a transcript and optical character recognition on video – that consists only of slides for sake of simplicity – yields the major keywords, equations, etc. This can result to a single document in which both indexing and searching makes sense.

Even in this oversimplified case, there are however some major issues: How to recognize that the incoming sound is in fact speech and we should transcribe it? And that the pictures we are getting on the input are really slides and character recognition will not be executed on objects only similar to letters?

One way would be to run really all methods we can and then select the ones with best accuracy. Although this is very wasteful, it is a possible solution.

In this case, it is also superfluous to run character recognition on all frames. Either framerate subsampling or detection of transitions can be used to eliminate most of the frames which are otherwise close to identical. But in the case of other multimedia content, text recognition may be required on a level of individual frames, so this decision has to depend on the particular input.

Therefore, we need an expert, that would recommend us beforehand, what subsections of the multimedia may be of our further interest. Based on this information, a set of algorithm pipelines may be prepared to process each pre-selected piece of the media. We will try to propose such expert in Section 4 of this paper.

1.2 Use of Meta-learning

Meta-learning helps the further processing to better understand the data it gets on input. As such, it creates and continually evolves a model, where the output is not directly connected to target classes, but rather to selection of methods how to extract the final information. As this is a classification problem, we will be using similar terminology, just with the "meta-" prefix. Therefore meta-features are the inputs to such classifiers and on output we gather a meta-knowledge.

In this paper, we will use meta-learning for two different purposes:

In Data Processing As it would not be practical to prepare each and every possible data extraction scenario by human expert, we better prepare a layer of data extraction, pre-processing and classification to behave as a recommender instead. As these classifiers will "learn how to learn" the subsequent layers of data processing, we may call this a meta-learning according to [2, section 1.2.3].

As the meta-knowledge can propose a relation between multiple modalities and final outcome, the further processes may benefit from a wider range of information for its decisions. Another advantage of using a meta-learning is, that the gathered meta-knowledge may be also relatively easy transferred to

other systems, opposed to classification models. Basically, as long as the system uses the same meta-features for the meta-learning as the original system.

Once a new meta-knowledge of the data extraction and processing graph is gathered, there may be also a possibility to share such information with specialized extractions used not for video, but for the individual modalities as well. For this, the features have to be also mapped to the individual modalities to enable the selection of appropriate ones.

In Prediction Modelling In the data processing, classifiers are commonly used for pattern recognition and data segmentation. Once we are able to assign a description to some subspace of a feature space, all the incoming items can be described in a same manner.

However, the space cannot be divided arbitrarily, as we have to keep generalisation properties of the classifier. For that, models have to be trained on the incoming data. Usually, we have to set-up classification algorithm and parameters tuned to optimal decision boundaries, which may be again a try-and-fail process.

In this case, meta-learning can be used for recommendation of those parameters, based on previously processed datasets.

2 Multimedia Modalities and Data Extraction

By definition, multimedia content combines multiple media delivering the message. Currently, the prevalent form of consumed multimedia includes audio and video, where one or both of the modalities carry the information. In some cases, the multimedia is also accompanied by text, either in form of an annotation (and so describing the multimedia as a whole) or as lyrics or subtitles, which adds the information about approximate correspondence timing.

As our goal is to extract information in form of text or other easily indexable and searchable data, text input can be transferred pretty much directly. We will consider implementation of some text-mining methods, such as [3], later in future. Currently, we will focus mainly on data extraction from "pure" multimedia, especially on audio and video.

2.1 Audio

Audio signal is actually an encoded sound pressure at a given time. Audio track may consist of multiple channels that are meant to be played together to create an illusion of space (mastered track), or may carry different content (separate instruments, individual microphones). Sometimes, there is also a possibility of multiple language mutations, but media containers usually carry the appropriate information and keep the audio separated.

When working with audio, we have to be also aware of few possible problems. Sound can contain a noise or hum captured during recording (background noise) or generated by bad amplification, storage and reproduction. Sound can be also

a subject to reverberation when recorded along with its reflections or distortion when the level of incoming sound exceeds recording threshold. On top of that, mastered records usually contain layering of multiple instruments that may be impossible to decompose back.

As it does not make sense to work with low-level audio signal, a set of descriptors and classifiers have been created throughout the time. The most widely used – MPEG 7 – has been also standardised [1].

The audio descriptors may contain, for example, following information:

Temporal from signal energy: Attack time, Decrease, Centroid, Effective Duration and others

Spectral from signal frequencies: Centroid, Skewness, Kurtosis, Slope, Decrease, Variation

Harmonic created by sinusoidal modelling: Fundamental Frequency, Noisiness, Odd-to-Even Harmonic Ratio

Perceptual computed using human hearing model: Mel Frequency Cepstral Coefficient, Loudness, Specific Loudness, Sharpness, Roughness

Processing of music and speech differs a lot. Even in our data extraction decision we will need to differentiate between these tasks. Such problem is discussed for example in [18], however spoken text with background music is commonly misclassified.

Music Combination of descriptors mentioned above are used for several tasks in music processing. For example, instrument detection [15], genre classification [9] or discovery of similar music [10].

Speech In speech signal, we may be also interested in the tonality of the speech, as this may help us in speaker distinguishment [8].

However we are usually far more interested in the content of the speech, and therefore methods of automatic speech recognition have been created. Such methods usually use Hidden Markov Models to transform the signal from a frequency spectra into individual phonemes or even words. Such extracted data can be almost directly indexed and used.

2.2 Video

Video signal is far more complex. Technically, we have to deal with amount of light hitting a particular section of a plane in time. Practically, we acquire such light through a Bayer mask usually in three channels: red, green and blue. Signal is then mostly stored in the YUV colour space and U (B−Y) and V (R−Y) channels are also usually subsampled.

Video also brings lot more troubles: colours may be shifted, because reference to white may be changing even during one shot, modern CMOS chips still induce a rolling shutter effect, older CCD chips were sensitive to burn-ins, optics of the camera induce distortions and vignetting and depending on a shutter time both camera shake and motion blur may be present.

Single Frame Many of the video processing approaches are based on processing single frame at a time. As there are many image processing methods, all you need to do is to run them on all frames and either use the result as a time sequence or use only some statistic of these data. Examples of such methods include classification of textures [19], bag-of-features classification [13], text recognition [12], object recognition [4] or face recognition [20].

Multiple Frames As the resolution of the video signal is usually significantly smaller than of static photos, some of the above-mentioned methods may require multiple video frames (or fields in case of interlaced video) to gather enough structural information. This approach is known as a super-resolution [14] and is used in multiple areas of image and video processing.

Sequence of multiple frames also introduces a concept of motion detection, object tracking and more precise object classification [6]. These are the methods that usually require a fixed viewing angle and position of the camera. On the other hand, there are available more and more intricate methods of motion stabilization or smoothing [11] that use the motion information for a completely different purpose.

3 Target Information for the Index

Information that we are trying to acquire from the multimedia for indexing may differ significantly according to the final use. Some of the extracted information are crucial for video editors, but not of much interest for target audience. Example of such information may be a shot size – with what level of detail is an object seen in the picture.

Also the extraction methods may differ based on the target audience, therefore we chose two main scenarios that we are working on:

The first use-case is an extension of search possibilities in published multimedia material. We propose that index should keep information about spoken text along with information about speakers, detected objects or people. Where applicable, human actions and events. We have to be aware, that some of the public videos consist only of a sound track and visualization. Such multimedia should be found to have no correlation between audio and video in the meta-learning, and therefore only the audio should be processed.

Second example of data to index is a raw or only partially processed material used in film making and documentary. It may be required to have the above mentioned features in the index, along with other features: visual classification of an indoor/outdoor or seasonality of the shot, camera shot size, angle and movement, sound layout or linkage between different versions of the material: unedited (raw), dubbed, colour corrected, cut, ...

Some of these features can be represented by a text-like label. In such cases, groups of authors use either an existing standard or agreement. For example, commonly recognised shot sizes are: Very Long Shot, Long Shot, Medium Long Shot, Medium Shot, Medium Close Shot, Close Shot, Close-up and Extreme

Close-up. Although there may be also different names, all artists will understand this scale. The same applies forboth camera angles and basic camera movement – or rather a structure the camera was on (jig, crane, rails, tripod, hand-held, helicopter, drone, ...)

Other features have to be stored as a vector, or other structure that is not human-readable but creates a possibility of indexing and searching. A very simple example may be a vector of visual concept presence in a shot. These concepts are usually abstract and thus not easily describable.

As a possible storage for all extracted data and platform for search, we will consider project NARRA[3]. This project is developed on Center for Audiovisual Studies, Film and TV School of Academy of Performing Arts in Prague as an Open Narrative platform, where artists are enabled to collaboratively create narratives by linking individual multimedia items (audio, video, image,text) together. Apart from manual annotation and linking, NARRA supports automatic meta-data and description generators. Directional or non-directional links between individual items in a collection can be then created with automated synthesizers.

4 Proposed Processing Flow

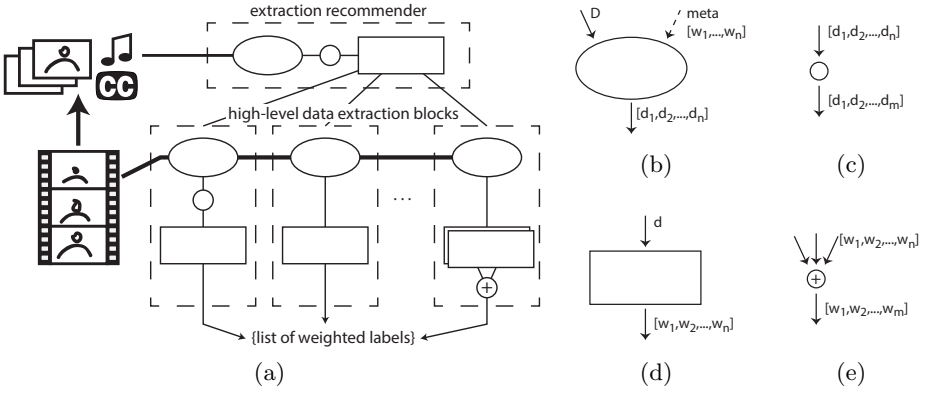

Fig. 1. Proposed data flow in the data processing (a) consists of an extraction recommender that proposes a set of high-level data extraction blocks. All developed functionality may be divided into four basic functionalities: Data extraction (b) takes input data D and possibly also a set of meta-information and outputs set of descriptors $[d_1, \ldots, d_n]$. Pre-processing (c) takes the descriptors and create a set of transformed descriptors $[d_1, \ldots, d_m]$. Descriptor or set of descriptors is then classified (d) and set of posterior probabilities or class weights $[w_1, \ldots, w_n]$ is returned. Result of multiple classifiers is joined by a late fusion (e), usually voting.

[3] `http://narra.eu`

As mentioned in the introduction, time and resource consumption is critical in most scenarios. Public media houses need to find illustrative material for current events as fast as possible, if not directly the recording of the event. In such hurry, multimedia is however usually poorly annotated by people and thus hard to discover.

Film-makers are commonly struggling to find pieces of their previous work that they know about, but forget the exact location. Or they have several versions of the footage, which may lead both to confusion which is the appropriate version, as well as to possible wasting of storage space.

In both cases, the media collections are large, and we need to gather as many relevant information as possible in reasonable time. We are therefore trying to deduce what the relevant information is, to eliminate wasteful extractions or training of classifiers.

4.1 Overall Structure

To achieve the best performance, we first extract the easiest descriptors from the multimedia. If there is a text information attached, we process it as soon as possible with keyword extraction and simple text-mining. Global audio descriptors are extracted to help distinguish sound and speech on a basic level. On video, multiple simple extractions are combined into a single pass. This is beneficial, as decoding of the video signal does take a fair amount of resources.

Although this first layer has to consist only from simple extraction methods, it may, however, provide also some information usable in the final indexing. Such as: if music or speech is present in the video, how many clips does the video consist of and where the cuts are, basic colour histogram, etc.

For the purpose of indexing, we assume each multimedia file as an item. Even if the multimedia have been mixed from multiple sources, we assume that the multimedia as a whole holds some meaning. If there are cuts detected in the video, each part is treated as a sub-clip for further analysis. This may introduce an information about an online-edited video from multiple cameras or generally enable linkage of similar sub-clips.

Based on the output from the first data extraction layer, we select a set of high-level extraction methods that will run in parallel to gather more detailed information about the multimedia. Such selection will be performed by a classifier, which is evolved thorough meta-learning.

4.2 Used Meta-features

The set of used meta-features in the first layer has to be large enough to be able to correctly predict methods used in further processing, however excessive number of features will slow down the process of such selection and defeat the purpose of multi-level classification.

Currently we are experimenting with following multimedia meta-features: average sound power, variance of the sound power, statistic properties of specific

loudness, number of detected video edits, statistic properties of edit length and colour histogram of each detected clip, spatially divided into four blocks (2×2).

This list is, however, not definite yet as the selection of high-level data extraction blocks is not complete either.

4.3 Meta-learning

The process of meta-learning is based on a feedback from the high-level extraction block, where the classifier proposes multiple of these blocks. After the full evaluation, each block returns its score back to the classification step, and if the score is higher than a threshold, we add another data point that can be used in further classification.

For simplicity, we are currently using a k-NN classifier that returns the k closest input data we have met so far (based on the meta-features) along with precision of the used blocks. k has to be at least a double of extraction blocks present in the system. Based on these information we select only the most successful extraction blocks and execute them.

With introduction of such loop in our meta-learning, we are trying to improve it over time and possibly also enable adaptation to new data and concept drifts.

5 High-level Data Extraction Block

These blocks have to be at least partially constructed with a preliminary notion of output and data it is able to process, because we are very much limited by the data extraction methods themselves, which already carry some semantics. We are also trying to gather some implementations of currently used high-level data extraction methods and use them "as they are" as our extraction blocks. However, these methods have to be usually re-set on each new sub-clip.

We are also experimenting with a genetic programming approach to select the appropriate extraction methods and classifiers to achieve extraction of certain multi-channel high-level features. For example, speech does not consist solely from a sound, but even humans tend to understand more if watching the face of the speaker. This way, one can easily distinguish between individual speakers as well. Therefore a combination of visual and auditory signal processing seems to be beneficial. However we will not consider such blocks in this paper.

With information from the first layer, each extraction block should be able to get access to all required information. If the specification of sub-clips is included, extraction block can also limit its function only to certain parts of the multimedia.

5.1 Extraction

This is the section we are currently working on the most. We are testing the media descriptors mentioned in the Section 2, in respect to the possible subdivision of the multimedia proposed by the first layer.

Also, some descriptors yield their results as a big set of values dependent on time. In this case, custom further processing is required.

5.2 Pre-processing

In case of descriptors of a lower-level, we are usually faced with a lot of high-dimensional data. As classifiers are generally very bad in coping with such data (due to "curse of dimensionality"), pre-processing methods, such as singular value decomposition or principal component analysis, can be used to reduce the dimension of original data. These pre-processing methods are usually costly and output dimensions are abstract, but smaller number of concepts is better-suited for classification tasks.

Other descriptors may create a sequences of data, which is also hard to be processed by a standard classifier. In such cases we may use either statistics of the data (minimum, maximum, first four empirical moments, ...) or some other transformation. We also consider a use of other classification algorithms that are designed to work with time sequences.

Most importantly, as there is usually a classifier hidden inside our block, meta-features may be extracted to help in selection of the classifier and/or its parameters. This will be discussed in next subsection.

5.3 Classification

Some of the data extraction algorithms are accompanied with preferred classifiers, as discussed in their own research papers. Music is for example commonly clustered with self-organising maps, whereas image features use classifiers based on nearest neighbours. There are also multiple approaches inherently using the deep convolution neural networks.

As we would like to use some of the low-level data extractors as well, we need to come up with some custom classifiers. For such cases, most suitable classification algorithms and their parameters need to be found.

This will possibly create a bottleneck and here the meta-learning principle may be used again. In this case, the individual classifiers will be learned beforehand outside of the system. Inside our high-level blocks, only the acquired meta-knowledge will be used to help selecting the most appropriate classifiers.

Actual creation of classification models will proceed for each dataset or collection in NARRA separately for better conformation to requirements of each segment of the data.

5.4 Post-processing

As the classifier may return multiple classes, or multiple classifiers will run in parallel, further processing may be required as well. Such processing may include selection of most possible classes, voting of multiple classifiers or text description of the output class where applicable.

6 Summary

We have proposed a use of meta-learning principles for multimedia processing and classification to induce faster indexing of multimedia content. The main benefit of our approach is that we are not running all possible extraction methods, but the first classification layer selects only the most relevant to be performed. In case of custom classifiers, possibly combining results from multiple extraction methods, meta-learning is also used to reduce time needed for selection of appropriate classifier and their parameters.

All of the presented work is currently under development and preliminary results are to be expected during 2Q2016.

Main part of the research will be hopefully conducted during the next two years of my doctoral studies.

References

1. Information technology – multimedia content description interface – part 4: Audio. Tech. Rep. ISO/IEC 15938-4:2002 (2002)
2. Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R.: Metalearning. Cognitive Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg (2009), http://link.springer.com/10.1007/978-3-540-73263-1
3. Cornelson, M., Greengrass, E., Grossman, R.L., Karidi, R., Shnidman, D.: Survey of Text Mining: Clustering, Classification, and Retrieval. Springer New York, New York, NY (2004), http://dx.doi.org/10.1007/978-1-4757-4305-0_7
4. Duygulu, P., Barnard, K., Freitas, J.F.G., Forsyth, D.A.: Computer Vision — ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV, chap. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, pp. 97–112. Springer Berlin Heidelberg, Berlin, Heidelberg (2002), http://dx.doi.org/10.1007/3-540-47979-1_7
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 580–587. IEEE (2014)
6. Javed, O., Shah, M.: Computer Vision — ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV, chap. Tracking and Object Classification for Automated Surveillance, pp. 343–357. Springer Berlin Heidelberg, Berlin, Heidelberg (2002), http://dx.doi.org/10.1007/3-540-47979-1_23
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
8. Lee, C.H., Soong, F.K., Paliwal, K.: Automatic speech and speaker recognition: advanced topics, vol. 355. Springer Science & Business Media (2012)
9. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 282–289. ACM (2003)

10. Lidy, T., Rauber, A.: Classification and clustering of music for novel music access applications. In: Lecture Notes in Applied and Computational Mechanics. pp. 249–285 (2008)
11. Liu, F., Gleicher, M., Jin, H., Agarwala, A.: Content-preserving warps for 3d video stabilization. In: ACM Transactions on Graphics (TOG). vol. 28, p. 44. ACM (2009)
12. Neumann, L., Matas, J.: Computer Vision – ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part III, chap. A Method for Text Localization and Recognition in Real-World Images, pp. 770–783. Springer Berlin Heidelberg, Berlin, Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-19318-7_60
13. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Computer Vision–ECCV 2006, pp. 490–503. Springer (2006)
14. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. Signal Processing Magazine, IEEE 20(3), 21–36 (2003)
15. Peeters, G., McAdams, S., Herrera, P.: Instrument Sound Description in the Context of MPEG-7. In: ICMC: International Computer Music Conference. pp. 166–169. Berlin, Germany (Sep 2000), https://hal.archives-ouvertes.fr/hal-01161319, cote interne IRCAM: Peeters00a
16. Ribeiro, P.C., Santos-Victor, J.: Human activity recognition from video: modeling, feature selection and classification architecture. In: Proceedings of International Workshop on Human Activity Recognition and Modelling. pp. 61–78. Citeseer (2005)
17. Robertson, N., Reid, I.: A general method for human activity recognition in video. Computer Vision and Image Understanding 104(2), 232–248 (2006)
18. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. vol. 2, pp. 1331–1334. IEEE (1997)
19. Selvan, S., Ramakrishnan, S.: Svd-based modeling for image texture classification using wavelet transformation. Image Processing, IEEE Transactions on 16(11), 2688–2696 (2007)
20. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on. pp. 586–591 (Jun 1991)
21. Wang, A.: The shazam music recognition service. Communications of the ACM 49(8), 44–48 (2006)