Towards AI-based Semantic Multimedia Indexing and Retrieval for Social Media on Smartphones

Stefan Wagenpfeil
University of Hagen
Faculty of Mathematics and Computer Science
58097 Hagen, Germany
stefan.wagenpfeil@studium.fernuni-hagen.de

Prof. Dr. Ing. Matthias Hemmje
University of Hagen
Faculty of Mathematics and Computer Science
58097 Hagen, Germany
matthias.hemmje@fernuni-hagen.de

***Abstract*—To cope with the vastly growing number of Multimedia Assets on Smartphones and Social Media, an integrated approach for Semantic Indexing and Retrieval is required. This paper introduces an approach towards a generic framework to fuse existing image and video analysis tools and algorithms into a Unified Semantic Annotation, Indexing and Retrieval Model resulting in a Multimedia Feature Vector Graph representing various levels of Media Content, Media Structures and Media Features. Utilizing Artificial Intelligence and Machine Learning, these Feature Representations can be used to provide accurate Semantic Indexing and Retrieval. This paper provides an overview of the Generic Multimedia Analysis Framework (GMAF) as well as the definition of a Multimedia Feature Vector Graph Framework (MMFVGF). As a third contribution we introduce AI4MMACCESS to detect differences, enhance Semantics and refine weights in the Feature Vector Graph. Combining this, we describe a solution for highly flexible Semantic Indexing and Retrieval that offers unseen possibilities for applications like Social Media or local Apps on Smartphones.**

***Index Terms*—semantic indexing; multimedia retrieval; framework; multimedia feature vector graph; MMFVGF, MMFVGG; AI4MMACCESS; GMAF;**

I. Introduction and Motivation

Every year more than 1.2 trillion digital photos and videos are taken, more than 85% of them on Smartphones and this number is still increasing [1]. This amount of media is neither manageable for the users nor for content providers and / or platforms like Social Media. The easiness to take pictures wherever and whenever you want is unseen in history of media creation. Additionally, cloud services and storage become cheaper and cheaper, which makes it very simple and affordable for users to store 100.000s of media assets on their Smartphone [2]. Smartphone vendors provide solutions allowing to access all the users' media assets directly through the device by transparently up- and downloading them to cloud-services [3]. The massive amount of Multimedia assets provides huge challenges for indexing, querying and retrieval algorithms. Although there is great progress in image analysis, object detection, and content analysis [4], there is still potential for improvements and further research to optimize the results of information retrieval algorithms.

978-1-7281-5919-5/20/$31.00 ©2020 European Union

Artificial Intelligence (AI) and Machine Learning (ML) including especially Deep Learning and Pattern Recognition have made a huge contribution to research and development during the last years [4][5][6][7]. With the help of these algorithms and methods, products and services became available, that provide a huge set of functionalities to detect content of images, videos, text or audio. One major example is Google's Vision AI – a service to access a fully trained and equipped neural network for object recognition and basic semantic feature detection in images [8]. Similar APIs are provided by Microsoft [9] or Amazon [10]. All these services provide basic object detection, text and pattern recognition and a basic semantic feature detection.

Today's digital cameras are able to automatically provide a huge set of additional information for each picture. Starting with Date, Time, GEO-Location, Aperture, Exposure-Time, Lens-Type, Focal-Length, etc. many of this kind of information can be used to enrich images and their metadata representation. The EXIF-standard is implemented in almost all camera models from Smartphone to Professional [11]. For videos, MPEG7 provides metadata about videos, shots, video content in a standardized and industry accepted way [12]. Semantic knowledge representation also evolved a lot in providing standards, algorithms, data structures, and frameworks [5] as well as products or projects like Google's Knowledge Graph [13] or the Semantic Web [14].

Summarizing this, we are in the situation, that an enormous set of technologies, standards and tools is available that provides a powerful opportunity to build new structures and algorithms on. This paper provides an outline and approach towards a generic framework to integrate various syntaxes, semantics, and metadata of Multimedia Content into a single indexing model, which can be used for direct retrieval or as API for other applications. To address the various levels of Multimedia Content (technical level, content level, semantic level, and intra-content level) [15][16], we introduce a *Multimedia Feature Vector Graph Framework (MMFVGF)*, which is able to integrate the various metadata features of Multimedia Assets into a unified semantic data structure. And we describe a ML-based indexing and retrieval component to refine the MMFVGF and access the resulting Multimedia Assets.

II. State of the Art and Related Work

This section provides an initial selection of the most relevant state of the art in science and technology and work related to this paper covering Multimedia Processing, Semantic Analysis, Multimedia Querying, and Retrieval, AI pattern matching and Social Media. Our *Generic Multimedia Analysis Framework (GMAF)* utilizes and integrates several of the algorithms and technologies presented in this section as building blocks. Thus, this section indirectly introduces the functionality of the GMAF as well.

A. Multimedia Processing

In Multimedia Processing, lots of tools, algorithms and APIs are available. The most common products are already defined in the previous section [8][9][10][11][13][17] and are based on years of research, development, and training. Deep Learning methods utilize neural networks to provide better accuracy [7], in addition massive sets of annotated training data are available [18]. Additionally, numerous standards have been defined in the last years including image annotations (like EXIF [11]) and video metadata (like MPEG7 [12]). There are still several vendor-specific metadata models, like Adobe's XMP standard [19] or the production industry standard MXF [20] and all of these can easily be integrated into GMAF by just writing a simple plugin and attach it to the framework (see Figure 5). Actually, there is very good progress in detecting objects, relevant regions, the color-footprint or even activities or moods of an image or a region, all the technical metadata are embedded in the image or video itself [11][12] and digital assistants like Siri, Cortana, or Alexa already provide the basic technology to use natural language for query input. Object and region detection provide good results [8], and also scene-detection and automatic extraction of a video's key frames is not a major challenge anymore [17]. Also, there are lots of standards to describe general metadata for video [12] and if all of this is applied recursively to a video's frames, any image processing algorithm is applicable for video as well (depending on the computation time).

B. Semantic Analysis

Vocabularies and Ontologies are the basis for semantic analysis, which are usually defined in RDF Schema. A good overview of the current state-of-the art is presented by [21]. Our work is based on these existing standards and will utilize them and contribute to them as much as possible. Spaciotemporal annotations in movies (for example) can be used to feed Composition Relationships (cr) into the MMFVG but in return, detected cr information can be written into the asset's corresponding MPEG7 sections. To explore the associations between terms and image regions as the visual context information, several models have been defined and developed. Examples are the co-occurrence model, the translation model, the continuous relevance model, or the multiple Bernoulli relevance model [22]. In addition, with Google's Knowledge Graph project [13], a huge collection of semantic relationships, synonyms, and metadata is available for public use. For training, classification, or automation purposes, several libraries of annotated media are available [18][23], which are widely used in development to train AI components in Multimedia Detection and Analysis topics. These datasets can also be applied to the concepts described in this paper.

C. Multimedia Querying and Retrieval

According to [22], three levels of visual information retrieval may be distinguished: Level 1 contains low-level information like color, shape, location of image elements, etc. Level 2 provides attributes or semantic content like the retrieval of objects of a given type or class as well as the retrieval of individual objects. Level 3 gives abstract attributes and includes search requests for named events or activities. As part of the concept-based image retrieval process, our approach utilizes existing segmentation algorithms to refine the image's or video's semantic information and utilize existing semantic representations of Level 1 to 3 to refine detected information. In general, this work will contribute to closing the semantic gap: *"the semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation"* [24].

D. AI Pattern Matching

During the last years, AI has made huge leaps forward. Especially Machine Learning has evolved to become largely relevant for research as well as industrial applications. The most relevant technology for our approach is Deep Learning, which is defined as a part of Machine Learning and consists of algorithms used to model high-level abstractions in data by using architectures composed of multiple nonlinear transformations [21][6]. Especially Convolutional Networks (CNN) are used as feed-forward mechanisms to increase the accuracy of feature detection. Actually, by using the GMAF, this technology is indirectly integrated into the presented solution, but the current CNNs are trained to detect the most relevant part of an image and therefore fail in some cases that out approach will be able to solve.

E. Social Media

In Social Media, a Multimedia Asset is much more than "just" a video or an image. In addition to the raw asset, metadata like posts, text, description, corresponding posts, comments, likes, or dislikes, and the users' personal settings can contribute to the relevance of a Multimedia Asset. Currently it is hard to combine all this information into a single model to semantically index all the corresponding parts. Therefore, in Social Media, usually independent searches based on keywords are applied [25].

F. Related Work

Multimedia Processing has been a relevant research topic for many years and lots of contributions to this topic have been made. [26] for example, provides a *"Standard Multimedia Processing Pipeline"* to which all the contributing topics can

be aligned. Therefore, we will align the related work on this model to illustrate the most relevant contributions.

So, the *"General Data Processing Pipeline"* can be mapped to this "Standard Multimedia Processing Pipeline" [26] providing a Multimedia specific representation of the eight typical processing steps of Multimedia Content (see Figure 1). According to this pipeline model, Multimedia Processing is separated into Creation, Composition and Delivery.

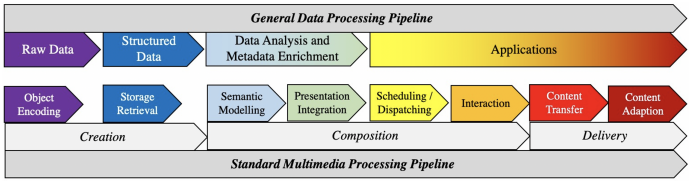

Fig. 1. Mapping of the General Data Processing Pipeline to the Standard Multimedia Processing Pipeline

"Multimedia Creation" is the process of object encoding and storage retrieval. In the context of this paper, the raw data on users' Smartphones (image, video, or text) and the corresponding structured metadata are used which are provided by the smartphone itself or by Apps used by users to annotate these assets. The assumption is, that metadata is directly attached to the asset like [11] or [12]. The Multimedia Creation process step also includes assets, that are received by other channels like Social Media or Messaging. In this case, the set of metadata can vary, but is still existing.

In the *"Multimedia Composition"* phase, lots of algorithms for the lower-level of feature extraction, like SIFT or Haralick have been defined, implemented and applied to Multimedia Content [21][27][28][15]. Segmentation algorithms like MSER or SIMSER [21][29][30][31] help to identify the correct bounding boxes of objects within images and several ontology-based models for semantic annotation have been proposed by [21][32][33][34][35]. Also, recent approaches regarding feature-fusion in special topic areas are contributing to our approach [36][37]. [38] proposes Confidence-Based Ensemble to provide Semantics and Feature Discovery to dynamically extend the traditionally static semantic classifiers and provides a mathematical model to resolve semantic conflicts and to discover new semantics.

The *"Multimedia Delivery"* phase as well as the *"Scheduling"* and *"Interaction"* phase are not relevant for this document.

G. Discussion and Summary

As stated in the previous paragraphs, there are lots of concepts, algorithms, and tools, that provide solutions for topics in the area of Multimedia Processing. But even by utilizing all these technologies, algorithms, and research results, it is currently not easy or even possible to retrieve Multimedia Content for queries like *"show me a picture where I got my new watch"* or *"show me the video where my daughter had her first concert"*. One reason for this is, that usually most of the described algorithms run independently from each other and do not contribute to a common metadata model. A second reason is, that there is lots of unused potential by focusing only on a single asset and not taking into account that related or similar assets can contribute to the metadata of the asset under investigation as well.

Our work will contribute to closing this gap by providing an approach towards a generic framework to integrate the current state-of-the-art algorithms (GMAF), a data structure to fuse existing features into a single unified model (MMFVGF), and a ML-component to refine the feature model and to provide semantic indexing, querying, and retrieval (AI4MMACCESS).

III. Conceptual Design

Our approach is aligned to the *"User Centered System Design"* defined by [39]. To formalize the model in a visual language, we will use the UML [40] and present an initial Use Case design for the most relevant activities.

In our context, two major actors can be identified, which are interacting with a Multimedia Asset Management System (see Figure 2). The Producer is creating or modifying Multimedia Content by taking pictures, videos, posting to Social Media, or by manually creating annotations. The Consumer will then search, filter, and retrieve the asset. Producer and Consumer can be the same person in real-world-scenarios. All Use Cases are part of the *"Data Analysis and Metadata Enrichment"* phase of the General Data processing Pipeline (see Figure 1).

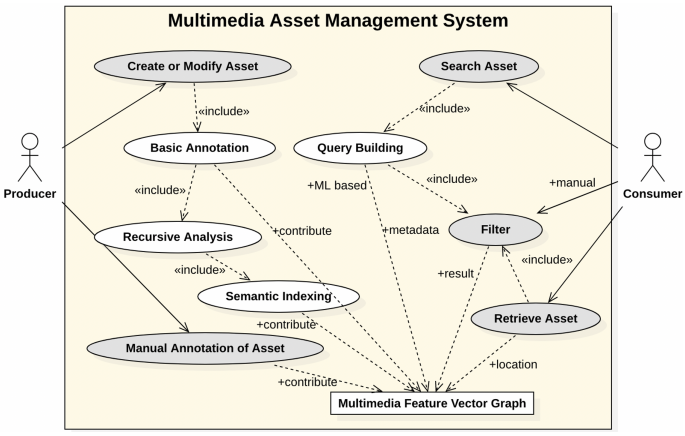

Fig. 2. Initial Use Case Design for the extensions of a Multimedia Asset Management System

The Use Cases *"Create or Modify Asset"*, *"Manual Annotation or Asset"*, *"Search Asset"*, *"Filter"* and *"Retrieve Asset"* are already in place on almost every Smartphone or Multimedia Application. Our approach will extend these basic Use Cases by providing additional functionalities. From the users' point of view, nothing will change, and existing applications can be used in the same way as before. We will enhance the search accuracy of these applications

by providing additional components, that will increase the semantic quality of Multimedia retrieval (see Figure 2).

"Basic Annotation" is already in place on many Smartphones, and can provide metadata like location, date and time, EXIF-data, and in some cases even built-in face-detection [41]. This Use Case is extended to feed its data into the Multimedia Feature Vector Graph (MMFVG) as part of the MMFVGF, where it will be enriched and fused with other metadata, which are generated in the subsequent processes. The *"Recursive Analysis"* applies a chain of filters to every detected object or region in order to refine the detection results. *"Semantic Indexing"* fuses all the different metadata into a unified model represented by the MMFVG. To take full advantage of the MMFVG, the *"Search"* Use Case has to be extended as well by introducing a *"Query Building"* component, that utilizes the MMFVGF.

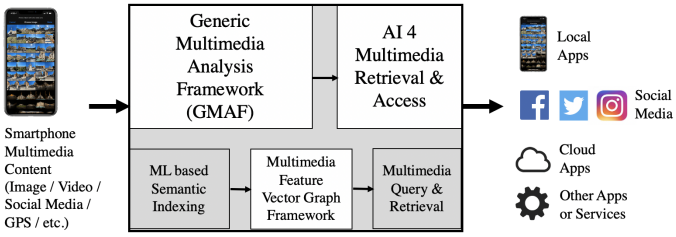

Fig. 3. Multimedia Asset Management System Architecture including the GMAF, MMFVG and AI4MMACCESS components

From a conceptual architecture perspective, three frameworks are required to fulfill the requirements (see Figure 3):

Generic Multimedia Analysis Framework (GMAF): a framework capable of combining and utilizing existing Multimedia processing systems for images, videos and text-information (like Social Media Posts). It can be regarded as a chain of semantic metadata providers, which will produce all the data that is required to construct a Semantic Index. The GMAF represents the Use Cases *"Basic Annotation"*, *"Recursive Analysis"* and *"Semantic Indexing"*.

Multimedia Feature Vector Graph Framework (MMFVGF): a framework that fuses all the semantic metadata into a large semantic Feature Vector Graph and assigns it to each media asset. Weights, nodes, and references are structured in a way, that the relevance of features within a Multimedia Asset can be determined very accurate.

AI for Multimedia Retrieval and Access (AI4MMACCESS): a ML-component to process and permanently adjust the Feature Vector Graphs of each image or video in order to refine the relevance-information. AI4MMACCESS provides a permanent re-processing in terms of the Use Case "Semantic Indexing" as well as a "Query Builder" component.

A central shared resource of this approach is the *Multimedia Feature Vector Graph (MMFVG)*, which fuses existing ML based Semantic Indexing methods into a single representation for Multimedia Query and Retrieval. The MMFVG is the result of the Generic Multimedia Analysis Framework (GMAF) and will be used within the AI for Multimedia Retrieval and Access (AI4MMACCESS).

To illustrate the expressive potentials of the MMFVG, let us look at some initial query types supported by this data structure, which could represent queries, where current algorithms are likely to fail and the reasons for this:

Q1: *"show me the picture from Florida, where Jane wore flipflops"* – some of the algorithms will be able to actually find pictures of Jane, and maybe also pictures where a person is wearing flipflops – but most of the algorithms will fail combining this correctly. One would also receive pictures of Jane, where another person next to her is wearing flipflops.

Q2: *"show me a picture where I got my new watch"* – this kind of question is problematic for most of the current algorithms. One reason is, that information from more than one picture is required to solve this query. A comparison of "old" pictures with the ones, where a "new" object appears, is required. The second reason is, that all the object detection algorithms are trained to find relevant objects, which a watch usually is not. Even if we would use the finest granularity of object detection, the "watch" won't get any relevance for indexing. Even more difficult to answer is the question, what happens if my "new" watch becomes "old" again? (Note: this is one reason why AI4MMACCESS does a permanent re-processing of Feature Vectors Graphs).

Q3: *"can you find the picture where I had my leg broken"* – is also difficult to solve with current technologies. How should you detect a broken leg? The semantics that broken legs might be detected by a white plaster cast in combination with the region-detection "leg" is hardly possible. Maybe I broke my arm at another time? How should current algorithms distinguish? Even if you would have a Twitter-Post with that picture, none of the current algorithms would be able to combine this properly.

The combination of GMAF and MMFVGF provides an approach towards a solution, how this kind of queries could be processed with accurate results. The GMAF produces and analyzes Feature Vectors, the MMFVGF integrates and provides the MMFVG data structure for querying.

The MMFVG has not been designed to be human-readable or -viewable, but to provide best AI-processing support. The basic concept is closely related to graph-theory, so you can provide visualizations of parts of the MMFVG, but in general a MMFVG for a given Multimedia Asset will be too large

to be visualized in a senseful way. To describe this data structure, the following terms are introduced:

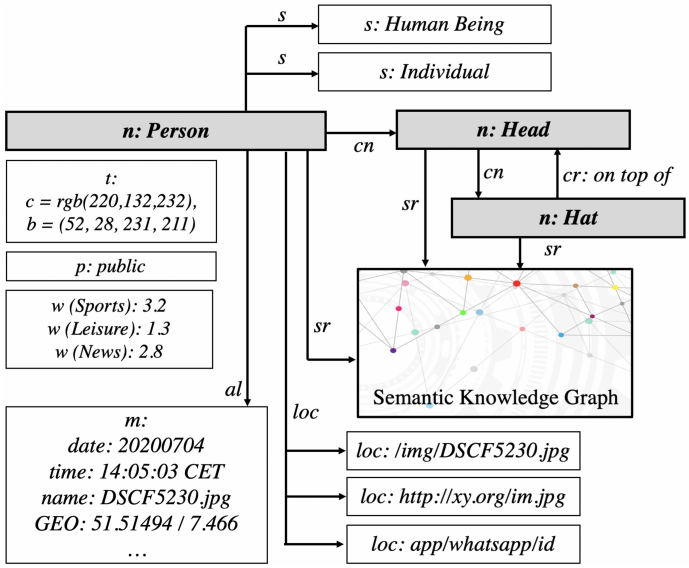

Fig. 4. Illustration of the Multimedia Feature Vector Graph (MMFVG) and its corresponding attributes. For illustration purposes only some of the relationships are shown in that figure

- Node *n*: nodes represent a single object, activity or region of relevance that has been detected in an asset.

- Weight *w*: each node has several assigned weights. The weight will represent the relevance of a node according to a special context. The context can be referred as surrounding metadata, which refine the later search- and query-scope. Best example is the timeline of a movie, where the very same object can have different contexts during the story of the movie. A weight results of the importance of the node in the overall image and the deviation from ”normal” image content compared to other images of a similar kind (remember the ”new watch” example). It is calculated initially by the GMAF, but constantly refined by AI4MMACCESS.

- Childnodes *cn*: these are sub-objects, that have been detected by recursive application of the GMAF. Example Person → Body → Arm → Left Hand → Fourth Finger → Ring. So, one of the Person’s child nodes would be the ”Body”, one of the body’s child nodes would be ”Arm”, and so on. Child nodes are produced by applying the bounding boxes of detected objects to the GMAF recursively.

- Technical Attributes *t*: each node can have numerous technical attributes. These can be the Color-Footprint of this node (c), its bounding box (b) within the image (defined by x, y, width, height), the DPI-resolution of the section, information about sharpness or blurring, etc.

- General Metadata *m*: this is a reference to the Asset’s general metadata object, which is extracted by EXIF or MPEG7 [11][12] and contains metadata like date, time, GEO-coding, Aperture, Exposure, Lens, Focal Length, Height, Width, Resolution, etc. In m also a unique identifier for the Multimedia Asset is stored.

- Synonym information *s*: each node usually defines a human understandable object, like ”Person”. To support flexible querying, the object’s description has to be normalized and aligned to a synonym graph, where the ”is a” relationship is modelled. So, for a ”Person”, we would find ”individual”, ”being”, ”human”, ”creature”, ”human being”, etc. as well.

- Composition Relationship *cr*: this attribute is used to provide a relationship between a Multimedia Asset’s objects. It contains information like ”next to”, ”in front of”, ”behind”, ”attached to”, etc. and is calculated by recursively applying the bounding boxes of the objects and measuring the distances between these objects.

- Semantic Relationship *sr*: each object is assigned to a position within the overall semantic knowledge graph, for example in [13].

- Asset Links *al*: if this image is part of a series or refines other images that have already been processed, it is linked to their MMFVG as well. Asset Links can also point to text information (like posts, comments, likes or dislikes in Social Media) and therefore provide important meta-information that is not included in the asset itself.

- Locations *loc*: an asset can be placed on multiple locations (local Smartphone, Cloud-Services, Messaging Apps, etc.). To show the asset to the users, we need to know, which locations are available, at what resolution the image is stored there and then determine the best option to retrieve the asset.

- Privacy and Security *p*: each asset has to be assigned to Privacy and Security settings, like to an Access Control List. This is important to ensure that the users will only process and receive their own content and content that has been shared to them by other users.

The MMFVG itself does not provide any functionality, it is a data structure for general use to recursively describe Multimedia Assets and their semantic relationships and to provide a generic structure where all the standards of existing specifications can be mapped to. A detailed data model and a formal representation of the MMFVG is part of our ongoing work.

It has to be investigated, at what point of time in the overall process the MMFVG has to be constructed and when the permanent re-processing of the AI4MMACCESS is performed.

As indexing structures might become quite huge and complex, when any possible query-context should be calculated, it might be better to process the relevant assets just at query-time. This is subject of further analysis in regard to performance, runtime, data volume and will refine the overall architectural topology.

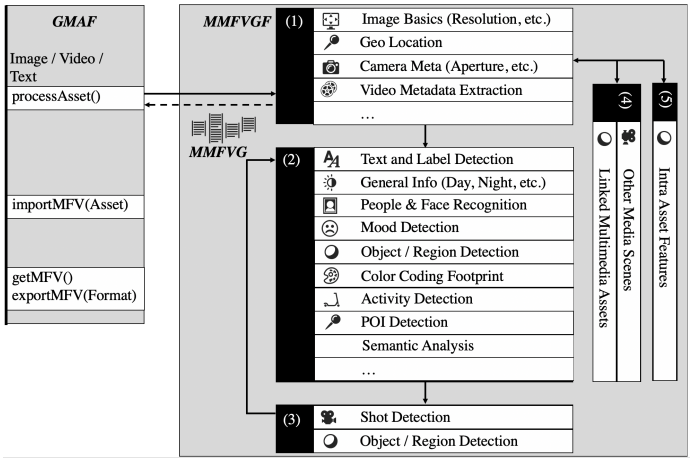

Fig. 5. Conceptual processing approach and basic structure of GMAF and MMFVGF

Figure 5 illustrates this conceptual approach and how the Multimedia Feature Vector Graph is constructed within GMAF and MMFVGF. The MMFVGF can deal with multiple dimensions of feature-extraction, in this way significantly extending and refining the standard Multimedia layers as defined in the so-called Strata Model [16], in which each media segment has its own layer distinguished by horizontal layers (mostly time-based) and vertical layers (synchronized media signals). Based on these layers, additional feature representations are introduced as dimensions within the MMFVG:

(1) Features within a single Multimedia Asset, like object detection, technical metadata, basic semantic metadata

(2) Features on higher semantic levels, like people or face recognition, mood detection, ontology links

(3) Recursive Features, which result from applying the GMAF recursively to subsets of the Multimedia Asset

(4) Indirect Features, which result from linked assets and refine the features of the current asset (other scenes in a video, corresponding Social Media posts)

(5) Intra-Asset-Features, which use information from Assets of the same series, location, or timestamp or information from generally similar Assets to refine an Asset's Features.

These feature levels can be classified into *"horizontal levels"*, which mostly contain non-temporal indirect features or Intra-Asset-Features (besides the temporal dimension), *"vertical levels"* containing all features within a single Asset including higher semantic levels and recursive features. GMAF and MMFVGF will produce a unified Multimedia Feature Vector Graph, which represents the fused information of all of these levels.

While this section introduced and presented our model of layers and dimensions as well as the corresponding basis technologies, the following section will illustrate our prototypical implementation of selected components.

IV. Initial Prototypical Implementation

In order to illustrate and partially proof the conceptual approach described in the previous section, we provided several selected implementations for the most relevant parts of GMAF and MMFVGF to show, that our approach is valid and further investigation is promising.

While the GMAF is based on standard technologies (Java, SOAP, REST, XML, and HTTP) and a smart way to combine and utilize them, the MMFVGF and the AI4MMACCESS cannot be achieved with standard technologies and have to be derived from [11][14][12][20]. These starting points are introduced in this section and their contribution to research and technology is illustrated. The prototype of the GMAF including the MMFVGF is available at [42] including further information of the implementation details.

A basic semantic analysis is performed by GMAF's plugins. In our prototype we used [8][11] as well as [43] to implement the plugins for GMAF shown also in Figure 4. Several different semantic models have to be combined, fused and mapped during the process of creating the MMFVG.

A. Generic Multimedia Annotation Framework (GMAF)

The GMAF is implemented as a framework that utilizes standard image processing technologies like object detection, basic semantic annotations and standard image metadata as described in the previous chapters of this document by applying [8][11][44]. New algorithms can be attached as plugins quite easily and a recursive feature enrichment is also part of the GMAF and the basis for the MMFVG (Figure 4). The usage of the GMAF is quite straightforward, as it already provides ready-to-use adapters for the major Multimedia Processing tools. In general, the GMAF will process each uploaded Multimedia Asset according to a processing chain, where any available plugin is applied to enrich the MMFVG. The GMAF is available as a Standard JEE Webapplication and can be deployed to any JEE compatible server or container [44][45]. Figure 6 shows a sample usage of the GMAF with an uploaded image, which is then internally processed into a MMFVG and exported for visualization purposes to GraphML (see Figure 7). Current results of GMAF-processing show, that the level of detail increases a lot due to the recursive application of algorithms. It is currently subject of investigation, what level

of detail will provide the best MMFVG-results for further processing.

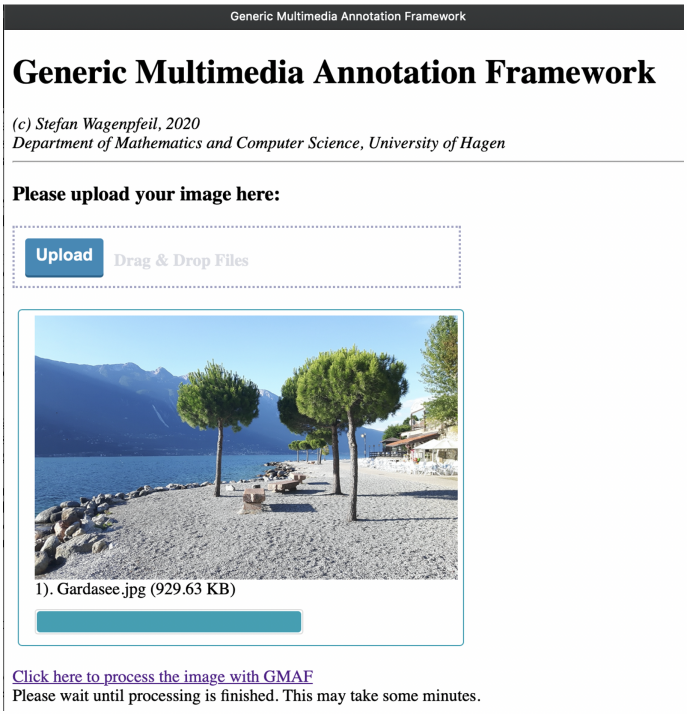

Fig. 6. GMAF Web Application with uploaded image of a seaview.

B. Multidimensional Feature Vector Framework (MMFVGF)

The MMFVGF calculates the Multidimensional Feature Vector (MMFVG) and provides structures to organize and maintain this data structure. MMFVGF is also implemented in Java and includes importers and exporters to various formats (like JSON, GraphML, XML, etc.) and APIs to use the MMFVGF and access the MMFVG via SOAP and REST. The implementation of the various objects of the MMFVG is based on current Multimedia Standards like [11][13][12][17][43][46].

C. AI4MMACCESS

To gain the ability to process queries like Q1, Q2 or Q3, AI4MMACCESS is introduced as a specialized AI to index and retrieve Multimedia Content. It is based on the IVIS reference model for Advanced Interfaces [47] and adapts this model to be compatible with the EDISON project [48].

The most important task for AI4MMACCESS is to detect deviations of an asset compared to typical other assets of this type. Whereas the GMAF calculates weights according to standard rules, AI4MMACCESS is trained to detect even minor changes of one asset like the ”new watch” of example Q2 – noticing, that usually the ”watch” itself would not be of any significant relevance for the picture’s description. For this deviation, a semantic comparison has to be implemented, which can include texts from Social Media (”look, I have a new watch”), other images (with the ”old watch”), location, date and time information (”me at the watch store”) and also a comparison of the MMFVG from different Multimedia Assets. As the semantic relevance directly correlates to the query-context, the relevance-information has to be calculated for each context.

AI4MMACCESS utilizes a Machine Learning compatible representation of the MMFVG to detect the relevant parts of the MMFVG according to the users’ or query’s context, the deviation from other Multimedia Content and supporting information like Social Media. These concepts can then be applied to applications on Smartphones, which still leaves further investigations on the capabilities to support AI4MMACCESS as native Smartphone application in respect to memory, computing power and security topics. Then, a great user experience can be provided. Adaptations of the AI4MMACCESS-API for Social Media or other services will be quite straight-forward.

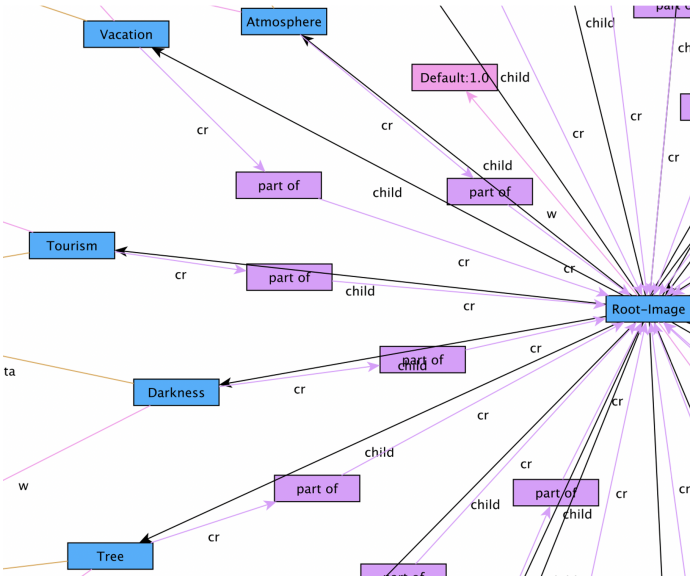

Fig. 7. A small detail of the seaview image after constructing its MMFVG and exporting it as GraphML, visualized with yED [46]

D. Querying

As the MMFVG has been designed to be represented in RDF [21] and therefore is compatible to all corresponding concepts [21], a query to the MMFVGF can be written in SPARQL [49] as a formal query language. For Q2 the corresponding SPARQL-query would look like this:

```
SELECT ?x ?y ?z
WHERE {
        ?x rdfs:subClassOf:watch .
        ?x mmfvg:attribute:new .
        ?y rdfs:subClassOf:Person .
        ?y mmfvg:name:Jim .
        ?x mmfvg:partof:?y .
        ?z mmfvg:type:Image
}
```

By enhancing the current functionality of [49] with MMFVG attributes, nodes and links, the result accuracy and functionality of current solutions can be increased a lot. Querying in the MMFVGF follows the *Query by Example (QBE)* paradigm [50], which will construct a $MMFVG_{Search}$ data structure representing the users' query. This query will be processed within the MMFVGF to find matching results.

E. Summary

Our prototypical implementation shows, that the basic concepts are a valid approach towards AI-based Semantic Multimedia Indexing and Retrieval for Social Media. Currently, our implementation analyzes *"vertical levels"* with some selected plugins and calculates a MMFVG. The *"horizontal levels"* including cross referencing and the AI component to detect context-based deviations are subject of further investigation and implementation and one goal of our work. Optimisation and validation of our frameworks is currently ongoing, as well as the definition of an approach towards the mapping of existing Multimedia services covering data redundancy, information uncertainty and the building of new semantic relations during the aggregation process of MMFVGs. After extending the prototype by further analysis plugins, we will start to train the AI4MMACCESS based on the calculated MMFVGs to detect context-based deviations within the MMFVGs. Finally, the prototype implementation should be adopted and optimized for Smartphones.

V. Relevance

Filling the semantic gap has been a research topic for years. Many contributions have helped to gain better basic analysis of Multimedia Content. Although, current technologies still cannot cope with the vast amount of generated content. The frameworks, concepts, architectures, and implementations described in this paper are designed to narrow down this gap and to provide an extensible solution for AI-based Indexing and Retrieval of Multimedia Content with a special focus on Social Media on Smartphones.

The GMAF Framework as well as the MMFVGF and AI4MMACCESS contribute to research and technology and provide basic components and solutions for other applications. Applications based on this technology will increase the quality of Multimedia Querying and Retrieval for the user and / or the provider of services like Social Media.

References

[1] M. Nudelman, "Smartphones cause a photography boom", Statista / Business Insider, http://www.businessinsider.com/12-trillion-photos-to-be-taken-in-2017-thanks-to-smartphones-chart-2017-8, Tech. Rep., Sep. 2020.

[2] O. Mazhelis, G. Fazekas, and P. Tyrväinen, "Impact of storage acquisition intervals on the cost-efficiency of the private vs. public storage", in *2012 IEEE Fifth International Conference on Cloud Computing*, 2012, pp. 646–653.

[3] Apple.com, "Icloud – the best place for photos, files and more", Apple.com, http://www.apple.com/icloud/, Tech. Rep., Jun. 2020.

[4] J. Beyerer, M. Richter, and M. Nagel, *Pattern Recognition - Introduction, Features, Classifiers and Principles*. Berlin: Walter de Gruyter GmbH & Co KG, 2017, ISBN: 978-3-110-53794-9.

[5] C. Beierle and G. Kern-Isberner, *Methoden wissensbasierter Systeme - Grundlagen, Algorithmen, Anwendungen*. Berlin Heidelberg New York: Springer-Verlag, 2019, ISBN: 978-3-658-27084-1.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016, ISBN: 978-0-262-03561-3.

[7] J. Heaton, *Deep Learning and Neural Networks*. Heaton Research Inc., 2015.

[8] Google.com, "Google vision ai – derive insights from images", Google.com, http://cloud.google.com/vision, Tech. Rep., Jul. 2020.

[9] Microsoft.com, "Machine visioning", Microsoft.com, http://azure.microsoft.com/services/cognitive-services/computer-vision, Tech. Rep., Jul. 2020.

[10] Amazon.com, "Amazon recognition", Amazon.com, http://aws.amazon.com/recognition, Tech. Rep., Jul. 2020.

[11] M. M. I. of Technology, "Description of exif file format", MIT - Massachutsetts Institute of Technology, http://media.mit.edu/pia/Research/deepview/exif.html, Tech. Rep., Jul. 2020.

[12] Shih-Fu Chang, T. Sikora, and A. Purl, "Overview of the mpeg-7 standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, 2001.

[13] Google.com, "Google knowledge search api", Google.com, http://developers.google.com/knowledge-graph, Tech. Rep., Jul. 2020.

[14] W3C.org, "W3c semantic web activity", W3C.org, http://w3.org/2001/sw, Tech. Rep., Jul. 2020.

[15] D. Avola, L. Cinque, G. Foresti, N. Martinel, D. Pannone, and C. Piciarelli, "Low-level feature detectors and descriptors for smart image and video analysis: A comparative study", in. Feb. 2018, pp. 7–29, ISBN: 978-3-319-73890-1. DOI: 10.1007/978-3-319-73891-8_2.

[16] M. S. Kankanhalli and Tat-Seng Chua, "Video modeling using strata-based annotation", *IEEE MultiMedia*, vol. 7, no. 1, pp. 68–74, 2000.

[17] FFMpeg.org, "Ffmpeg documentation", FFMpeg.org, http://ffmpeg.org, Tech. Rep., Jul. 2020.

[18] "Overview of open images v6", Open Images Dataset, http://storage.googleapis.com/openimages/web/factsfigures.html, Tech. Rep., Jul. 2020.

[19] Adobe.com, "Work with metadata in adobe bridge", Adobe.com, http://helpx.adobe.com/bridge/using/metadata-adobe-bridge.html, Tech. Rep., Jul. 2020.

[20] E. Recommendations, "Material exchange format", EBU Recommendations R121, http://mxf.irt.de/information/eburecommendations/R121-2007.pdf, Tech. Rep., Jul. 2007.

[21] H. Kwaśnicka and L. C. Jain, *Bridging the Semantic Gap in Image and Video Analysis*. Berlin, Heidelberg: Springer, 2018, ISBN: 978-3-319-73891-8.

[22] E. Spyrou, D. Iakovidis, and P. Mylonas, *Semantic Multimedia Analysis and Processing*. Boca Raton, Fla: CRC Press, 2017, ISBN: 978-1-351-83183-3.

[23] M. dataset, "The mirflickr retrieval evaluation", LIACS Medialab at Leiden University, http://press.liacs.nl/mirflickr, Tech. Rep., Jul. 2020.

[24] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[25] S. Sami, *Seo For Social Media: It ranked first in the search engines*. Amazon Press, 2020, vol. 1.

[26] R. G. D. Bultermann, "Socially-aware multimedia authoring: Past, present and future", *ACM Transactions on Multimedia Computing Communications and Applications.*, 2013.

[27] S. Krig, "Interest point detector and feature descriptor survey", in. Sep. 2016, pp. 187–246, ISBN: 978-3-319-33761-6. DOI: 10.1007/978-3-319-33762-3_6.

[28] R. Hannane, A. Elboushaki, A. Karim, P. Nagabhushan, and D. M. Javed, "An efficient method for video shot boundary detection and keyframe extraction using sift-point distribution histogram", *International Journal of Multimedia Information Retrieval*, vol. 5, Mar. 2016. DOI: 10.1007/s13735-016-0095-6.

[29] A. Śluzek, *Local Detection and Identification of Visual Data*. LAP LAMBERT Academic Publishing (29 Sept. 2013), 2013.

[30] J. S. Sevak, A. D. Kapadia, J. B. Chavda, A. Shah, and M. Rahevar, "Survey on semantic image segmentation techniques", in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 306–313.

[31] G. Wang, K. Gao, Y. Zhang, and J. Li, "Efficient perceptual region detector based on object boundary", Jan. 2016, pp. 66–78, ISBN: 978-3-319-27673-1. DOI: 10.1007/978-3-319-27674-8_7.

[32] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Event detection and recognition for semantic annotation of video", *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 279–302, 2011. DOI: 10.1007/s11042-010-0643-7. [Online]. Available: https://doi.org/10.1007/s11042-010-0643-7.

[33] R. Arndt, R. Troncy, S. Staab, and L. Hardman, "Comm: A core ontology for multimedia annotation", in. Dec. 2008, pp. 403–421. DOI: 10.1007/978-3-540-92673-3_18.

[34] J. Ni, X. Qian, Q. Li, and X. Xu, "Research on semantic annotation based image fusion algorithm", in *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*, 2017, pp. 945–948.

[35] N. Gayathri and K. Mahesh, "An efficient video indexing and retrieval algorithm using ensemble classifier", in *2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, 2019, pp. 250–258.

[36] F. Zhao and w. zhao, "Learning specific and general realm feature representations for image fusion", *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[37] Q. Hu, C. Wu, J. Chi, X. Yu, and H. Wang, "Multi-level feature fusion facial expression recognition network", in *2020 Chinese Control And Decision Conference (CCDC)*, 2020, pp. 5267–5272.

[38] K. Goh, B. Li, and E. Y. Chang, "Semantics and feature discovery via confidence-based ensemble", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, no. 2, pp. 168–189, May 2005, ISSN: 1551-6857. DOI: 10.1145/1062253.1062257. [Online]. Available: https://doi.org/10.1145/1062253.1062257.

[39] D. A. Norman and S. W. Draper, *User Centered System Design - New Perspectives on Human-computer Interaction*. Justus-Liebig-Universität Gießen: Taylor & Francis, 1986, ISBN: 978-0-898-59872-8.

[40] M. Fowler, *UML Distilled - A Brief Guide to the Standard Object Modeling Language*. Boston: Addison-Wesley Professional, 2004, ISBN: 978-0-321-19368-1.

[41] Apple.com, "Face recognition in apple fotos", Apple.com, https://support.apple.com/de-de/guide/photos/phtad9d981ab/mac, Tech. Rep., Aug. 2020.

[42] S. Wagenpfeil, "Gmaf prototype", University of Hagen, Faculty of Mathematics and Computer Science, http://diss.step2e.de:8080/GMAFWeb/, Tech. Rep., Jul. 2020.

[43] A. S. Foundation, "Apache commons imaging api", Apache Software Foundation, https://commons.apache.org/proper/commons-imaging/, Tech. Rep., Aug. 2020.

[44] Oracle.com, "Java enterprise edition", Oracle.com, https://www.oracle.com/de/java/technologies/java-ee-glance.html, Tech. Rep., Aug. 2020.

[45] Docker.inc, "What is a container", Docker.inc, https://www.docker.com/resources/what-container, Tech. Rep., Aug. 2020.

[46] yWorks GmbH, "Yed graph editor", yWorks GmbH, https://www.yworks.com/products/yed, Tech. Rep., Aug. 2020.

[47] M. X. Bornschlegl, K. Berwind, M. Kaufmann, F. Engel, P. Walsh, M. L. Hemmje, and R. Riestra, "Ivis4bigdata: A reference model for advanced visual interfaces supporting big data analysis in virtual research environments", in *BDA@AVI*, 2016.

[48] *Edison project, european union's horizon 2020 research, grant agreement no. 675419*.

[49] W3C.org, "Sparql query language for rdf", W3C.org, https://www.w3.org/TR/sparql11-overview/, Tech. Rep., Aug. 2013.

[50] I. Schmitt, N. Schulz, and T. Herstel, "Ws-qbe: A qbe-like query language for complex multimedia queries", in *11th International Multimedia Modelling Conference*, 2005, pp. 222–229.