Automatic Annotation of Scientific Video Material based on Visual Concept Detection

Christian Hentschel
Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
christian.hentschel@hpi.uni-potsdam.de

Ina Blümel
German National Library of Science and Technology
Hannover, Germany
ina.bluemel@tib.uni-hannover.de

Harald Sack
Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
harald.sack@hpi.uni-potsdam.de

ABSTRACT

Rapid growth of today's video archives along with sparsely available editorial metadata and too few capacities of libraries and archives for manual annotation demand for efficient approaches of automated metadata extraction. In addition, editorial and non-authoritative metadata is usually not fine-grained enough to describe video on a segment level, which is often required for efficient pinpoint search and retrieval. We consider the use case of the AV Portal provided by the German National Library of Science and Technology – a web based video search engine that offers access to educational video content from various areas of engineering and natural sciences. User studies that have been conducted during the conceptional design stage of the AV Portal have indicated a strong interest of potential users to search for specific visual concepts, like e.g. "landscape", "drawing", "animation", within videos of a particular domain. We present an approach that supports automatic content-based classification of video segments that is tailored to the special requirements of the AV Portal regarding its technology oriented content and academic users. We furthermore show that semantic analysis of the generated metadata not only allows for better retrieval goal definition but also offers explorative search within the archive using visual concepts.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Video Retrieval, Visual Concept Detection, Semantic Analysis

1. INTRODUCTION

Video recordings of academic lectures and scientific experiments provide a valuable source of information for students as well as for scientists. Libraries that collect these data and intend to provide access to a broader audience through newly created web portals report a rapidly growing interest in video data. At present, a major drawback of these collections, however, is the comparatively low availability of text-based editorial metadata that describe the contained information, which is mandatory to provide efficient search and retrieval within these collections. Moreover, even if available, editorial metadata often describe video recordings only on a general level, as e.g., in terms of a video title and an author name, which typically is too coarse for content determination by a non-expert user. With libraries usually maintaining only little capacities for manual annotation and with regard to the sheer mass of available video data today (not to mention those assets newly created every day), considerably large amounts of information are currently rendered useless due to the mere fact that the respective videos are not retrievable and thus invisible for the user.

Inevitably, new approaches for machine supported metadata generation are required to satisfy user needs and reduce the amount of manual labor necessary for content annotation. The *AV Portal* project of the German National Library of Science and Technology[1] is a web based video search engine, which offers access to scientific and educational video recordings such as lectures and experiments from various areas of engineering and natural sciences such as information technology and chemistry. Access to these data is currently limited by the aforementioned lack of metadata. Research conducted within the AV Portal project therefore focuses on enrichment and generation of metadata from all available domains, i.e. on the one hand by analyzing and processing the sparsely available authoritative metadata and on the other hand by automatic extraction of metadata from audio and video streams. Through automated classification of the available audio-visual data streams non-authoritative metadata can be extracted with relatively little manual intervention. In this paper, we focus on detection of *visual concepts* for automatic metadata generation.

During the last years, there have been numerous efforts to open up audiovisual content for efficient retrieval and ac-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
i-Know '13 Graz, Austria
ACM 978-1-4503-2300-0/13/09 ...$15.00
http://dx.doi.org/10.1145/2494188.2494213

[1] AV Portal – `http://www.tib-hannover.de/en/research-and-development/projects/av-portal/`

cess. Yovisto[2] is an academic video search engine collecting video recordings of university lectures and courses as well as recordings of scientific presentations from international conferences. Yovisto enables pinpoint content-based search via state-of-the-art video analysis technologies, as e.g., shot boundary detection, video OCR, and automated speech analysis (ASR) [10]. Users are able to tag and comment video segments for educational purposes. Moreover, Yovisto also applies semantic analysis to annotate videos with semantic metadata that are published for further usage as Linked Open Data and enable semantic exploratory video search [15]. In difference to yovisto, the AV Portal also applies visual concept detection to open up video content. The mediaglobe project established the prototype of a semantic video search engine for documentary video content of a limited domain. Besides shot boundary detection, video OCR, and ASR, also manually trained visual concept detection has been applied to open up the video content for retrieval. Semantic search in mediaglobe is supported via facetted browsing [4]. Although mediaglobe also applies visual concept detection, the proprietary scientific and technical content of the AV Portal demanded special requirements wrt. to determining the appropriate visual concepts according to the users' demands. In [5] Kobilarov et al. present the BBC's approach to join up all of its resources using linked data principles and a tailored ontology for BBC's program data. Later on, the NoTube[3] project harnessed BBC's program information by analyzing it with a NLP-tool to extract named entities and to map them to linked data resources [11]. Both approaches are limited to editorial program data and consider the annotated asset as a whole rather than its temporal segments. In contrast to the BBC program data and the NoTube project the limiting aspect in the AV Portal is the lack of rich editorial material. As a large-scale European library project the Europeana[4] concentrates various research efforts in media archiving, analysis and retrieval in order to protect Europe's cultural heritage. Within this context the authors in [6] recognize the importance of automatic visual content classification and propose an approach for analysis of Europeana images based on principles very similar to the methods applied in this paper for video data. The European AXES project[5] aims at enabling users to explore audiovisual archives. Research efforts target the exploitation of visual concepts for metadata generation and the combination with textual metadata derived from speech transcripts [9].

Visual concept detection (VCD) has been successfully applied to images or video frames to automatically assign labels depending on the presence or absence of depicted objects (e.g. "diagrams") and scenes (e.g. "lecture"). We apply visual concept detection at video segment level and thus are able to provide a comparatively fine granular generation of metadata, having the strong advantage for the user to retrieve only those parts of a video recording, showing the respective aspects of interest. The implemented approach (see Section 3) follows the well-known Bag-of-Features approach using aggregated local histogram of gradient features for video frame description and supervised machine learning based on support vector machines. We have identified relevant visual concepts by conducting a small user survey and carefully selected representative video frames as training material. Finally, we map visual concept labels to unambiguous semantic entities via Named Entity Mapping in order enrich the semantic metadata context and to provide more accurate and complete search results.

This paper is structured as follows: We first give a brief introduction to the AV Portal, the contained videos and the covered domains. We continue with a description of the survey that was designed to find out, which specific visual concepts a user of a particular discipline would find useful if supported as filter or search target by the AV Portal search engine. Furthermore, we describe the process for ground truth data generation. Section 3 presents related work in the domain of visual concept detection in visual data and briefly sketches the approach applied here. Based on a manually created training and testing dataset we evaluate our approach in section 4. In Section 5 we describe the advantages of mapping visual concepts to semantic entities. Finally, we conclude the paper by giving a brief summary of the obtained results as well as an outlook to future work.

2. THE AV PORTAL

The German National Library of Science and Technology (TIB) ranks as one of the largest specialized libraries worldwide covering more than 6 million media units pertaining to all areas of engineering, as well as architecture, chemistry, information technology, mathematics and physics. The TIB's task is to comprehensively acquire and archive literature from around the world in order to provide access to students and scientists as well as to preserve cultural heritage. Next to textual data such as books, journals and patents the TIB provides access to non-textual media formats such as 3D models, research data, or audio-visual media [2]. While archiving and access of text-based information relies on well-established principles for indexing, search, and retrieval, video data demands for completely new techniques for extracting the comprised information and to provide access via appropriate retrieval interfaces. Currently, the TIB archives 2,000 hours of video data (e.g. computer animations, and video recordings of university lectures, scientific conferences and experiments) with an approximated growth of 1,000 hours added every year. The TIB AV Portal is a current research project established in order to create workflows and to develop tools that allow academic libraries to treat audiovisual data in the same way as text documents within the library processing chain and to make it as easy for users to locate and use the growing stock of non-textual material.

A major focus of the AV Portal project lies in the exploitation of techniques for automatic indexing of the *visual content* depicted in video recordings. The user will be able to retrieve specific video segments that *depict* a particular aspect (object or scene) of the respective scientific subject by keyword based retrieval methods. Research efforts in machine vision target the growing demand for efficient automatic visual content classification. The task usually is to automatically recognize categories of depicted objects and scenes (i.e. the visual concepts) – very similar to the requirements posed by the AV Portal. Recognition is usually

[2]Yovisto – `http://www.yovisto.com`
[3]NoTube – `http://notube.tv`
[4]Europeana – `http://http://www.europeana.eu/`
[5]AXES – `http://www.axes-project.eu`

considered as a classification problem of separating positive from negative examples of a given visual concept. Most approaches rely on supervised machine learning techniques that require a set of manually annotated data for training a model of a specific concept.

As for the AV Portal project, we follow a similar approach. During the initial stage we decided to focus on video recordings of 6 subject areas: architecture, chemistry, information technology, mathematics, physics, and engineering. New videos to be archived by the TIB can be easily related to one of these subjects since the content provider is always known and thus the respective discipline. Analysis of the video content therefore concentrates on the recognition of visual concepts per subject meaning that we take advantage of the domain knowledge by training subject specific visual concept classifiers. Our intention was to provide temporal video annotations associated with those video segments that depict the corresponding concept. The selection of visual concepts is based on a user survey and in close cooperation with TIB experts responsible for one of the aforementioned subject areas.

2.1 Concept Definition

In order to evaluate the user needs for video retrieval based on visual concepts we have conducted a small survey among 23 potential users. Using the example of the subject 'architecture', we intended to find out, which specific visual categories a user interested in a particular subject would find useful if supported as filter or search target by the AV Portal search engine. Together with subject experts at the TIB we have devised a number of different 'architecture'-related visual concepts that we proposed as search targets to the survey participants. We divided the available video material (e.g. computer animations, recordings of lectures, scientific conferences or experiments) into animated film vs. real film and queried the desired concepts for both types separately. Figure 1(a) shows the votes for particular concepts for graphical AV material. As depicted, users draw a huge interest in different kinds of (technical) drawings in videos. Figure 1(b) shows the same evaluation for real film recordings. Here, the concepts "interior" and "building" are favored filters to narrow down the search. We have also identified concepts ("urban open space" and "painting") that were considered to be interesting according to the participants' vote, but could hardly be found in the video material under consideration and for which model training therefore had to be postponed until more material becomes available.

Based on the results obtained for the subject 'architecture' and together with the respective subject experts at the TIB we have decided on a final list of visual concepts for the remaining five subjects. Our goal was to obtain concepts that are fully disjoint as we assumed this would simplify the ground truth generation process since the decision of whether a particular key frame belongs to a specific concept would then be unambiguous. A few cross-subject concepts have been considered as equally important for all 6 subjects. Additional subject-specific concepts have been defined by the TIB subject experts by taking into consideration the results of the 'architecture-survey'. Table 1 shows the list of visual concepts used for video annotation. The visual concept label names have been carefully aligned with labels

Cross-Subject	Computer animation, Drawing, Graph, Interview, Lecture/Conference, Screencast, Technical drawing
Architecture	Construction Indoor, Construction site, Interior, Facade Detail, Building, Landscape, Model, Cityscape, 3D/Perspective, Object
Chemistry	Experiment indoor, Experimental lecture, Molecular drawing, Microscopy, Molecular structure, Techniques/ Methods
Information Tech.	Electronic Components
Mathematics	Computer simulation
Physics	Experiment indoor, Molecular structure, Microscopy
Engineering	Experiment indoor, Bridge, Construction site, Meeting, Microscopy, Machine, Shipping, Aircraft technology, Agricultural machine, Landscape

Table 1: Subject-specific and cross-subject concepts in the AV Portal

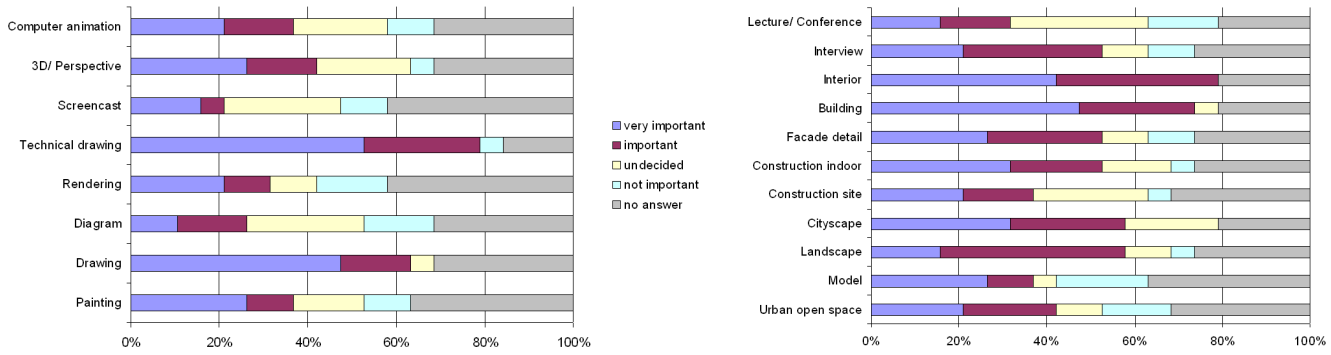
provided by the German Authority File *Gemeinsame Normdatei (GND)*[6] used for cataloging in library context in order to guarantee proper integration of the video meta data in the library process in future. We have specified 7 cross-subject genre and one (mathematics, information technology) to 10 (architecture, engineering) subject-specific genres. Several visual concepts are shared by related disciplines like "molecular structure" and "microscopy" by chemistry as well as by physics. In contrast to visually rich subjects such as engineering or architecture, concept identification for more theoretical subjects such as 'mathematics' has proven to be difficult. Here, object and scene recognition seems to be less important than, for example, transcription of mathematical formulas.

2.2 Ground Truth Data Generation

In the next step, a representative ground truth per visual concept needed to be compiled to serve as training examples in the machine learning process. This was achieved by manually selecting single video key frames depicting the respective concept as positive examples. As this process has been performed by several assistants a detailed description for each of the concepts was required in order to guarantee a homogeneous selection of key frames per concept. Therefore, a number of visual characteristics has been assembled, attributing each concept meaningfully. These descriptions have been provided by the subject experts who had a clear idea in mind when selecting the concepts as potential retrieval candidates. This process not only was important in order to guarantee training data quality but also as a verification step in which humans have crosschecked whether a set of characteristics and thus a concept itself was mutually exclusive from all other concepts as it was assumed that this would increase classification accuracy.

We have started by selecting up to 80 different videos per subject with varying run time. In an automated key frame extraction process (see Section 3) a number of representative frames for each video has been selected. An average amount of 100 key frames has been annotated per visual concept. The number of examples required eventually de-

[6]Gemeinsame Normdatei – `http://d-nb.info/gnd`



(a) Computer graphics

(b) Real film

Figure 1: Illustration of user demands concerning the classification of AV material

pends on the AV material and the visual variance within the respective concept. Visually more complex concepts require more key frames to cover the overall variance. Missing coverage would most likely lead to false dismissals in the classification process since the classifier did not *learn* the concept in its entirety. Special attention has been payed to concepts that were particularly desired by users and subject experts in order to provide a comprehensive ground truth. Key frames that could not be assigned to any of the defined concepts have been left out and were thus not considered as training or testing examples. As one might expect, not every visual concept appears equally frequent in the video material (see Section 4). Thus, the obtained examples sets vary largely wrt. the total number of key frames.

3. CONTENT-BASED CLASSIFICATION

In this Section we briefly present the applied approach for content-based classification of video data. As mentioned in Section 2, we aim at fine-granular, i.e. segment level, concept detection. We therefore start by automatically structuring a video into segments of homogeneous visual characteristics. We especially identify *shot boundaries*, i.e. transitions resulting from varying camera positions and video editing processes. Based on the obtained shot boundaries, representative key frames are extracted, i.e. single video frames that represent the content of the respective video segment. First, these key frames are used as training data by annotating them with the respective depicted concept (cf. 2). Subsequently, they serve as the entity based on which the classifiers will decide for new videos whether the originating video segment should be annotated with the associated concept label. We therefore consider content-based video classification as a special case of image classification, since key frames are nothing but still images extracted from video streams.

Content-based classification of image data has been subject of research for many years and Bag-of-Visual-Words (BoW) image representations have emerged as a successful state of the art when aiming at re-usable visual descriptors capable to represent a wide (potentially unlimited) range of visual concepts ([3, 7, 12]). Their advantage lies in the aggregation of local image features (typically histograms of gradients) in order to describe statistical properties of depicted visual concepts, i.e. by counting representative local features similar to counting words in text retrieval.

In our approach, we extract SIFT (Scale-Invariant-Feature-Transform, [8]) features at a fixed grid of 6×6 pixels on each channel of a key frame in RGB color space. By concatenating these features we obtain a 384-dimensional feature vector at each grid point. Based on these features a visual vocabulary is computed via k-means clustering that delivers a set of representative visual words (codewords). For our approach $k = 4,000$ cluster centers are computed on the RGB SIFT features computed on all key frames from a particular subject. By assigning each of the extracted RGB-SIFT features of a key frame to its most similar codeword (or cluster center) using a simple approximate nearest neighbor classifier, a normalized histogram of codeword frequencies is computed, i.e. a Bag-of-Words, representing this key frame. The combination of SIFT for local key frame description and the BoW model makes the approach invariant to transformations, changes in lighting and rotation, occlusion, and intra-class variations [3].

Once the key frame descriptors have been computed the problem of visual concept recognition can be approached by standard machine learning techniques. Kernel-based Support Vector Machines (SVM) have been widely used in image classification scenarios (cf. [3, 13, 16]). For our approach, we apply a Gaussian kernel based on the χ^2 distance measure, which has proven to provide good results for histogram comparison. Following [16] the kernel parameter γ is approximated by the average distance between all training key frame BoW-histograms. Therefore, the only parameter to be optimized in a 4-fold cross-validation is the cost parameter C of the support vector classification. New key frames can be classified using the aforementioned Bag-of-Words feature vectors and the trained SVM model.

We consider the classification task a one-against-all approach – one SVM per given visual concept is trained to separate the key frames from this concept from all other given concepts. Hence, the classifier is trained to solve a binary classification problem, i.e., whether or not a key frame depicts a specific visual concept. The ground truth generated in Section 2.2 is split into 50% training and 50% testing data. Since video

data tend to be visually rather similar when considering different segments taken from the same sequence and in order to avoid testing on training data we have split our data based on video files, i.e. considering *all* key frames taken from one video either as training *or* as testing data, which however may lead to imbalanced training/testing datasets since the various videos may differ in length and number of segments and thus number of extracted key frames. As mentioned before, due to archiving regulations of the TIB, the video author is always known and thus the originating subject can be automatically derived. This domain knowledge has been introduced into the classification process by training classifiers on a per subject area level in order to slightly alleviate the classification task to result in better classifier accuracy. Furthermore, we have decided not to merge the cross-subject concepts (e.g. 'lecture/conference', cf. Table 1) into a single training and to train a cross-subject classifier. This decision was made due to the fact that e.g. a lecture in chemistry significantly differs from a lecture in architecture (one shows a scientific experiment, the other shows a power point presentation of buildings) and it is assumed that the variance within all disciplines would distort the classifier resulting in inferior accuracy. Negative training examples for a given concept are obtained by assembling all key frames of the respective subject that do *not* depict the concept.

4. EVALUATION

Based on the testing data generated by splitting the available ground truth data, the classification performance of the trained concept models has been evaluated. As already discussed, we have trained subject-specific models and thus tested the performance on subject-specific testing data likewise. By thresholding the classifier output, precision and recall values have been computed for each visual concept. Figures 2(a) - 2(f) visualize the harmonic mean of both values in terms of the F_1Score.

As can be seen, the classification performance varies largely. Figure 2 also shows the number of positive training ('train+') and testing ('test+') examples available. As stated in Section 2 in the current selection of videos used to extract the ground truth not every visual concept is equally well presented. Despite our initial assumption, not all of the 7 cross-subject concepts are truly of equal prominence for all subjects. While 'interview' scenes as well as 'screencast' are concepts, for which we have found representative key frames in videos from almost all subjects, key frames depicting 'computer animation', 'graph', 'drawing', and 'technical drawing' as well as 'lecture/conference' did not provide enough training material (we have decided to limit classifier training to an availability of at least 16 positive training samples) in order to learn models for all subjects.

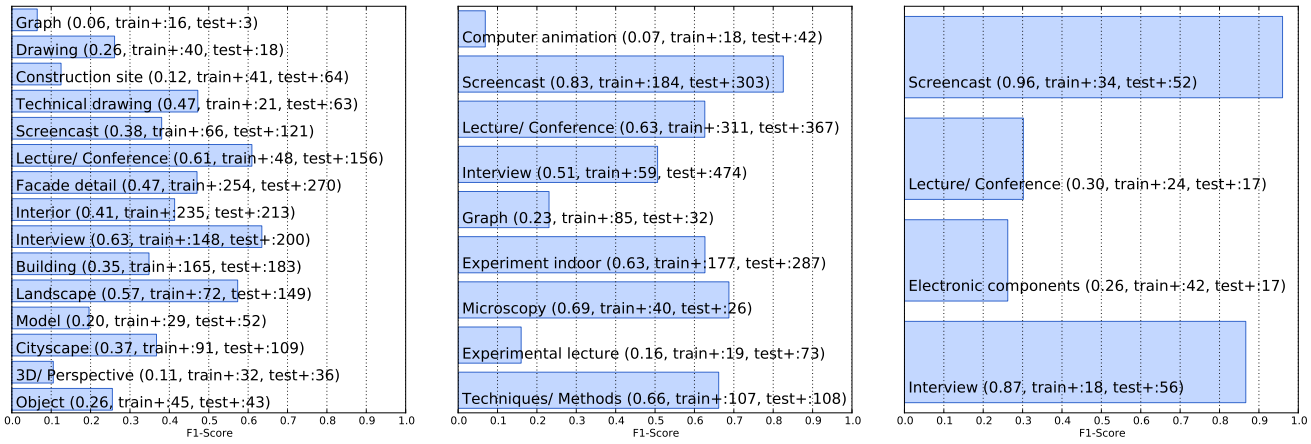
Concept classifiers for the different 'interview'-scenes have shown a rather satisfying performance (average $F_1Score = 0.58$) with a strong exception for mathematics, which should be attributed to a comparatively small set of training examples. Similarly, lower classification performance of 'lecture/conference' (information technology as well as physics), 'graph' (architecture, physics), and 'technical drawing' (chemistry, engineering) can be explained. In general, it should be noted that often a correlation between low training set sizes and low classification performance can be observed. Extremely low classification accuracy (with an $F_1Score < 0.1$) always is attended by very few training examples (< 20) as can be seen for 'engineering/bridge', 'architecture/graph' and 'chemistry/computer animation'. Few training examples typically do not cover sufficient visual variance within the concept and thus the trained classifiers do not recognize respective examples within the testing set. This can be demonstrated, as e. g., in Fig. 3, where the example taken from the training set for the concept 'bridge' clearly differs from the one in the testing set. These findings are in line with the results presented in [1], were the authors have shown that classification performance correlates with trainings set size. On the other hand, comparatively high classification performance, which is based on a small set (see e.g. 'Physics/Technical drawing') should be doubted as they are for once typically due to high recall values. Moreover, visual variance in small test sets is necessarily low and results could have been based on the mere coincidence of these few examples covered in the training data.

Figure 3: Visual variance for example key frames of the visual concept 'bridge' taken from training (left) and testing (right) data sets.

We consider a few examples for visually similar concept classes that may distort the classification accuracy. In Fig. 4, one can clearly observe that examples from both concepts show dominant vertical and horizontal lines necessarily affecting gradient based classifiers. By adding features such as color histograms extracted from color spaces that separate artificial from natural lighting condition, these concepts should possibly be better distinguishable. Similarly, Fig. 5 can easily be confused due to common features such as hard contrasts and large homogeneous background areas. Since both concepts currently only compete in the subject architecture, a general statement is yet difficult to make. However, once they become important for other subjects too, a closer investigation is required.

Finally, Fig. 6 shows visual concepts, where the aimed idea of fully disjoint concepts proved to be difficult to maintain. Buildings are usually part of cityscapes, but sometimes also are present in landscape scenes. While this is not a significant problem to the classification approach presented (multiple classifiers can predict different concepts for the same key frame) it shows that often an initial clear idea of a visual concept is distorted by real-world examples. This, on the other hand, requires fine tuning of the manual annotation process since otherwise an annotator cannot decide whether, as e.g.. Fig. 6(c) is a 'landscape' *or* 'building' example.

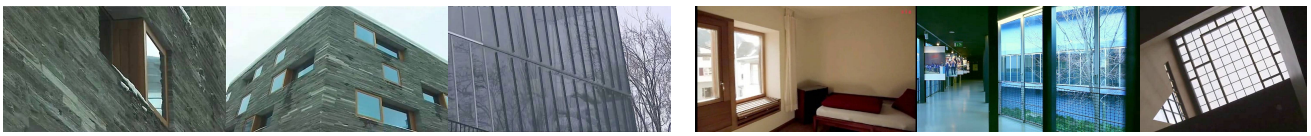
Future work needs to focus on incrementing the ground truth data in order to be able to validate the current classification results on a larger test set. Moreover, visual concepts like



(a) Architecture (b) Chemistry (c) Information Technology

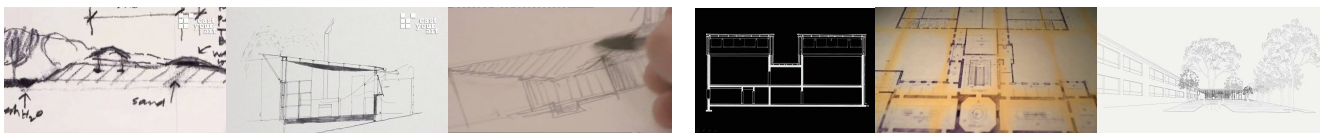
(d) Mathematics (e) Physics (f) Engineering

Figure 2: Visual Concept Evaluation Results ($F_1 Score$)



(a) Facade (b) Interior

Figure 4: Examples taken from concepts "Facade" and "Interior" show similar dominant horizontal and vertical lines.



(a) Drawing (b) Technical drawing

Figure 5: The cross-subject concepts "Drawing" and "Technical drawing" exhibit similar features in contrast and background.

(a) Cityscape (b) Building (c) Landscape

Figure 6: Visual concepts where examples cannot uniquely be mapped to either concept.

'drawing' and 'technical drawing' or 'computer animation' and 'screencasts' that have proven to be difficult to distinguish in an automatic process should be merged into a single superclass. Special attention should also be payed to the selection of negative training examples. As described in Section 3, negative examples currently are obtained by taking all non-positive examples of a given visual concepts. Since many key frames extracted from a video, however, cannot be assigned to any of the described visual concepts, negative training data sampling remains incomplete. Integrating these frames as negative examples in the training process will most like improve the classification accuracy.

5. SEMANTIC CONCEPT LABELING

The label names of visual concepts, as specified in Section 2.1, are rather limited when aiming at semantic interpretation and understanding. Therefore, we have devised a mapping of visual concepts to *semantic entities*. As in philosophy, ontology denotes the study of the nature of being, while epistemology on the other hand is concerned with the nature and scope of knowledge by questioning what knowledge is, how it is acquired, and the possible extent to which a given subject or entity can be known and experienced via our senses. Therefore, we also distinguish pure semantic entities that represent an ontological concept or individuals from visual concepts that merely depict semantic entities, but are not the same as the semantic entities they do depict. Nevertheless, to support the ontological existence of visual concepts for further processing, we have decided to introduce URIs (Uniform Resource Identifiers) for the visual concepts under consideration within our own namespace, as e.g. the visual concept 'building' has been be assigned the URI `http://av.getinfo.de/resource/Building`.

For further semantic processing, the fact that a visual concept also represents a semantic entity being depicted must be expressed. This can simply be achieved by making use of the well known FOAF vocabulary[7] and DBpedia[8] entities to create appropriate RDF statements, as e.g.,

```
av:Building foaf:depicts dbpedia:Building .
```
[9]

This relation enables the inclusion of the visual concepts into the context-sensitive semantic processing of the other text-based video metadata, such as e.g., audio transcripts, video OCR, or authoritative archival data, to increase the accuracy of the Named Entity Mapping. All visual concepts have been manually mapped to semantic entities, which are generally depicted by these visual concepts. In this way, semantic entities represented via visual concepts define part of the context in which potential ambiguities of the text-based metadata can be solved. The process of context-based Named Entity Mapping is explained in further detail in [14].

The AV portal utilizes semantically mapped visual concepts as facet filters to enable exploratory search on the video content, i.e. besides for keywords or semantic entities, also visual concepts can be applied as search query or query refinement. As e.g., the search query for the famous architect 'Norman Foster' can be endorsed by applying the visual concept 'building' as a filter facet to achieve search results of videos that are dealing with Norman Foster while simultaneously depicting buildings.

6. SUMMARY AND OUTLOOK

By means of content-based classification manual annotation of video material in the AV Portal is minimized to an initial training stage of visual classifiers. This enables automatic metadata generation supporting user needs in search and retrieval. Confirmation of the derived visual concepts finally will be provided by a user study with a first integrated AV Portal prototype. We have shown that current approaches for visual concept detection provide satisfying results but require a sufficient amount of training examples to cover the visual variance. In order to validate our results on larger corpora of video data, additional manual annotation work is required to put our findings on a firm footing. Furthermore, we have given evidence that careful concept definition is necessary in order to provide unambiguous descriptions for annotation to generate clean training data. From the perspective of a user, fine granular video annotation at segment level adds a significant value by providing additional metadata that allows for immediate determination of video segments depicting the aspect of interest. This new search experience for video portals remains completely illusive when limited to manual annotation processes. Thereby, an architect is able to narrow down his search by the subject-specific filter "building", or a chemist is able to conduct a targeted search for videos depicting chemical experiments.

Finally, we have briefly outlined the idea of combining visual concept classification and semantic analysis in order to provide semantic unambiguous search results by formalizing the *meaning* targeted by a particular concept label. However, future work is required in order to fully exploit the mapping of semantic visual concepts to knowledge bases such as DBpedia, as e.g., also to enable reasoning over visual concepts.

7. REFERENCES

[7]FOAF vocabulary – `http://www.foaf-project.org/`
[8]DBpedia – `http://dbpedia.org`
[9]the following namespace prefixes are used:
`av:` for `http://av.getinfo.de/resource/`,
`foaf:` for `http://xmlns.com/foaf/0.1/`, and
`dbpedia:` for `http://dbpedia.org/resource/`

[1] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, pages 26–33, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[2] J. Brase and I. Bluemel. Information supply beyond text: non-textual information at the German National Library of Science and Technology (TIB) - challenges and planning. *Interlending & Document Supply*, 38(2):108–117, June 2010.

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, and D. Maupertuis. Visual Categorization with Bags of Keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[4] C. Hentschel, J. Hercher, M. Knuth, J. Osterhoff, B. Quehl, H. Sack, N. Steinmetz, J. Waitelonis, and H. Yang. Open up cultural heritage in video archives with mediaglobe. In G. Eichler, L. W. M. Wienhofen, A. Kofod-Petersen, and H. Unger, editors, *IICS*, volume 204 of *LNI*, pages 190–201. GI, 2012.

[5] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. P. B. Simperl, editors, *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737. Springer, 2009.

[6] I. Kollia, Y. Kalantidis, K. Rapantzikos, and A. Stafylopatis. Improving Semantic Search in Digital Libraries Using Multimedia Analysis. *Journal of Multimedia*, 7(2):193–204, Apr. 2012.

[7] T. Leung and J. Malik. Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[8] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004.

[9] K. McGuinness, R. Aly, S. Chen, M. Frappier, K. Martijn, H. Lee, R. Ordelman, R. Arandjelovic, M. Juneja, C. V. Jawahar, A. Vedaldi, J. Schwenninger, S. Tschopel, D. Schneider, N. E. O'Connor, A. Zisserman, A. F. Smeaton, and H. Beunders. AXES at TRECVid 2011. In *TRECVid Workshop*, Dec. 2011.

[10] H. Sack and J. Waitelonis. Integrating social tagging and document annotation for content-based search in multimedia data. In K. Möller, A. de Waard, S. Cayzer, M.-R. Koivunen, M. Sintek, and S. Handschuh, editors, *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, Athens (GA), USA, November 2006.

[11] B. Schopman, D. Brickley, V. Buser, L. Miller, C. van Aart, L. Aroyo, S. Linguerri, V. Malaisé, M. Minno, M. Mostarda, L. Nixon, D. Palmisano, Y. Raimond, and R. Siebes. Notube: making the web part of personalised tv. volume websci10, 2010.

[12] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, number Iccv, pages 1470–1477. IEEE, 2003.

[13] C. Snoek and M. Worring. Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2009.

[14] N. Steinmetz and H. Sack. Semantic multimedia information retrieval based on contextual descriptions. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 382–396. Springer Berlin Heidelberg, 2013.

[15] J. Waitelonis and H. Sack. Towards exploratory video search using linked data. *Multimedia Tools and Applications*, pages 1–28, 2011.

[16] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73(2):213–238, Sept. 2006.

APPENDIX

Acknowledgments

This work was supported in part by means of the German National Library of Science and Technology under the project AV-Portal. We would like to thank architekturclips.de for providing AV material serving as training data.