Semantic Context Inference in Multimedia Search

Qianni Zhang and Ebroul Izquierdo

School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
{qianni.zhang, ebroul.izquierdo}@elec.qmul.ac.uk

Abstract. Multimedia content is usually complex and may contain many semantically meaningful elements interrelated to each other. Therefore to understand the high-level semantic meanings of the content, such interrelations need to be learned and exploited to further improve the search process. We introduce our ideas on how to enable automatic construction of semantic context by learning from the content. Depending on the targeted source of content, representation schemes for its semantic context can be constructed by learning from data. In the target representation scheme, metadata is divided into three levels: low, mid, and high levels. By using the proposed scheme, high-level features are derived out of the mid-level features. In order to explore the hidden interrelationships between mid-level and the high-level terms, a Bayesian network model is built using a small amount of training data. Semantic inference and reasoning is then performed based on the model to decide the relevance of a video.

Keywords: Multimedia retrieval, context inference, mid-level features, Bayesian network

1 Introduction

In realistic multimedia search scenarios, high-level queries are usually used and search engines are expected to be able to understand underlying semantics in content and match it to the query. Researchers have naturally started thinking of exploiting context for retrieving semantics. Content is usually complex and may contain many semantically meaningful elements interrelated to each other, and such interrelations together form a semantic context for the content. To understand the high-level semantic meanings of the content, these interrelations need to be learned and exploited to further improve the indexing and search process. In latest content-based multimedia retrieval approaches, it is often proposed to build some forms of contextual representation schemes for exploiting the semantic context embedded in multimedia content. Such techniques form a key approach to supporting efficient multimedia content management and search in the Internet. In the literature, such approaches usually incorporates domain knowledge to assist definition of the context representation. However, this kind
of schemes is often limited as they have to rely on specifically defined domains of knowledge or require specific structure of semantic elements. In this chapter we introduce our ideas on how to enable systems to automatically construct a semantic context representation by learning from the content. Depending on the targeted source of content, which could be online databases, a representation scheme for its semantic context is directly learned from data and will not be restricted to the pre-defined semantic structures in specific application domains.

Another problem hampering bridging the semantic gap is that it is almost impossible to define precise mapping between semantics and visual features since they represent the different understandings about the same content by humans and machines. This problem can be approached by dividing all types of metadata extracted from multimedia content into three levels - low, mid and high - according to their levels of semantic abstraction and try to define the mapping between them. Low-level features include audio-visual features, the time and place in which content was generated, etc.; mid-level features can be defined as content-related information that involves some degree of semantics but is still not fully understandable by users; while high-level features contains a high degree of semantic reasoning about the meaning or purpose of the content itself. The mid-level features are relatively easy to obtain by analysing the metadata, but usually are not directly useful in real retrieval scenarios. However, they often have strong relationships to high-level queries while these relationships are mostly ignored due to their implicitness. In this way, defining the links between objects on low- and high- levels becomes more tangible through dividing the mapping process in two steps. While linking low-level features to mid-level concepts are relatively easy to solve using the well-defined algorithms in the state-of-the-art, the mapping between mid- and high-level features are still difficult. In the proposed research, in order to explore the hidden interrelationships between mid-level features and the high-level terms, a Bayesian network model is learned from a small amount of training data. Semantic inference and reasoning is then carried out based on the learned model to decide whether a video is relevant to a high-level query. The relevancies are represented by probabilities and videos with high relevancies to a query are annotated with the query terms for future retrieval. The novelty of the proposed approach has two aspects: using a set of mid-level features to reduce the difficulty in matching low-level descriptors and high-level semantic terms; and employing an automatically learned context representation model to assist the mapping between mid- and high-level representations.

Although a full video annotation and retrieval framework using the proposed approach is mentioned in this chapter, the focus will be on the context learning and automatic inference for high-level features. The mid-level features are assumed to be available and are extracted using algorithms with reasonable performance.

The rest of this chapter is organised as follows: Section 2 gives a review on the state-of-the-art techniques on context reasoning for multimedia retrieval task; Section 3 describes the three representation levels of multimedia content and the implemented mid-level features; Section 4 presents the proposed technique
for semantic context learning and inference for mid-level to high-level matching; Section 5 shows selected experimental results and the chapter is concluded with Section 6.

2 Related Works

The problem of high-level decision making often consists of reasoning and inference, information fusion, and other common features. Popular techniques related to storing and enforcing high-level information include neural networks, expert systems, statistical association, conditional probability distributions, different kinds of monotonic and non-monotonic, fuzzy logic, decision trees, static and dynamic Bayesian networks, factor graphs, Markov random fields, etc [15, 21, 9, 14, 6]. A comprehensive literature review on these topics can be found in [17]. Many object-based image classification frameworks rely on probabilistic models for semantic context modelling and inference [13]

In [16], the authors propose a probabilistic framework to represent the semantics in video indexing and retrieval work. In this approach, the probability density function which denotes the presence of a multimedia object is called a multiject. The interaction between the multijects is then modelled using a Bayesian network to form a multinet. The experiments in this research are restricted to a few particularly selected concepts and may not work in general retrieval scenarios. In [5], authors have developed a framework based on semantic visual templates (SVTs), which attempts to abstract the notion of visual semantics relating particular instances in terms of a set of examples representing the location, shape, and motion trajectories of the dominant objects related to that concept. Bayesian relevance feedback was also employed in the retrieval process. In [18], authors concentrate on analysing wildlife videos that capture hunting. Their approach models semantic concepts such as vegetation, terrain, trees, animals, etc. Neural networks are employed in this approach to classify features extracted from video blobs for their classification task. A coupled HMM approach is presented in [3] for detecting complex actions in Tai Chi movies, a very interesting application. In literature, the Bayesian approach is mainly used in region or object-based image retrieval systems [11, 10, 12], in which the object’s likelihood can be calculated from the conditional probability of feature vectors. Some systems use probabilistic reasoning to exploit the relationship between objects and their features. In [2], the authors propose a content-based semi-automatic annotation procedure for providing images with semantic labels. This approach uses Bayes point machines to give images a confidence level for each trained semantic label. This vector of confidence labels can be exploited to rank relevant images in case of a keyword search. Some other systems employ Bayesian approach in scene classification e.g., indoor/outdoor or city view/landscape [20, 12]. As we have already stated, this kind of classification has been restricted to mutually exclusive categories, and so is only suitable for images that have one dominant concept. But in more realistic scenarios in image classification and retrieval, images are complex and usually consist of many semantically meaningful objects. There-
fore the relationships between semantically meaningful concepts of the objects cannot be ignored and need to be explored to a greater degree.

3 Three-Level Multimedia Representation

The aim of multimedia retrieval techniques is to elicit, store and retrieve the audio- and imagery-based information content in multimedia. Taking images as an example of various digital multimedia content types, the search for a desired image from a repository might involve many image attributes including a particular combination of colour, texture or shape features, a specific type of object; a particular type of event or scene; named individuals, locations or events; subjectively associated emotions; metadata such as who created the image, where and when. According to these attributes, the types of query in a multimedia retrieval scenario can be categorised into three levels of increasing complexity corresponding to the used features [7]:

“**Level 1**: Primitive features such as colour, texture, shape, sound or the spatial location of image elements.”

“**Level 2**: Derived features involving some degree of logical inference about the identity of the objects depicted in the image.”

“**Level 3**: Abstract attributes involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted.”

The first level of primitive features can be directly derived from the visual content, and does not require employing any knowledge base. It is relatively easy to generate this type of features but they are mostly restricted to applications about retrieving strong visual characters, such as trademark registration, identification of drawings in a design archive or colour matching of fashion accessories. The second and third levels are more commonly demanded in real-world scenarios as semantic image retrieval. However, the semantic gap lying between levels 1 and 2 hampers the progress of multimedia retrieval area [19]. The semantic gap is commonly defined as “the discrepancy between low-level features or content descriptors that can be computed automatically by current machines and algorithms, and the richness, and subjectivity of semantics in high-level human interpretations of audiovisual media”. To bridge the semantic gap has become the most challenging task in semantic multimedia retrieval. Thus in this paper, by dividing this task into two steps: matching **Level 1** (low-level features) to **Level 2** (mid-level features) and matching **Level 2** (mid-level features) to **Level 3** (high-level features), we hope to minimise the difficulty in the task and try to find feasible solutions.

4 Semantic Inference for Video Annotation and Retrieval

The proposed semantic context learning and inference approach analyses the inter-relationships between the high-level queried concepts and mid-level features by constructing a model of these relationships automatically and use it to infer high-level terms out of mid-level features in future search.
Figure 1 shows the work flow of this approach. There are two processes in the work flow, the learning process and the inference process. In the learning process which is usually carried out off-line. First, several mid-level features are extracted using any specifically designed classifiers. A subset of the database randomly selected for training purpose is then manually annotated on the high-level query concept. Then extracted mid-level features and the manual annotations on the training subset is used to derive the semantic context model. In this research Bayesian network is used for modelling the semantic context involving mid-level features and the particular query concept. The learning process concerns learning of both the network structure and probability tables of nodes.

One important feature in this module is that the Bayesian network model is constructed automatically using a learning approach based on K2 algorithm, which is basically a greedy search technique. It is able to automatically learn the structure of a Bayesian model based on a given training dataset. In this algorithm, a Bayesian network is created by starting with an empty network and iteratively adding a directed arc to a given node from each parent node. The iteration is terminated when no more possible additions could increase a score which is calculated as:

\[
K2\text{score} = \prod_{j=1}^{p} \frac{\Gamma(\sum_k a_{ijk})}{\Gamma(\sum_k a_{ijk} + \sum_k b_{ijk})} \prod_{k=1}^{q_i} \frac{\Gamma(a_{ijk} + b_{ijk})}{\Gamma(a_{ijk})}
\]
where \(i, j\) and \(k\) are respectively the index of the child node, the index of the parents of the child node, and the index of the possible values of the child node. \(p\) is the number of different instantiations of parent nodes, \(q_i\) is the number of possible values of the \(i\)th child node, \(b\) is the number of times that the child node has the value of the \(k\)th index value of the node, and \(a\) is the number of times that the parents and the child correlate positively in discrete cases. This selection criterion is basically a measure of how well the given graph correlates to the data. Due to the scope of this paper, we give only a brief introduction to \(K2\) algorithm here. If the reader is interested in more details about this algorithm, please refer to \(\S\).

Then in the inference stage, when an un-annotated data item is present, the Bayesian network model derived from the training stage conducts automatic semantic inferences for the high-level query. The Bayesian network in this case represent the semantic context involving the mid-level features of this video and their underlying links to the high-level query.

Assume that, for dataset \(D\), a total number of \(n\) pre-defined mid-level features are available: \(M = \{m_i| i = 1, 2, ..., n\}\). For the high-level query \(q\) in the same dataset \(D\), each \(m_i\) are supposed to be more or less interrelated to \(q\), and \(M\) together form the semantic context of \(D\) and can be used as the ‘semantic evidence’ for carrying out context reasoning for \(q\). Consider an example case in which five mid-level features are used, \(n = 5\), then the joint probability that \(q\) exists in a piece of multimedia content considering the semantic context can be calculated as:

\[
P(M,q) = P(q) \cdot P(M|q) = P(q) \cdot P(m_1,m_2,\ldots,m_5|q) \\
= P(q) \cdot P(m_1|q) \cdot P(m_2,m_3,m_4,m_5|m_1,q) \\
= P(q) \cdot P(m_1|q) \cdot P(m_2|m_1,q) \cdot P(m_3,m_4,m_5|m_1,m_2,q) \\
= P(q) \cdot P(m_1|q) \cdot P(m_2|m_1,q) \cdot P(m_3|m_1,m_2,q) \\
\quad \cdot P(m_4,m_5|m_1,m_2,m_3,q) \\
= P(q) \cdot P(m_1|q) \cdot P(m_2|m_1,q) \cdot P(m_3|m_1,m_2,q) \\
\quad \cdot P(m_4|m_1,m_2,m_3,q) \cdot P(m_5|m_1,m_2,m_3,m_4,q)
\]

Assuming the components of \(M\) are independent to each other, the joint probability takes the form of

\[
P(M,q) = P(q,m_1,m_2,\ldots,m_5) = P(q) \cdot P(m_1|q) \cdot P(m_2|q)\ldots \cdot P(m_5|q)
\]

A similar expansion can be obtained for an arbitrary number of classes \(n\):

\[
P(M,q) = P(q,m_1,m_2,\ldots,m_n) = P(q) \cdot \prod_{i=1}^{n} P(m_i|q)
\]

## 5 Experiments

The experiments were carried out on the good sized un-edited video database. Videos were segmented into shots and each shot was treated as an independent
multimedia element for analysis and annotation. Several mid-level features were selected by observing the content and they belong to Level 2 of the three-level representation scheme as described in Section 3. These mid-level features include: ‘Regular shapes’, ‘2D flat-zone’, ‘Human faces’, ‘Flashlight’, ‘Text’, ‘Vegetation’, ‘Ship’, ‘Music’, ‘Speech’.

On top of that, two high-level queries have been carefully selected considering those commonly exist in the database with reasonable proportions and have relatively rich connections to the mid-level features: *City view from helicopter* (CVH) and *Football stadium*.

![Bayesian model for CVH](image)

**Fig. 2.** Bayesian model for CVH

For the sake of illustration, an example of the learned model structures is shown in Figure 2 considering the case for *City view from helicopter* and some examples of probability distribution (PD) tables are given in Table II.

In an attempt to read the networks manually, one may note that these learned structures and parameters are sometimes not meaningful in the understanding of human-beings. This is because they only reveal the underlying relationships between the query concept node and feature nodes from the machine’s perspective, as the overall aim in the proposed semantic inference approach is to build a context network from available metadata rather than those containing full human intelligence in brains. The considered mid-level features and high-level queries are merely a small subset in the specific scenario due to the lack of rich subjects in the used video content. However, the experiment carried out based on this specific scenario shows that suitable high-level queries can be obtained purely based on semantic reasoning on a small subset of mid-level features.

The training processes for each semantic inference model were performed on a randomly selected subset of less than 10% of the whole dataset and this process
took only a few seconds on a PC with Pentium D CPU 3.40GHZ and 2.00GB of RAM.

Ground-truth annotations of the two high-level terms have been made available on the whole experiment datasets. Based on these ground-truth annotations, we managed to obtain statistical evaluations on the results to show the performance of the proposed approach, as given in Table 2.

### Table 1. Examples of PD tables for CVH

| Camera motion |           |           |
|---------------|-----------|-----------|
| CVH True      | 0.169     | 0.831     |
| True          |           |           |
| False         | 0.002     | 0.998     |

| Text          |           |           |
|---------------|-----------|-----------|
| CVH True      | 0.101     | 0.899     |
| True          |           |           |
| False         | 0.236     | 0.764     |

| Vegetation    |           |           |
|---------------|-----------|-----------|
| CVH True      | 0.764     | 0.236     |
| True          |           |           |
| False         | 0.966     | 0.034     |

### Table 2. Evaluation results

| City view from helicopter | TP | TN | FP | FN | Precision | Recall | Accuracy | ROC area |
|----------------------------|----|----|----|----|-----------|--------|----------|----------|
| 10                         | 2891 | 4 | 63 |    | 71.40%    | 13.70% | 97.74%   | 63.1%    |

| Football stadium       | TP | TN | FP | FN | Precision | Recall | Accuracy | ROC area |
|-------------------------|----|----|----|----|-----------|--------|----------|----------|
| 4                       | 2950 | 3 | 11 |    | 57.10%    | 26.70% | 99.53%   | 97.1%    |

As it can be observed from Table 2, the detection performance of the proposed approach is reasonably good, given the limited accuracy of the mid-level feature extractors and the abstractness and sparse distribution of the query terms throughout the dataset.

### 6 Conclusions

In this chapter an approach for semantic context learning and inference has been presented. The basic idea was to organise the representations for multimedia content in three semantic levels, by adding a mid-level feature category
between low-level visual features and high-level semantic terms. Thus, the mapping between low-level content and high-level semantics could be achieved in two steps, which were easier to achieve than a single step. This research has concentrated on the second step, high-level semantic extraction based on the semantic context involving a number of mid-level features. This was achieved by automatically building a context model for the relationships between the mid-level features and the query. Modelling and inference in this case were carried out using the $K_2$ algorithm. The proposed approach was tested on a large size video dataset. The obtained results have shown that this approach was capable of extracting very abstract semantic terms that were scarcely distributed in the database.

Acknowledgments. The research that leads to this chapter was partially supported by the European Commission under the contracts FP6-045189 RUSHES and FP7-247688 3DLife.

Open Access. This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Boutell, M., Luo, J.: Beyond pixels: Exploiting camera metadata for photo classification. Pattern recognition 38(6), 935–946 (2005)
2. Bradshaw, B.: Semantic based image retrieval: a probabilistic approach. In: Proceedings of the eighth ACM international conference on Multimedia, pp. 167–176 (2000)
3. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 994–999 (1997)
4. Chang, E., Goh, K., Sychay, G., Wu, G.: Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology 13(1), 26–38 (2003)
5. Chen, W., Chang, S.F.: Generating semantic visual templates for video databases. In: 2000 IEEE International Conference on Multimedia and Expo, 2000. ICME 2000, vol. 3 (2000)
6. De Jong, F.M.G., Westerveld, T., De Vries, A.P.: Multimedia search without visual analysis: the value of linguistic and contextual information. IEEE Transactions on Circuits and Systems for Video Technology 17(3), 365–371 (2007)
7. Eakins, J.P., Graham, M.E.: Content-based image retrieval: A report to the jisc technology applications programme. Tech. rep., Institute for Image Data Research, University of Northumbria at Newcastle (1999), http://www.jisc.ac.uk/uploaded_documents/jtap-039.doc
8. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine learning 9(4), 309–347 (1992)
9. Fan, J., Gao, Y., Luo, H., Jain, R.: Mining multilevel image semantics via hierarchical classification. IEEE Transactions on Multimedia 10(2), 167–187 (2008)
10. Fei-Fei, L., Fergus, R., Perona, P.: A bayesian approach to unsupervised one-shot learning of object categories. In: Proc. ICCV, vol. (2003)
11. Fergus, R., Perona, P., A., Zisserman, o.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (2003)
12. Hoiem, D., Sukthankar, R., Schneiderman, H., Huston, L.: Object-based image retrieval using the statistical structure of images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (2004)
13. Kherfi, M.L., Ziou, D.: Image collection organization and its application to indexing, browsing, summarization, and semantic retrieval. IEEE Transactions on multimedia 9(4), 893–900 (2007)
14. Koskela, M., Smeaton, A.F., Laaksonen, J.: Measuring concept similarities in multimedia ontologies: Analysis and evaluations. IEEE Transactions on Multimedia 9(5), 912–922 (2007)
15. Lavrenko, V., Feng, S., Manmatha, R.: Statistical models for automatic video annotation and retrieval. In: IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP’04, vol. 3, IEEE Computer Society Press, Los Alamitos (2004)
16. Naphade, M.R., Huang, T.S.: A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Transactions on Multimedia 3(1), 141–151 (2001)
17. Naphade, M.R., Huang, T.S.: Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. IEEE Transactions on Neural Networks 13(4), 793–810 (2002)
18. Qian, R., Haering, N., Sezan, I.: A computational approach to semantic event detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 200–206 (1999)
19. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on pattern analysis and machine intelligence 22(12), 1349–1380 (2000)
20. Vailaya, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.J.: Image classification for content-based indexing. IEEE Transactions on Image Processing 10(1), 117–130 (2001)
21. Zhu, X., Wu, X., Elmagarmid, A.K., Feng, Z., Wu, L.: Video data mining: Semantic indexing and event detection from the association perspective. IEEE Transactions on Knowledge and Data engineering, 665–677 (2005)