Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Inferring semantics from textual information in multimedia retrieval

Mats Sjöberg *, Jorma Laaksonen, Timo Honkela, Matti Pöllä [1]

Adaptive Informatics Research Centre, Helsinki University of Technology, P.O. Box 5400, 02015 TKK, Finland

ARTICLE INFO

Available online 1 May 2008

Keywords:
Multimodal databases
Multimedia semantics
Content-based retrieval
Self-Organizing Maps
Symbol grounding problem

ABSTRACT

We propose a method for inferring semantic information from textual data in content-based multimedia retrieval. Training examples of images and videos belonging to a specific semantic class are associated with their low-level visual and aural descriptors augmented with textual features such as frequencies of significant words. A fuzzy mapping of a semantic class in the training set to a class of similar objects in the test set is created by using Self-Organizing Maps (SOMs) trained from the low-level descriptors. Experiments with two databases and different textual features show promising results, indicating the usefulness of the approach in bridging the gap from low-level visual features to semantic concepts.

1. Introduction

In the last decade the amount of digital information available to the public has been increasing ever more rapidly. Particularly the volume of multimedia and multimodal data has been growing in recent years. This development has been driven by the increasing availability and uptake of digital cameras, mobile phones with camera capabilities, and digital video cameras. This trend can also be seen in the increasing popularity of multimedia sharing web sites such as YouTube, Flickr and Google Video.

This exciting development puts an increasing emphasis on the development of automated content-based retrieval methods that index and retrieve multimedia information based on its contents. Such methods, however, suffer from a serious problem: the *semantic gap*, i.e. the wide gulf between the low-level features used by computer systems and the high-level semantic concepts understood by human beings. In this article we propose a method of using different textual features for inferring semantics from textual information to help bridge the semantic gap from visual features to semantic concepts.

We have used our PicSOM [21] content-based information retrieval (CBIR) framework to test our proposed method on two very different visual databases. The first one is a set of videos and semantic classes from the NIST's TRECVID 2005[2] development data set. The TRECVID video set contains TV broadcasts in different languages and textual data acquired by using automatic speech recognition and machine translation software where appropriate. The video shots in the annotated set are accompanied with verified semantic ground truth concepts such as "videos depicting explosions or fire". Our initial experiments with this data have been published earlier in [42].

The second data set is from the "Pockets full of memories" art installation that was on display in the Centre Pompidou National Museum of Modern Art, Paris, France from April 10 to September 3, 2001 [24]. The visitors contributed over 3300 objects by digitally scanning images of them and describing them with names and keywords. These descriptions are in different languages, mostly French and English. The visitors could also add semantic descriptions of the objects by moving continuous-valued sliders between eight different property pairs, such as old–new, functional–symbolic, etc.

In the approach we use, several Self-Organizing Maps (SOMs) [16] are trained in an unsupervised manner with visual, aural and textual feature data calculated from the objects of a multimodal database. Then all objects in the training set that belong to a given semantic class are mapped onto these SOMs. This mapping generates relevance value fields on the SOM surfaces, which can further be mapped to the objects of the test set. These relevance values can be interpreted as membership values of a fuzzy set corresponding the given semantic class. The objects of the test set can then be ordered according to these relevance values gained from the SOM mapping.

In this article we complement the basic set of low-level visual and aural features with different types of statistical features calculated from the text associated to the visual objects. The motivation for this is that the textual data utilizes human language and is therefore closer to human semantic perception.

* Corresponding author. Tel.: +358 9 451 3267; fax: +358 9 451 3277.
E-mail address: mats.sjoberg@tkk.fi (M. Sjöberg).

[1] Supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Centre of Excellence in Adaptive Informatics Research*.

[2] http//www-nlpir.nist.gov/projects/trecvid/

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.01.029

In the current experiments we have used more textual features than in our earlier experiments. We present experiments using word histogram and keyword frequency features using SOMs and a binary keyword method using an inverted file. This setting is actually more general since any method which generates an ordering of objects of any modality in the test set can be used. As long as different modalities have well-defined associations, e.g., keyframe images as parts of video shots that contain them, they can be combined in the multimodal fusion stage of our system.

The main contributions of this article are twofold. First, we propose a novel method for combining different modalities in an efficient and natural way, even incorporation of different types of indices (e.g. SOM, inverted file). Secondly, we compare different textual features used both in isolation and in combination with visual and aural features in semantic retrieval. In doing this we also gain an insight into the performance of the SOM-based methodology in comparison with the computationally more heavy inverted file index.

The rest of the article is organized as follows. First, Section 2 introduces the question of extracting the "statistical" semantic content from both natural language and visual data domains. Then, Section 3 describes our PicSOM framework for CBIR and discusses how textual features can help in bridging the semantic gap between visual features and high-level concepts. The feature extraction methods are explained in Section 4 and the two databases used in our experiments in Section 5. The results of the experiments are presented in Section 6 and, finally, the conclusions are drawn in Section 7.

2. Semantics in natural language and visual data

Traditionally, *semantics* is an area of linguistics that deals with meaning. Semantics is usually used to refer to (i) the relation that some sign has to objects or events and (ii) the relation that a sign has to other signs. Often the signs, like words in some natural language, have hierarchical relationships and these structures are referred to as *taxonomies*. Notions of *semantic categories* and *concepts* are also often used.

In the context of this article, the most interesting question is whether the semantic information expressed in some natural language could be automatically inferred to the extent that would prove to be useful for the purpose of multimedia retrieval. For this goal, it can be assumed that very precise formalization of the natural language will not be needed.

Aside from the practical needs of improving the existing content-based retrieval techniques, a more profound question is associated with the *symbol grounding problem* [9]. If one could implement a statistical framework for analyzing the visual and textual aspects of a multimodal object conjointly, then it might be viable to find common groundings for some semantic concepts in the language and visual domains. Or, it could be possible to ground some language concepts visually or vice versa.

2.1. Statistical presentation of language semantics

Serious efforts to develop computerized systems for natural language understanding have taken place for more than half a century. However, the more general the domain or complex the style of the text, the more difficult it is to reach a high quality of understanding. All systems need to deal with problems like ambiguity and lack of semantic coverage and utilize pragmatic insight. The methodological realm of semantic processing of language is still largely dominated by predicate logic.

Maybe the most striking example of formalization of natural language is the work of Montague [31]. Examples of the language considered in his work include sentences like "Bill walks", "every man walks", "the man walks", and "John finds a unicorn". It may be fair to say that most of the linguistic phenomena are set aside. The idea of being rigorous may often lead to the negligence of the original complexity of the phenomenon being considered [47].

Handling a computerized form of written language rests on the processing of discrete symbols. Similarity in the appearance of the words does not usually correlate with the content they refer to. As a simple example one may consider the words "window", "glass" and "widow". The words "window" and "widow" are phonetically close to each other, whereas the semantic relatedness of the words "window" and "glass" is not reflected by any simple metric. This motivates the selection to use symbolic presentation of natural languages such as English on the word-level instead of character, phoneme or syllable levels.

Contextual information has been widely used in statistical analysis of natural language corpora (consider, e.g., [3,41]). One useful numerical representation can be obtained by taking into account the sentential context in which the words occur. First, we represent each word by a vector in an n-dimensional space, and then code each context as an average of vectors representing the words in that context. In the simplest case, the dimensionality n can be taken equal to the number of different words, and each word is represented by a vector with one element equal to one and others equal to zero. Then the context vector simply gives the frequency of each word in the context. In information retrieval, a similar approach is called as the *bag-of-words technique*, applied in methods related, e.g., to the vector space model [40]. For computational reasons the dimension may be reduced by different methods, e.g. random projection [14] or Latent Semantic Indexing (LSI) [6].

The main tool used in our experiments for dimension reduction is the SOM. Earlier, the SOM has been used in the analysis of word context data, e.g., by [37] (artificially generated short sentences), and [12] (Grimm's fairy tales). It has also been used for finding semantic relationships between document titles in a small document collection [25] and cluster documents according to their textual similarity [10], and even for very large document collections [17]. In [8], a SOM analysis of word contexts was performed with a one-dimensional map in order to find synonymous words (see also [12]). The result can be called a SOM of words, or a *word category map*. Earlier, the name Self-Organizing Semantic Map has also been used. Similar results have also been presented in [27–29]. In the current article, a new approach for extracting contextual information is employed based on the occurrence frequencies of a set of statistically computed keyphrases.

2.2. Image semantics

Image semantics can be defined as understanding the conceptual content of images. The understanding is tightly intertwined with the abilities (i) to segment an image in a relevant manner, (ii) to detect invariant features and (iii) to relate these features with semantic categories.

The definition of semantics, as outlined above, is typically based on the linguistic level. On the other hand, the final or reference relation of many words is with some visually perceivable object or event. For instance, one may attempt to define the meaning of the word "horse" by indicating its position in a hierarchical system as an animal and a living object, or through some features that can be expressed in language. However, the meaning of the word "horse" is also strongly related to its shape or the variety of shapes as a living object.

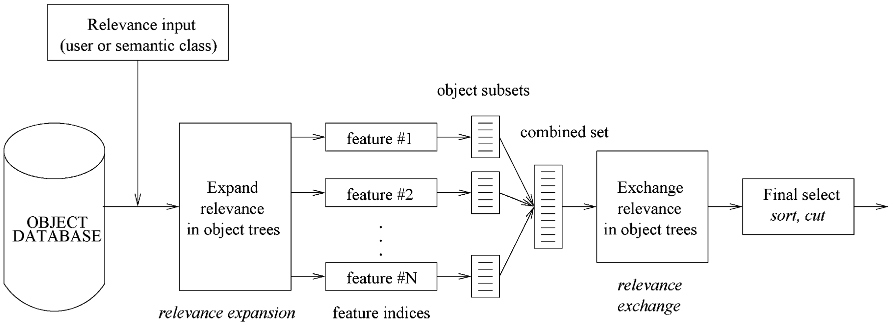

Fig. 1. Processing stages in PicSOM.

Traditionally there has not been any particularly useful way to create, for instance, a functional shape or texture description framework. Therefore, the semantic descriptions of visual content have focused on representations that can be expressed in some linguistic form. Purely linguistic or symbolic models of semantics can, though, be considered deficient as, e.g., pointed out by Harnad as the symbol grounding problem [9]. This means that defining word meaning purely in terms of other words or symbols produces circular definitions as in a dictionary. Humans, however, solve this problem naturally, by grounding many words in perception and physical experience. Many computational systems now try to incorporate sensory data, such as vision and hearing, to improve learning [38]. This has lead us to explore methods using visual data in conjunction with textual data in a machine learning system.

The relationship between image content and semantic concepts has become a subject of recent intensive study in the field of CBIR. The goal has been given various names ranging from "image-to-word transformation" [32], "matching words and pictures" [1], "image auto-annotation" [30], "automatic image captioning" [35], to "automatic image annotation" [7], depending on the selected viewpoint and the specific tasks the authors have been addressing. Various different technical methods and their combinations have been applied, including co-occurrence statistics [32], expectation maximization (EM) [1], support vector machines (SVM) [7], latent semantic analysis (LSA) [30], and Markov random fields (MRF) [2]. The SOM have also been applied in our previous works for inferring semantics from automatically segmented images [23,45,22,46].

In summary, we claim that both images and linguistic expressions provide semantic information and these sources of information are partially complementary. This idea is also empirically supported by the results of the experiments reported in this article, where we show that in most cases visual and textual features perform best in semantic retrieval when used together and not in isolation.

3. PicSOM CBIR system

The CBIR system PicSOM [21] has been used as a framework for the research described in this article. *Query by example* (QBE) is the main interactive operating principle in PicSOM, meaning that the user provides the system a set of example objects of what he or she is looking for, taken from the existing database. This *relevance feedback* information is then expanded to related objects. This would expand relevance both from a video to its constituent keyframe images, and also from individual keyframes to the video shot that contains it.

The PicSOM system has originally been used in interactive mode where the user influences the retrieval of the system with relevance feedback and the results will improve in each iteration. In this article, however, there is no interaction as the experiments have been performed in an offline mode, where pre-defined semantic classes of the training set are used for ordering a separate test set in the order of decreasing relevance or similarity to each semantic class. This is because we are here interested in the mapping abilities of the SOMs for semantic concepts using a mixture of textual and visual features.

3.1. Retrieval with low-level visual features

PicSOM uses several SOMs [16] in parallel to index and determine the similarity and relevance of database objects for retrieval. These parallel SOMs have been trained with different data sets acquired by using different feature extraction algorithms on the objects in the database. This results in each SOM arranging the same objects differently, according to similarities with respect to the corresponding feature. A visual overview of the PicSOM processing stages is shown in Fig. 1.

In the interactive mode each database object receives a relevance value based on user input. In the offline mode, these values are initialized by using pre-defined semantic classes of objects. For each object type (e.g. video, image, text), all relevant-marked objects in the database of that type get a positive weight inversely proportional to the total number of relevant objects of the given type. Similarly the non-relevant objects get a negative weight inversely proportional to their total number. The grand total of all weights is thus always zero for a specific type of objects. On each SOM, these values are summed into the best-matching units (BMUs) of the objects. This results in the formation of sparse relevance value fields on the map surfaces.

The sparse relevance values on the maps are low-pass filtered or "blurred" to spread the relevance information between neighboring units. Due to the topology preserving property of the SOM we can expect neighboring map units to be similar, which motivates this procedure. The filtering is performed by convolving the value fields by a tapered kernel function. This produces to each map unit a *qualification value*, which is given to all objects that are mapped to that unit (i.e. have it as the BMU). Map areas with a mixed distribution of positive and negative values will even out in the blurring, and get a low-average qualification values. Conversely in an area with a high density of mostly positive values, the units will reinforce each other and spread the positive values to their neighbors. This process automatically weights the maps according to their ability to map relevant objects coherently and densely in specific map areas.

The next processing stage is to combine the qualification values gained from each map to the corresponding objects. These values are again shared with related objects. For example the relevance of a video clip is obtained as the sum of the values of its keyframe images, audio and possible textual content. In the final stage, the test set objects are ordered in the decreasing order of qualification values. This ordering, we argue, estimates the decreasing order of relevance of the objects to the semantic target of the retrieval task. In the interactive retrieval mode, a specific number of objects with the highest qualification values would be returned to the user as retrieval results of that round. In the offline mode a performance measure is calculated as the final result.

3.2. Bridging the semantic gap with textual features

The PicSOM system was initially designed for images, and particularly using low-level statistical visual features only. Such features describe images on a very low-abstraction level, for example local color distributions, and do not generally correspond very well with the human perception of an image. In the TRECVID experiments, to be described in this article, we have also used motion and sound features, but the problem remains the same: a very low-level feature description cannot match abstract human understanding.

However, textual features have a close relationship to semantic concepts, as they describe the human language, which has a much closer relation to human understanding and meaning of concepts than any low-level visual features. By including textual features we hope to bring the feature and concept levels closer to each other and thus help to bridge the semantic gap. By using SOM techniques this is done in a fuzzy manner, providing only semantic class membership values for each object. This fuzziness is appropriate as such relationships can never be defined exactly, even by human beings.

A notable philosophical advantage can be seen to result from using SOM techniques for both visual and textual data: the different information domains will become commensurable when the extracted statistical features are mapped on the SOM surfaces. Consequently, the interplay of the information modalities will not be hindered by the fact that the information originates from conceptually different sources. It is important to note that this is a general property of the PicSOM system, and not restricted to the specific feature types and modalities mentioned in this article.

4. Feature extraction

In the PicSOM system, several feature extraction methods can be applied for each object type in the database. For example, from all images one can calculate color, shape and texture descriptors and from videos aural and motion descriptors. For each feature extraction method a separate SOM is trained with the resulting feature vectors.

Each feature will then organize the database objects differently on the SOM surface according to its discriminative properties. These features will, in general, perform differently for different semantic classes and different sets of objects. However, the automatic weighting performed in PicSOM will reward features that discriminate relevant and non-relevant objects well, i.e. SOMs that map semantic classes into well-defined clusters. Conversely, features that perform clustering badly will get a very low weight, ensuring that their impact on the final result will be very small.

In the following sections we will describe shortly the different visual and aural features used in our experiments, and then the textual features in more detail.

4.1. Visual and aural features

As still image features we have used both our own non-standard features as well as the standardized MPEG-7 descriptors [13] calculated using the MPEG-7 Experimentation Model (XM) Reference Software [33]. The following MPEG-7 still image features were used: Edge Histogram, Homogeneous Texture, Color Structure and Color Layout.

If we treat the values in the different color channels of the HSV color space as separate probability distributions, we can calculate the three first central moments: mean, variance and skewness. This produces our simple non-standard color moments feature. The Zernike moments [15] feature describes the overall shape of the border of an object. This border can be obtained from the segmentation mask of a segmented image. Texture neighborhood is a simple textural feature that examines the luminance values of the eight-neighborhood of each inner pixel in an image. The values of the feature vector are then the estimated probabilities for each 8-neighborhood position that the corresponding neighbor pixel is brighter than the central pixel.

For the video content we used the standard MPEG-7 Motion Activity descriptor and our own non-standard temporal features of color and texture data. A temporal video feature is calculated as follows. Each frame of the video clip is divided into five spatial zones: upper, lower, left, right and center. A still image feature vector is calculated separately for each zone and then concatenated to form frame-wise vectors. The video clip is temporally divided into five non-overlapping video sub-clips or slices of equal length. All the frame-wise feature vectors are then averaged within the slices to form a feature vector for each slice. The final feature vector for the entire video clip is produced by concatenating the feature vectors of the slices. For example, using the three-dimensional average RGB color still image feature we would get a temporal video feature vector with a dimensionality of $3 \times 5 \times 5 = 75$.

The purpose of the concatenations in our temporal features is to capture how the averaged still image features change over time in the different spatial zones. Such features make sense since the video clips used in our experiments are short enough for the averages within the slices to be meaningful, while still having some variations between different slices. The variations result, for example, from some object moving from a spatial zone to another.

As an audio feature the Mel-scaled cepstral coefficient (MFCC), or shortly Mel cepstrum, was used. This feature is commonly used for speech recognition, but can be used with other sounds as well [5]. Mel cepstrum is the discrete cosine transform (DCT) applied to the logarithm of the Mel-scaled filter bank energies.

The Mel cepstrum feature should be able to detect speech of different persons and particularly separate speech from other natural sounds, and also music, such as theme music in a TV program or commercial. In news broadcasts this might; however, be problematic since there can be non-natural sounds, such as explosions. Furthermore, many sounds may be overlapped by others, such as a reporter narrating. This feature is calculated using an external program created by the Speech recognition group at the Laboratory of Computer and Information Science at the Helsinki University of Technology.[3]

4.2. Word histogram

The word histogram feature is a statistical textual descriptor which is calculated in three stages. First a histogram is calculated for each textual object (document) in the database giving the

[3] http://www.cis.hut.fi/projects/speech/

frequencies of all the words in that text excluding stop words (i.e. commonly used words such as "the"). Then the document-specific histograms are combined into a single histogram or dictionary for the whole database. The final word histogram feature vectors are calculated for each document by comparing its word frequencies to the dictionary, i.e. the words in the database-wide histogram. For each word in the dictionary we calculate the *term frequency-log-inverse document frequency* (tf-log-idf) weight [39] for the document. The tf-log-idf weight is commonly used in information retrieval and is given as the product of the term frequency and the logarithm of the inverse document frequency.

The feature extraction procedure can be formulated mathematically as follows: The term frequency for a word k in one document is calculated as

$$\mathrm{tf}_k = \frac{n_k}{\sum_{j \in K_\mathrm{D}} n_j}, \tag{1}$$

where n_k is the number of occurrences of the word k in that document. The denominator gives the number of occurrences of all dictionary words K_D in that same document (again excluding stop words). The corresponding document frequency is calculated as

$$\mathrm{df}_k = \frac{N_k}{N}, \tag{2}$$

where N_k is the number of documents where the word k appears, and N is the total number of documents in the collection. The tf-log-idf is then given as the product of Eq. (1) and the log-inverse of Eq. (2):

$$\mathrm{tf\text{-}log\text{-}idf}_k = \frac{n_k}{\sum_{j \in K_\mathrm{D}} n_j} \log \frac{N}{N_k}. \tag{3}$$

The resulting feature vector is then composed of the tf-log-idf values of all dictionary words k for that document. The final dimensionality this vector can be very large, even for moderately sized databases. Dimensionality reduction can then be employed, for example in our experiments we reduced the dimensionality to 100 by using singular value decomposition.

4.3. Keyword frequency

Another textual feature used in our experiments is the keyword frequency which is calculated as follows. First, all the texts of all objects belonging to a given semantic class are concatenated into a class-specific corpus. Then, a list of keyword candidates is extracted based on the frequencies of the keywords in that corpus. The potential keywords may consist of one or more words, in our experience the length is limited to three or four words.

Next, a reference corpus is utilized to select such keywords from the frequency list that are common in the class corpus under examination, but more rare in the reference corpus; i.e. keywords that best distinguish the class-specific corpus from the reference corpus. The purpose of the reference corpus is to represent a certain language in general, and typically a very large and well-balanced corpus is chosen for the task. Thus, comparing a particular-domain corpus to the reference corpus should reveal keywords that are specific to that domain only.

The keyword lists of both corpora are sorted according to the keyword frequencies, and all keywords are given their rank in the list (all keywords that have an equal frequency receive an equal rank). Then, the keywords of the class-specific corpus are processed one by one, calculating for each keyword the ratio of the ranks of the keyword in the two corpora.

Finally, the keywords are sorted in an ascending order according to their rank ratios. Keywords that receive a low ratio are the ones we are interested in, since they were more common in the particular corpus than in the reference corpus. On the other hand, the middle ground of the keyword list is now occupied by general, probably non-specific keywords that were common in both corpora, and the end of the list has keywords that were more frequent in the reference corpus [11]. An idea very similar to the keyword frequency method has also been presented in [4].

In this article we used the keyword frequency feature only for the TRECVID database experiments, since the "Pockets full of memories" database had only a few keywords per object which was not deemed sufficient for this method. We used the text data from the entire TRECVID database as the reference corpus. The semantic classes were quite small in comparison to the entire database and it could thus be seen as sufficiently neutral. The overall TRECVID database corpus represents mostly news broadcast type text with a considerable amount of "noise" from the automatic speech recognition and possible machine translation. If we would have used a completely separate reference corpus such peculiarities may been deemed as significant. This, we fear, was the case in our previous experiments, where we used an external reference corpus with worse results [44].

4.4. Binary keywords

The binary keyword method uses an inverted file index that contains a mapping from words to all the database objects containing them. As such, the inverted file index is a very commonly used method for information retrieval, but is computationally very heavy. A recent extension of the PicSOM system allows the usage of such an inverted file as a replacement for SOM indexing by BMUs [18]. Instead of mapping the textual objects of a given semantic class to BMUs, we instead seek the most informative keywords of a class and map them directly to other objects using the inverted file.

The binary keyword features were generated by gathering concept-dependent lists of the most informative terms or keywords. Let us denote the number of objects in the training set associated with semantic class c as N_c, and assume that of these objects, $N_{c,k}$ contain the keyword k in the textual data. Using words not found in the stop word list, the following measure can be calculated for keyword k regarding the class c:

$$S_c(k) = \frac{N_{c,k}}{N_c} - \frac{N_k}{N}, \tag{4}$$

where N_k is the total number of documents that contain the keyword k and N is the total number of documents in the collection. If this measure has a high positive value it means that the keyword k has a higher frequency in the class c than in the collection as a whole. A negative value with high magnitude would similarly indicate a word that appears less commonly in the class than in the collection generally. For every semantic class c, we record the 10 or 100 most informative keywords k according to the $S_c(k)$ measure. The number of used keywords depends on which one gives better retrieval performance for that class in cross-validation. A class-specific inverted file is then created as mapping from these informative words to the database objects that contain them. In the PicSOM system a measure indicating the closeness of a textual object i to the semantic class c used in generating the inverse file can be calculated as

$$S_{i,c} = \sum_k \frac{\delta_{i,k}}{N_k}, \quad \text{where } \delta_{i,k} = \begin{cases} 1 & \text{if } k \text{ exists in } i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In Eq. (5) the sum is taken over all words k in the inverse file. N_k is the total number of documents that contain the keyword k. The higher the value of $S_{i,c}$ for a specific textual object is, the closer it is

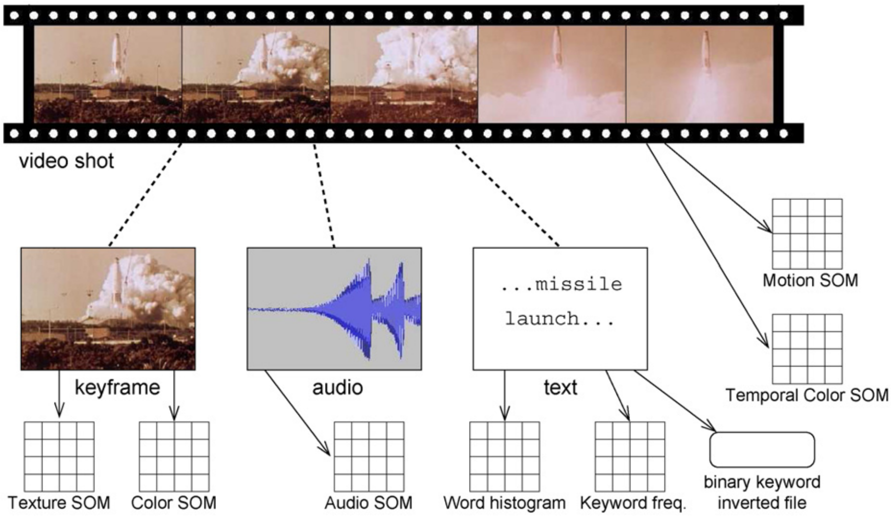

Fig. 2. The hierarchy of videos and examples of multi-modal SOMs.

deemed to be to the given class *c*. The value of this measure is then added to the qualification values of objects produced by the other, SOM-based, features.

5. Data

We have used two rather different data sets in our experiments. The TRECVID video database is a large and rich multimedia database with videos, keyframe images, audio and text. In contrast the "Pockets full of memories" database contains simple single-object images with corresponding keywords. These databases will be presented in more detail in the following sections.

5.1. *TRECVID video data*

Our research group at the Helsinki University of Technology has taken part in the NIST's TRECVID video retrieval evaluations in 2005 [19], 2006 [43] and 2007 [20]. In the experiments described in this article we have used the TRECVID 2005 data, which contains about 790 videos divided into a total of almost 100 000 video clips. From the set of videos originally used as the TRECVID development data we picked only those that had some associated textual data and semantic classifications, resulting in a set of about 35 000 video clips. These video clips were used for the experiments described in this article. Each video clip has one or many keyframes, which are representative still images taken from the video. Also the sound of the video was extracted as audio data. NIST provided textual data acquired by using automatic speech recognition software and machine translation from Chinese (Mandarin) and Arabic to English.

In the PicSOM system the videos and the parts extracted from these were arranged as hierarchical trees as shown in Fig. 2, with the main video as the parent object and the different extracted media types as child objects. In this way the relevance assessments can be transferred bidirectionally between related objects in the PicSOM algorithm as described in Section 3. From each media type different features were extracted, and SOMs were trained from these as is shown with some examples in the figure.

As image features for the video key frames we used Edge Histogram, Homogeneous Texture, Color Structure and Color Layout from MPEG-7. Additionally we used a Canny edge detection feature which was provided by NIST. From the videos we calculated the MPEG-7 Motion Activity feature, as well as separate non-standard temporal features based on average RGB color, texture neighborhood and color moments (see Section 4.1). From the audio data we calculated the MFCCs.

Table 1
Semantic classes from the TRECVID 2005 data set

Semantic class description	A priori (%)	Training set	Test set
An explosion or a fire	1.08	109	265
Regional territory graphic (map)	1.90	376	282
A US flag	0.79	123	151
An exterior of a building	7.28	1578	943
Waterscape or waterfront	2.30	375	420
A captive person	0.16	23	32
Any sport in action	2.59	460	437
A car	7.27	1279	1239
All objects	100	17 230	17 407

As textual features we used word histogram, keyword frequency and binary keywords. The texts were stemmed beforehand using the Porter stemming algorithm [36]. The feature vectors initially produced by the word histogram feature had a dimensionality of about 27 000, which was reduced to 100 by using singular value decomposition.

A total of 39 semantic sets were provided with the TRECVID 2005 development data. These are each a set of video clips that belong to a given semantic class, for example videos depicting "an exterior of a building". These video clips were cooperatively annotated during the TRECVID evaluation using semantic class definitions provided by LSCOM [34]. For these experiments we divided the original TRECVID development set into training and test sets. Table 1 shows the eight semantic classes that were used in our experiments. The semantic class description, shown in the first column, is a shortened version of the one that was used in TRECVID. The second column gives the a priori probability of the class. The third and fourth columns in the table give the number of videos in the training set and in the test set, respectively.

5.2. *"Pockets full of memories" data*

The "Pockets full of memories" data set consists of 3327 objects originating from an art installation that was on display in

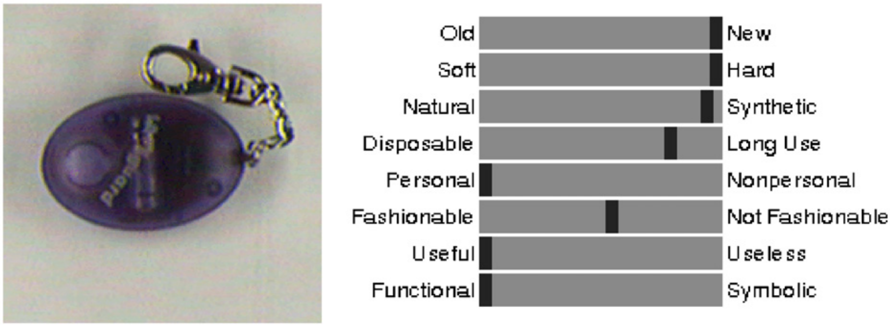

Fig. 3. An "attackalarm" and the property values given by its owner.

the Centre Pompidou National Museum, Paris, France in 2001. During the installation visitors were encouraged to contribute their personal items, which were scanned to produce a digital image of them. The visitors provided the objects with a name, a set of keywords and an evaluation of eight semantic property pairs. These pairs were: old–new, soft–hard, natural–synthetic, disposable–long use, personal–nonpersonal, fashionable–not fashionable, useful–useless and functional–symbolic. The quantifications were given using a touch sensitive screen, where the visitors could select a continuous value between the two extremes. In our experiments we have scaled the resulting values to the range $[-1, 1]$. In Fig. 3 an example is shown of a scanned image and the corresponding property evaluations.

In the original art installation the images were then organized by the SOM algorithm that positioned objects of similar descriptions near each other in a two-dimensional "wall of objects" which was displayed in the gallery [24]. In later research we also looked at the visual contents of the images and the correlation between the visual and semantic information in this database using SOMs [44]. In this article we have used this same data to study the mapping of a set of semantic classes, using both visual and textual features.

Before the visual feature extraction we applied automatic segmentation on the images to separate the main object from the background. This was a comparatively easy task, since most backgrounds were relatively homogeneous as can be seen in the example in Fig. 3. The segmentation algorithm used is described in more detail in [44]. The final visual features were thus extracted only from the area of the image which the segmentation algorithm had identified as belonging to the object (i.e. not the background). The segmentation also enabled the use of the Zernike moments feature for describing the shape of the objects.

As visual features MPEG-7 Edge Histogram, Zernike moments and color moments were used, the same set of features as in our previous experiments with this database. As textual features we used word histogram and binary keywords. We decided to do without stemming of the words since the keywords had been written in at least two different languages, English and French. The vector length of the word histogram feature was 3924, which could be used directly without any reduction because the database was quite small, only 3327 objects.

The range of the owner-given semantic properties was divided into three equal parts and we selected the top and bottom parts as two semantic classes. For example objects with old–new property values in the range $[-1, -\frac{1}{3}]$ were selected as belonging to the semantic class *old*, and those in the range $[\frac{1}{3}, 1]$ to *new*. Table 2 summarizes all classes that could be created in this way. The second column gives the a priori probability of the class. The third and fourth columns in the table give the number of objects in the training and test sets, respectively. The percentages are quite large, which indicates that people tend to pick values in the extremes of the property ranges.

Table 2
Semantic classes from the "Pockets full of memories" data set

Semantic class description	A priori (%)	Training set	Test set
Old	27.1	501	400
New	60.6	938	1077
Soft	32.8	541	551
Hard	53.8	894	895
Natural	23.2	343	429
Synthetic	70.4	1207	1134
Disposable	24.0	422	375
Long-use	67.3	1083	1155
Personal	70.7	1135	1217
Nonpersonal	21.3	373	336
Fashionable	54.8	860	963
Not-fashionable	24.5	426	389
Useful	74.6	1212	1270
Useless	17.9	332	264
Functional	58.5	903	1044
Symbolic	31.5	571	476
All objects	100	1678	1649

The first image in Fig. 4 shows a SOM organized according to the MPEG-7 Edge Histogram feature. Each SOM unit is represented by a visual label which is the most similar image of the database in that feature space. Similar objects can be seen to form clusters, and within the clusters the object properties change continuously, thus retaining the topographical ordering of the feature space.

Below in Fig. 4, the distributions of the semantic classes *soft* and *natural* have been mapped onto the Edge Histogram SOM. The dark areas represent map units to which many objects from that semantic class have been mapped to. One immediately notes a clear correlation between the *soft* and *natural* classes. There seems to be a large set of objects that are both soft and natural, roughly in the middle of the Edge Histogram SOM. Visual inspection of the SOM labels indicates that these are mostly human hands. In addition, the two distributions cluster quite cleanly, indicating that the feature is very discriminative when evaluating these semantic properties.

6. Experimental results

When evaluating the performance of a content-based retrieval system it is important to measure how well the system manages to rank the relevant objects before the non-relevant ones. These two aspects can be evaluated by the two basic information

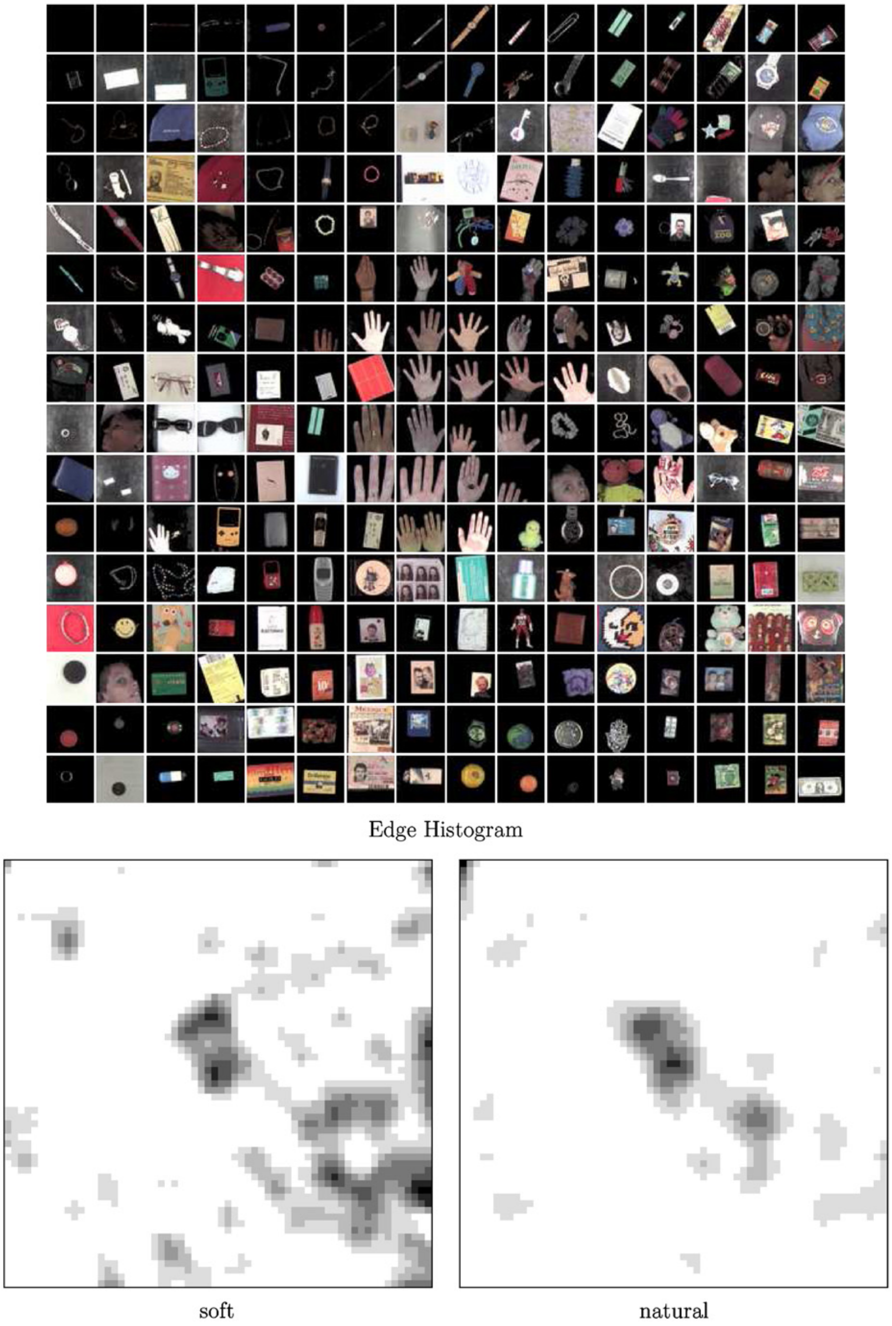

Fig. 4. The MPEG-7 Edge Histogram SOM and the distribution of two semantic classes.

retrieval measures: *precision* and *recall*. Precision is the percentage of relevant objects in the set of returned objects thus far, while recall is the percentage of all relevant objects that are in the returned set. In general both aspects are important, and one way to combine them both into a single measure is to use the non-interpolated average precision of retrieval [26].

The non-interpolated average precision is formed by calculating the precision after each retrieved relevant object, thus implicitly including also the recall. The final per-class performance measure is obtained by averaging these precisions over the total number of relevant objects. In this calculation, the precision is defined to be zero for all non-retrieved relevant objects. In the TRECVID experiments, only the 2000 objects deemed to be the most relevant ones were evaluated, while in "Pockets full of memories" all 1649 objects in the test set were ranked each time. The per-class average precision was finally averaged over all semantic classes of that database to generate an overall average precision.

Several experiments were run with the two databases. Each experiment was performed separately for each of the semantic classes. In the TRECVID runs there were seven experiments, each for a different combination of features: only non-textual features (nt), the three textual features: word histogram (wh), keyword frequency (kwf) and binary keywords (bkw) used alone, and then in combination with the non-textual features (wh + nt, kwf + nt, bkw + nt). The binary keyword method used different inverted files for each semantic class as explained previously. In the "Pockets full of memories" runs we performed only (nt, wh, bkw,

Table 3
Average precision results for TRECVID experiments

Semantic class	nt	wh	kwf	bkw	wh + nt	kwf + nt	bkw + nt
Explosion or fire	0.0567	0.0061	0.0104	0.0285	0.0595	0.0583	**0.0779**
Map	0.3396	0.0061	0.0411	0.0049	0.3402	0.3402	**0.3433**
Depicting US flag	0.0713	0.0023	0.0052	0.0059	0.0763	0.0825	**0.0848**
Exterior of building	0.0988	0.0108	0.0042	0.0068	0.0985	**0.0996**	0.0990
Waterscape or front	0.2524	0.0056	0.0090	0.0053	0.2482	0.2484	**0.2515**
Captive person	0.0054	0.0043	0.0025	0.0000	**0.0161**	**0.0161**	0.0158
Sport in action	0.2240	0.0108	0.0338	0.0970	0.2207	0.2328	**0.2666**
A car	0.2818	0.0177	0.0065	0.0114	**0.2824**	0.2820	0.2800
Average	0.1662	0.0080	0.0141	0.0200	0.1677	0.1700	**0.1774**

Table 4
Average precision results for "Pockets full of memories" experiments

Semantic class	nt	wh	bkw	wh + nt	bkw + nt
Old	0.262	0.302	0.344	0.309	**0.350**
New	0.594	0.630	**0.706**	0.636	0.699
Soft	0.337	0.398	0.514	0.405	**0.522**
Hard	0.546	0.562	0.704	0.573	**0.706**
Natural	0.260	0.353	**0.462**	0.343	0.458
Synthetic	0.677	0.723	**0.852**	0.727	**0.852**
Disposable	0.237	0.302	**0.394**	0.297	0.393
Long-use	0.641	0.698	**0.821**	0.695	0.820
Personal	0.652	0.729	**0.806**	0.717	0.804
Nonpersonal	0.248	0.255	0.399	0.270	**0.400**
Fashionable	0.502	0.586	**0.650**	0.578	0.644
Not-fashionable	0.253	0.269	0.348	0.264	**0.354**
Useful	0.692	0.758	**0.842**	0.753	0.839
Useless	0.188	0.211	0.327	0.229	**0.331**
Functional	0.557	0.639	**0.740**	0.631	0.737
Symbolic	0.322	0.354	0.496	0.368	**0.503**
Average	0.436	0.486	**0.588**	0.487	**0.588**

wh+nt, bkw+nt), since the keyword frequency was not suited for this case where we had only a very limited number of words per object.

The experiment results for the TRECVID runs are summarized in Table 3 and those for the "Pockets full of memories" in Table 4. The best results for each semantic class are indicated in bold face. The results show how the retrieval performance generally increases when the textual features are used in addition to the visual and aural ones. In TRECVID the textual features alone perform very poorly, but still give a significant advantage when combined with non-textual ones. On the other hand, in the "Pockets full of memories" experiments, the opposite occurs: the textual features are better than the visual ones. In this case using the textual features combined with the non-textual features does not improve the results significantly; however, it does not reduce the results even though the visual features perform badly by themselves. Overall the binary keywords make a substantial improvement, while the keyword frequency (in TRECVID) and word histogram features lead to smaller improvements.

The textual data in TRECVID has been produced using speech recognition and in some cases machine translation from Chinese or Arabic. This results in very "noisy" textual data, since speech recognition is never perfect, and machine translation reduces the quality even further. A manual inspection of the texts reveals many unintelligible words and sentences. Still, a sufficient number of important keywords seem to get through to make a significant difference in the results. This, however, means that the process is largely based on finding a small set of relevant keywords for each semantic class. In this sense the TRECVID case is similar to "Pockets full of memories" where the textual data is just a set of keywords. This may explain why the inverted file-based binary keyword method works in most cases better than the SOM-based textual features. The binary keyword method directly increases the qualification values of the test set documents with the correct keywords. The SOM features and indices; however, work in a more indirect way, increasing the qualification values of nearby keyword vectors, which may be suboptimal in this case.

The success of the textual features in the "Pockets full of memories" data can to some extent be explained by the relatively limited variety of different objects depicted in the database. For example, keys and phones, which are commonly found in people's pockets, are always hard and synthetic. Therefore, many of the mappings from keywords to semantic classes are quite unambiguous.

In these experiments, the keyword frequency feature shows a significant improvement compared to our earlier results [42]. This is due to improvements in the method, and the fact that the method now compares the words of the videos belonging to the semantic class to the entire database itself, not an entirely external corpus, as explained in Section 4.3.

7. Conclusions

In this article, we have studied how low-level visual features extracted from images and videos could be complemented with textual information. The nature of the used textual data has been twofold: semantic keywords for images and speech recognition output for videos. In both cases, we have analyzed how the incorporation of the textual information domain effects the average precision of content-based retrieval and compared it to the baseline of purely non-textual retrieval. As textual features have a closer relation to the semantic concepts as expressed in natural language they can be used, we believe, to narrow the semantic gap.

The texts and keywords were used in our experiments in three different ways and the relative performance increases were measured. Statistical word histogram features and the binary keyword-based inverse file method could be used for both the speech recognition and keyword data. In addition, statistical keyword frequency features were used for the speech recognition output. The central finding of the performed evaluations was that the inclusion of the textual information always improved the retrieval accuracy. In a majority of cases, the binary keyword method was the best one. The two SOM-based statistical methods, the word histograms and keyword frequencies, provided almost similar results.

Looking at the results from the opposite direction, the inclusion of the visual (and aural when available) low-level features always increased the average retrieval precision obtained with purely textual methods. This suggests that the two information domains complement each other, which, of course, is beneficial for inferring further semantic knowledge from the interplay of these two different modalities.

This reasoning returns us to the symbol grounding problem discussed in Section 2, i.e. that some symbols that humans use, e.g. some words, have to be grounded in a physical experience such as vision. Aside from the philosophical questions, these issues also become important when trying to build intelligent machines that are able to understand humans and their environment. Being able to infer connections between words and images is a step in this direction.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2003) 1107–1135 (special issue on Machine Learning Methods for Text and Images).

[2] P. Carbonetto, N. de Freitas, K. Barnard, A statistical model for general contextual object recognition, in: Proceedings of the 8th European Conference on Computer Vision, Prague, 2004.

[3] K. Church, P. Hanks, Word association norms, mutual information and lexicography, Comput. Linguist. 16 (1990) 22–29.

[4] F. Damerau, Evaluating domain-oriented multiword terms from texts, Inf. Process. Manage. 29 (1993) 433–447.

[5] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, in: A. Waibel, K. Lee (Eds.), Readings in Speech Recognition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, pp. 65–74.

[6] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.

[7] J. Fan, Y. Gao, H. Luo, Multi-level annotation of natural scenes using dominant image components and semantic concepts, in: Proceedings of the 12th annual ACM International Conference on Multimedia, New York, NY, 2004.

[8] S. Finch, N. Chater, Unsupervised methods for finding linguistic categories, in: I. Aleksander, J. Taylor (Eds.), Artificial Neural Networks, vol. 2, North-Holland, Amsterdam, 1992.

[9] S. Harnad, The symbol grounding problem, Physica D 42 (1990) 335–346.

[10] T. Honkela, S. Kaski, K. Lagus, T. Kohonen, WEBSOM—self-organizing maps of document collections, in: Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6, Neural Networks Research Centre, Helsinki University of Technology, Espoo, Finland, 1997, pp. 310–315.

[11] T. Honkela, M. Pöllä, M.-S. Paukkeri, I. Nieminen, J.J. Väyrynen, Terminology extraction based on reference corpora, Technical Report, Laboratory of Computer and Information Science, Helsinki University of Technology, Espoo, Finland, 2007.

[12] T. Honkela, V. Pulkki, T. Kohonen, Contextual relations of words in Grimm tales analyzed by self-organizing map, in: Proceedings of ICANN-95, International Conference on Artificial Neural Networks, vol. 2, EC2 et Cie, 1995.

[13] ISO/IEC, Information technology—Multimedia content description interface—Part 3: Visual, 15938-3:2002(E) (2002).

[14] S. Kaski, Dimensionality reduction by random mapping: fast similarity method for clustering, in: Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN 98), vol. 1, Anchorage, AK, USA, 1998.

[15] A. Khotanzad, Y.H. Hong, Invariant image recognition by Zernike moments, IEEE Trans. Pattern Anal. Mach. Intell. 12 (5) (1990) 489–497.

[16] T. Kohonen, self-Organizing Maps, in: Springer Series in Information Sciences, vol. 30, 3rd ed., Springer, Berlin, 2001.

[17] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive text document collection, IEEE Trans. on Neural Networks 11 (3) (2000) 574–585.

[18] M. Koskela, J. Laaksonen, E. Oja, Use of image subset features in image retrieval with self-organizing maps, in: Proceedings of Third International Conference on Image and Video Retrieval (CIVR 2004), Dublin, Ireland, 2004.

[19] M. Koskela, J. Laaksonen, M. Sjöberg, H. Muurinen, PicSOM experiments in TRECVID 2005, in: Proceedings of the TRECVID 2005 Workshop, Gaithersburg, MD, USA, 2005. Available online at: ⟨http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html⟩.

[20] M. Koskela, M. Sjöberg, V. Viitaniemi, J. Laaksonen, P. Prentis, PicSOM experiments in TRECVID 2007, in: Proceedings of the TRECVID 2007 Workshop, Gaithersburg, MD, USA, 2007. Available online at: ⟨http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html⟩.

[21] J. Laaksonen, M. Koskela, E. Oja, PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions, IEEE Trans. Neural Networks 13 (4) (2002) 841–853 (special issue on Intelligent Multimedia Processing).

[22] J. Laaksonen, V. Viitaniemi, Emergence of ontological relations from visual data with self-organizing maps, in: T. Honkela, T. Raiko, J. Kortela, H. Valpola (Eds.), Proceedings of the 9th Scandinavian Conference on Artificial Intelligence (SCAI 2006), Espoo, Finland, 2006.

[23] J. Laaksonen, V. Viitaniemi, M. Koskela, Emergence of semantic concepts in visual databases, in: Proceedings of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 05), Espoo, Finland, 2005.

[24] G. Legrady, T. Honkela, Pockets full of memories: an interactive museum installation, Visual Commun. 1 (2) (2002) 163–169.

[25] X. Lin, D. Soergel, G. Marchionini, A self-organizing semantic map for information retrieval, in: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '91, ACM, New York, NY, USA, 1991.

[26] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[27] R. Miikkulainen, A distributed feature map model of the lexicon, in: Proceedings of 12th Annual Conference of the Cognitive Science Society, Lawrence Erlbaum, Hillsdale, NJ, 1990.

[28] R. Miikkulainen, Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory, MIT Press, Cambridge, MA, 1993.

[29] R. Miikkulainen, Self-organizing feature map model of the lexicon, Brain and Language 59 (1997) 334–366.

[30] F. Monay, D. Gatica-Perez, On image auto-annotation with latent space models, in: Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA, 2003.

[31] R. Montague, The proper treatment of quantification in ordinary english, in: J. Hintikka, J. Moravcsik, P. Suppes (Eds.), Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics, D. Reidel, Dordrecht, 1973.

[32] Y. Mori, H. Takahashi, R. Oka, Image-to-word transformation based on dividing and vector quantizing images with words, in: Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

[33] MPEG, MPEG-7 visual part of the eXperimentation Model (version 9.0), iSO/IEC JTC1/SC29/WG11 N3914 (January 2001).

[34] M.R. Naphade, L. Kennedy, J.R. Kender, S.-F. Chang, J.R. Smith, P. Over, A. Hauptmann, A light scale concept ontology for multimedia understanding for TRECVID 2005, Techinical Report, IBM, 2005.

[35] J.-Y. Pan, H.-J. Yang, P. Duygulu, C. Faloutsos, Automatic image captioning, in: Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 2004.

[36] M. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[37] H. Ritter, T. Kohonen, Self-organizing semantic maps, Biol. Cybern. 61 (4) (1989) 241–254.

[38] D. Roy, Grounding words in perception and action: computational insights, Trends Cognitive Sci. 9 (8) (2005) 389–396.

[39] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, in: Computer Science Series, McGraw-Hill, New York, 1983.

[40] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (11) (1975) 613–620.

[41] H. Schütze, Dimensions of meaning, in: Proceedings of Supercomputing, 1992.

[42] M. Sjöberg, J. Laaksonen, M. Pöllä, T. Honkela, Retrieval of multimedia objects by combining semantic information from visual and textual descriptors, in: Proceedings of 16th International Conference on Artificial Neural Networks (ICANN 2006), Springer, Athens, Greece, 2006. Available online at: ⟨http://dx.doi.org/10.1007/11840930_8⟩.

[43] M. Sjöberg, H. Muurinen, J. Laaksonen, M. Koskela, PicSOM experiments in TRECVID 2006, in: Proceedings of the TRECVID 2006 Workshop, Gaithersburg, MD, USA, 2006. Available online at: ⟨http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html⟩.

[44] M. Sjöberg, V. Viitaniemi, J. Laaksonen, T. Honkela, Analysis of semantic information available in an image collection augmented with auxiliary data, in: I. Maglogiannis, K. Karpouzis, M. Bramer (Eds.), Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations, IFIP, Springer, Athens, Greece, 2006. Available online at: ⟨http://dx.doi.org/10.1007/0-387-34224-9_70⟩.

[45] V. Viitaniemi, J. Laaksonen, Keyword-detection approach to automatic image annotation, in: Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT 2005), London, UK, 2005.

[46] V. Viitaniemi, J. Laaksonen, Focusing keywords to automatically extracted image segments using self-organising maps, in: M. Nachtegael, D.V. der Weken, E.E. Kerre, W. Philips (Eds.), Soft Computing in Image Processing: Recent Advances, Studies in Fuzziness and Soft Computing, vol. 210, Springer, Berlin, 2006, pp. 121–156 (Chapter 5).

[47] H. von Foerster, Understanding Understanding, Springer, New York, 2003.

Mats Sjöberg received his M.Sc. (Tech.) degree in engineering physics and mathematics in 2006 from Helsinki University of Technology (TKK), Finland. Presently he works as a researcher at the Adaptive Informatics Research Centre at TKK. He is the author of many publications on image and multimedia analysis. He has acted as a reviewer for a number of conferences and journals. His current research interests include content-based multimedia retrieval and analysis, in particular the emergence of semantics in multimodal databases.

Jorma Laaksonen received his Dr. of Science in Technology degree in 1997 from Helsinki University of Technology, Finland, where he is presently a permanent teaching research scientist at the Laboratory of Computer and Information Science. He is an author of about 20 journals and 90 conference papers on pattern recognition, statistical classification, and neural networks. His research interests are in content-based information retrieval and computer vision. Dr. Laaksonen is an IEEE senior member, a founding member of the SOM and LVQ Programming Teams and the PicSOM Development Group, and a member of the International Association of Pattern Recognition (IAPR) Technical Committee 3: Neural Networks and Machine Learning. Dr. Laaksonen has instructed 10 Master's Theses and five Doctoral Theses since 1999.

Timo Honkela is currently a chief scientist at Adaptive Informatics Research Center of Helsinki University of Technology (TKK). The unit is a center of excellence appointed by the Academy of Finland. In the center, Honkela is the head of one of the five research groups called Computational Cognitive Systems. Earlier he has served as a professor at the Laboratory of Computer and Information Science at TKK and as a professor at the Media Lab of University of Art and Design Helsinki. He is also a docent in three universities and has given hundreds of presentations and invited talks.

Honkela has conducted research on several areas related to knowledge engineering, cognitive modeling and natural language processing including a central role in the development of the Websom method for visual information retrieval and text mining based on the Kohonen self-organizing map algorithm. The current research interests include cognitive modeling, statistical machine translation, adaptive knowledge representation and reasoning based on continuous formal systems, and the underlying cognitive, linguistic and philosophical issues.

Honkela is a former long-term chairman of the Finnish Artificial Intelligence Society, the vice chair of Finnish Cognitive Linguistics Association and currently the chair of the IFIP working group on knowledge representation and reasoning (WG 12.1).

Matti Pöllä received his M.Sc. (Tech.) degree in electrical and communications engineering in 2005 from Helsinki University of Technology (TKK), Finland. Presently he works as a researcher at the Adaptive Informatics Research Centre at TKK. His current research interests include text data mining using bio-inspired algorithms.