Abstract—We argue that current data mining systems have failed to keep up with the rapid increase in multimedia data available to the public and that improved content-based indexing methods should be utilized to solve this problem. This entails bridging the semantic gap between the high-level semantic concepts of humans and the low-level statistical representations used by computer systems. Furthermore, the problem of multimodal fusion of information coming from different data types must be solved in a multimedia scenario. PicSOM, a content-based information retrieval system described in this paper, addresses both these problems in a consistent manner by introducing multimodal object hierarchies and utilizing the strong data mining and discovery properties of Self-Organizing Maps (SOMs). We demonstrate this by providing two real-world examples. First we show how semantic associations emerge from images and subjective evaluations of personal items collected at an art installation. Secondly we review our results from the TRECVID video retrieval evaluations, and how our methodology helped in finding semantically similar objects from a large multimodal database.

I. INTRODUCTION

In the modern world information is produced at an ever increasing rate. For example, in 2003 it was estimated that we collectively produce around 5 exabytes of new information in that year alone [1]. Also the amount of multimedia data is rapidly increasing, especially with the growing uptake of digital cameras and mobile phones which can acquire photos and videos. Furthermore, the social need to share, browse and find such data is demonstrated by the popularity of image and video sharing services on the web such as Flickr1 and YouTube2. Sadly the methods of retrieval and data mining provided by such services have not caught up with this development.

Multimedia browsing and retrieval systems are not very efficient, and people have difficulty in finding relevant information without the help of additional metadata, such as tags and textual descriptions. Manually added metadata is, however, highly subjective, slow and cumbersome to input. In many cases it is simply impossible to acquire metadata of a good quality due to the sheer amount of media data. To address this problem, indexing methods need to be improved so that data mining can be done based only on the contents of the multimedia objects themselves. This is the approach taken for example in content-based information retrieval (CBIR) where the objects are indexed by statistical features calculated purely from their contents.

In this paper we describe our content-based multimedia mining and retrieval framework PicSOM [2], [3], which uses the Self-Organizing Map (SOM) artificial neural network algorithm to index and retrieve objects from multimedia databases. Multimedia object hierarchies are used to facilitate an intelligent integration of different modalities (text, images, speech, videos) using a priori knowledge of object relationships and relevance provided by the user. We also demonstrate how semantic associations in the data emerge from studying the distribution of the feature vectors on the SOMs. Thus producing strong multimedia browsing and mining capabilities.

The rest of this text is organized as follows. First, in Section II we describe how content-based indexing of multimedia can be done. In Section III we look at the PicSOM system and how the properties of the Self-Organizing Maps can be utilized for data mining purposes. Then, in Section IV we look at how interesting conceptual correlations were found in the “Pockets Full of Memories” image database which has been augmented with semantic features. In Section V we look at high-level feature extraction results from the NIST TRECVID 2005 and 2006 video retrieval evaluations. Finally, in Section VI we present concluding remarks and have a look at future prospects.

II. CONTENT-BASED INDEXING OF MULTIMODAL DATA

We shall begin by looking at the two main challenges for content-based mining and indexing of multimodal data. The fundamental unsolved question for all content-based methods is the semantic gap and how to bridge it. Multimedia data mining in turn faces the problematic task of multimodal fusion.

A. The semantic gap

Three semantic levels of image queries, which can be applied more generally for any media information query, can be identified [4]: retrieval by primitive features, such as colors of an image or words or character combinations in a text, retrieval by logical features or semantic attributes, which can be more high-level descriptions such as “a picture of a car” or “a medical text”, and retrieval by abstract attributes such as “a picture depicting happiness” or “an ironic text”. This division
into semantic levels clearly illustrates the semantic gap, i.e. the large separation between the high-level semantic description used by humans and the low-level features available to computer systems [5]. Object descriptions in CBIR systems are mostly of semantic level 1, while humans naturally use levels 2 or 3 when describing their query target. For example in the case of a digital image, the automatically extracted low-level features only “see” local phenomena, such as colors, patterns and textures. A human analysis of the same image is without a doubt more holistic, describing for example the objects seen in the image and their relationships, not necessarily even noticing particular colors or textures.

The core of the problem lies in the difficulty of automatically extracting semantic content from a data object. The same semantic concept might have many totally different low-level representations, and also objects with similar low-level features might have distinct semantic meanings. This has been pointed out specifically in the domain of digital images [6]. Humans on the other hand are experts in recognizing semantic content. And from what we know, low-level information, such as specific words in a text or colors in an image, is only one part of the information used in the recognition process of humans. A lot of a priori knowledge is involved, such as previous experience of similar situations, cultural context and so on. Also the current situation and the context is important. Such information is hard to incorporate into a computer system.

One approach is to take advantage of known relationships and context of data items in the retrieval process. This is what we have tried to do in PicSOM using hierarchical object structures, which will be described in more detail in Section III-B. In a multimedia context this approach touches the problem of multimodal fusion when the given relationships consist of objects of different modalities.

B. Multimodal fusion

Information coming from different modalities, like text, audio or video, can have some relevance to the query target. This problem, sometimes called multimodal information fusion, is not trivial and there is no obvious general solution. The information from different modalities may often correlate in some manner which can be beneficial to the information retrieval, but in other cases the information from different sources may even contradict each other. A general multimodal information retrieval system needs to be able to handle both situations and combine the information from different modalities in a useful way.

There are several ways to implement multimodal fusion in information retrieval, and the problem can be addressed on different levels. On the feature level, feature vectors from different modalities can simply be concatenated, essentially creating a new combined multimodal feature. This is the approach taken for example in ImageRover [7], where visual and textual features are combined into one unified vector. Another strategy is to process the different modalities separately and then merge the retrieval results in some manner in the end.

Finally, the most promising technique is to implement cross-modality into to the information retrieval process itself. This usually results in associating objects across different modalities. This cross-modality can, for example, be based on context that has been stored beforehand, for example that a certain sound clip was recorded at the same time and place as a certain photograph was taken. Another way to implement cross-modality is to do it with statistical correlation, using for example latent semantic indexing (LSI) or cross-modal factor analysis [8] and Bayesian network models [9]. Most of the existing multimodal information retrieval systems are highly specialized to work with specific types of media, for example audio and video, or even very specific domains, e.g. videos of sporting events. The PicSOM system however is more general and integrates cross-modality into to the information retrieval process using object hierarchies which are not domain specific.

III. PICSOM FRAMEWORK

In this section we will describe the content-based multimedia mining and retrieval system PicSOM [2], [3] developed since 1998 at the Laboratory of Computer and Information Science of the Helsinki University of Technology. We will also look in more detail how the system can be utilized for multimedia data mining with semantic class concepts.

A. Self-Organizing Maps and relevance feedback

The unique approach used in PicSOM is to have several Self-Organizing Maps (SOMs) [10] in parallel to index and determine the similarity of data objects. These parallel SOMs have been trained in an unsupervised manner with separate data sets obtained by using different feature extraction algorithms on the same objects. So each SOM arranges the same objects differently, according to its particular multidimensional feature vectors, and forms a two-dimensional index of the objects in the database.

In the interactive mode, PicSOM uses the principles of query by example [11] and relevance feedback [12], [13]. This means that the system shows the user a set of database objects, which the user then indicates as relevant or non-relevant to the current query, i.e. close to or far from what he is looking for. Based on this relevance feedback information PicSOM modifies its internal state so that it will display better objects in the next round. This is done by increasing the influence of those SOMs that give the most valuable similarity evaluation according to the current relevance feedback information. The user thus becomes an integral part of the query process, which can be seen as a form of supervised learning, where the user steers the system by providing feedback. A CBIR system implementing relevance feedback essentially tries to learn the optimal correspondence between the high-level human concepts and the low-level internal features used in the system.

B. Object hierarchies and relevance sharing in PicSOM

PicSOM models relationships between objects in a database by grouping the objects involved into object hierarchies. We focus mostly on structures that can be represented in a hierarchical manner, and thus can be organized in an object tree in the database. In some situations an object can have
many parents, and then the structure is technically no longer
a tree but rather a graph. In this paper we focus only on
one parent and its (possibly many levels of) children at a
time, which motivates the use of the term “object tree”. The
object hierarchies can be multimodal, i.e. consist of objects
of different data types, and the object tree can be of any depth.
An example of a real-world multimedia object is shown in
Figure 1; a typical media-rich e-mail message with image,
audio and video attachments, where the different parts have
been highlighted and numbered. A hierarchical object tree
created from this example is shown in the same figure.
The properties of each object in the hierarchical tree, i.e. the
calculated feature vectors (using one or many different feature
extraction methods), can be considered to be characteristic not
only of the object itself, but to some extent also of its parents,
children and siblings in the tree structure. We call this idea
relevance sharing, which means that the relevance will be
transferred from the object to its parents, children and siblings.
For example, if a certain e-mail message is considered relevant
by the user in a query, its attachments will also get elevated
relevance values.

Figure 2 shows an illustration of relevance sharing in the
previous e-mail example. In the top part the user has indicated
a child video-clip object as relevant and the relevance goes
upwards to the parent. In the bottom part the relevance is
spread downwards from the parent to its children. This
process will result in the multimodal fusion of the relevance
information concerning different object types.

C. Low-level features

The PicSOM system implements a number of methods for
extracting different low-level features, such as statistical visual
features from images and image segments, and aural and
motion-based features from video clips. These features include
a set of MPEG-7 content descriptors [3], [14], and additionally
some non-standard descriptors for color, shape, texture and
motion [15].

1) Color: The following MPEG-7 color descriptors have been
used in the experiments described in this paper: Color
Layout, Color Structure, Dominant Color and Scalable Color.
In addition to these the Average Color feature based on the
CIE L*a*b* color space [16] and Color Moments based on
the three first central moments of the color distribution are
also used as color features.

2) Texture: MPEG-7’s Edge Histogram descriptor has been
used to describe the statistical texture in images. In another
non-standard description of a region’s texture the YIQ color
space Y-values of the region pixels are compared with the
values of their 8-neighbors. The feature vector describes the
statistics of the resulting distribution and is called Texture
Neighborhood. The Edge Histogram feature is the histogram
of four Sobel edge directions and it is different from the
MPEG-7 descriptor with the same name. The Edge Co-
ocurrence feature gives the co-occurrence matrix of four Sobel
edge directions.

3) Shape: Besides the MPEG-7 Region Shape, the shape
features include two non-standard descriptors. The first, Edge
Fourier, consists of the set of the Fourier descriptors for the
region contour [17]. In the feature vector we include a fixed
number of low-order Fourier expansion coefficients of the
contour, interpreted as an arc-length parameterized complex
function. The coefficients are normalized against affine image
transformations. The second non-standard shape descriptor,
Zernike moments, is formed from the Zernike moments [18]
of the overall binary region shape. This descriptor is made
invariant to affine transformations as well.

4) Audio: The Mel-scaled cepstral coefficients (MFCC)
were used as an aural feature called Mel Cepstrum. It is the
discrete cosine transform (DCT) applied to the logarithm of
the mel-scaled filter bank energies. Mel cepstrum is commonly
used for speech recognition, but can be used with other sounds
as well [19].

5) Motion: The Motion Activity feature standardized by
the MPEG-7 group was used for video clips. The descriptor
tries to capture the intuitive notion of “intensity” or “pace” of a video clip using the quantized standard deviation of motion vectors for classification. For example the scoring of a goal in a soccer game is usually perceived as fast paced, while a normal TV interview is slow.

Another set of motion features were based on the following non-MPEG-7 still image features: Average Color, Color Moments, Texture Neighborhood, Edge Histogram and Edge Co-occurrence. The frame-wise feature values are averaged over the frames contained within each one of five non-overlapping temporal video slices. In this way we get a final feature vector that describes the changes of the still image descriptors over time in different spatial areas of the video.

6) Text features: Unlike the other features, an inverted file instead of a SOM index was used for the text in the TRECVID experiments. The extension of the PicSOM system for using such indices in parallel with the SOMs was presented in [20]. In addition, an n-gram statistical feature made of character triplets and word histograms based on stemmed words was studied.

D. Data mining with SOMs

Multimedia databases always contain sets of objects semantically related to some high- or mid-level concept. In many cases, one will even have a smaller set of examples of such objects available. Feature vectors can then be extracted from these examples and mapped into the corresponding feature space. The feature vectors will then function as a non-parametric model for the concept. In a favorable situation objects belonging to the same semantic class will reside close to each other in the feature space, at least locally.

We can now map the feature vectors of the concept objects on a trained SOM by finding the best matching unit (BMU) for each vector. The ordering on the SOM map has emerged from unsupervised training using feature vectors from all the database objects. Thus the BMUs of the feature vectors of the semantic class form a discrete probability distribution over the two-dimensional SOM surface which characterizes the object class. Due to the topology preserving property of the SOM, we can then expect that objects close in the feature space will be close on the SOM map as well. This provides a good model for the semantic concept. Figure 3 visualizes this process, i.e. how the original very-high-dimensional pattern space is first projected to the feature space and then to the SOM grid.

When doing multimodal data mining, kernel smoothing is then performed on the sparse value fields on the SOMs. This process spreads the concept information and also to eases visual inspection of large SOMs. Some examples of such concept distributions will be displayed in Sections IV-C and V-C. These class-conditional and feature-specific distributions or class models can be considered as estimates of the true distributions of the semantic concepts in question on the discrete two-dimensional grids. Thereby, instead of modeling probability densities in the high-dimensional feature spaces, we are essentially performing kernel-based estimation of discrete class densities over the SOM grid. Depending on the variance of the smoothing function, the kernels will overlap and weight vectors close to each other will partially share each other's probability mass. The concepts may be evaluated on different scales by varying the kernel variance, thus generating a discrete scale space of the semantic concepts.

Qualitatively different distributions can be obtained from the same data by using different feature extraction techniques, leading to a collection of different two-dimensional class models. The mapping of a semantic class on a specific SOM gives, as a by product, insight into how well the corresponding feature can cluster the vectors of that class. Such modeling of mid-level semantic concepts is a very useful step in supporting high-level mining of multimedia data.

The most representative objects of a given semantic concept can be obtained by first locating the SOM units that have highest responses on the estimated class distribution. The objects mapped to these units can then be selected as the most representative ones for that concept. Combining the responses of multiple features can be performed similarly as in the retrieval stage, after which we obtain the overall most representative objects of a specific concept regarding all the used features. This property is also beneficial for data mining, where we can find semantically similar objects by projecting examples on SOM feature maps, as will be demonstrated in Section IV-E.

IV. EMERGENCE OF SEMANTICS FROM IMAGE AND AUXILIARY DATA

An interactive art installation “Pockets Full of Memories” was on display in the Centre Pompidou National Museum of Modern Art in Paris, France in 2001 [21]. Visitors were encouraged to contribute with their own personal objects, which were photographed, adding keywords and quantified semantic features such as age or hardness. This data was projected in real-time onto a Self-Organizing Map (SOM) in the gallery as a “wall of objects”, i.e. a two-dimensional map of the contributed items. The distribution of the objects depended on both their visual, textual and semantic features, placing items with similar attributes near each other. The self-organized ordering emerged with time, created by a large number of local interactions on the map, rather than specified by hand.

In later research [22] we gathered this intriguing data of over 3300 objects into a multimodal database by incorporating the relationships between the images and the auxiliary textual and semantic information as object hierarchies in PicSOM. To be able to focus more accurately on the visual properties of the
objects, automatic image segmentation was used to separate the centrally-located objects from the nearly constant-colored backgrounds [22]. Several Self-Organizing Maps were then trained with different sets of statistical features calculated from the images and auxiliary data.

### A. Value features

For each object, eight values were given by its owner, quantifying its properties with regard to specific attribute pairs. The property pairs were old—new, soft—hard, natural—synthetic, disposable—long use, personal—nonpersonal, fashionable—not fashionable, useful—useless, and functional—symbolic. We have scaled the resulting values to the range $[-1, 1]$ and collected them as components of an 8-dimensional values vector.

### B. Creation of SOM maps from low-level data

Fig. 4 shows two SOMs created with different features. Each SOM unit is represented by a visual label, which is the image that is closest to the SOM unit’s model vector in feature space. The SOM surface in Fig. 4(a) is organized according to the MPEG-7 Edge Histogram feature. Objects with similar shapes and orientations form clusters. Within the clusters the object shapes change continuously, thus retaining the topographical ordering of the shape feature space.

The SOM of Fig. 4(b) is organized according to the values feature. Some clear clusters are formed, for example to the left of the center we find many plush toys and teddy bears with values that reveal softness and very personal items. Additionally, for example, in the upper right corner there are many watches and mobile phones with values indicating items that are useful, functional, new and synthetic.

These observations (and many others in our studies) clearly demonstrate the usefulness of the SOM model. The input data is of a high dimensionality, making it very hard to analyze, let alone to understand and visualize it in any real sense. In contrast, the two-dimensional SOM maps of different features provide intuitive views of the data, revealing meaningful relations within the distribution. This provides strong data mining capabilities for the PicSOM system, especially in a multimodal context.

### C. Correlations of visual maps and semantic concepts

From the different attribute quantifications given by the owners of the objects we generated a set of semantic classes. The value ranges $[-1, 1]$ were divided into three equal parts, where the low-end and high-end parts correspond to the semantic extremes. For example, for the hardness property, objects with values in the range $[-1, -\frac{1}{3}]$ belong to the semantic class soft and those with values in the range $[\frac{1}{3}, 1]$ belong to the class hard.

In Fig. 5 we see the distributions of three different semantic classes mapped onto the MPEG-7 Edge Histogram SOM: soft, natural and fashionable. 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 class on this feature SOM. There seems to be a large set of objects that are both soft and natural, roughly in the middle of the MPEG-7 Edge Histogram SOM. Visual inspection of the SOM labels in Fig. 4(a) reveal 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.

The observed correlation is intuitively easy to understand...
as many natural objects are also soft. Besides, MPEG-7 Edge Histogram, being a texture feature sensitive to local edges in the image, should be good at discriminating soft edges from hard ones. The distribution of fashionable items shows an example where MPEG-7 Edge Histogram does not discriminate well as the distribution is relatively disperse, signifying that the very abstract concept “fashionable” has no correlation with the edge shape of the image.

In Fig. 6 the class disposable has been mapped on three different SOMs: MPEG-7 Edge Histogram, Zernike moment and color moment. All three maps show good or very good clustering, with the disposable objects cleanly mapped into contiguous areas of the SOMs. Upon inspection of the visual labels on the SOM of the rotation-invariant Zernike moment shape descriptor, we could notice that in the upper-right corner of the map, where many disposable items have clustered, there are mostly rectangular objects, like candy boxes, and pieces of paper, like bus tickets.

**D. Matching of English and French words**

One interesting aspect of the keywords in the “Pockets Full of Memories” collection is that they consist of both English and French words. One typical class of objects in the database is referred to as pen or pencil in English and as crayon in French. From the Zernike moment SOM one could see that the areas with the densest object distributions in all these three are mostly located in the bottom part of the SOM surface. This result supports our hypothesis that meaningful relationships between both intra-lingual synonyms and inter-lingual word translations can be mined from a multimodal database.

**E. Finding best representatives of object classes**

As explained in Section III-D, it can be instructive to seek the most representative objects of a given semantic class by combining the responses of the class distributions on several feature maps. As an example we have looked at a small subset of objects that were labeled as watches, clocks or as montre in French. These objects were then mapped to a set of SOMs trained only on visual features, and the responses combined into a single ordered list of most representative objects.

In Fig. 7 one can see the eight most typical objects, of course excluding those in the example set. Clearly, even by using just visual features, and not taking advantage of the semantic values features or keywords, representative images can be found. In the shown set only one is not a watch, while the other images display watches of different types and orientations. Again, this is very useful for data mining, for example for finding similar objects in an unannotated database.

**V. HIGH-LEVEL SEMANTIC FEATURE DETECTION IN VIDEO**

The PicSOM system has been evaluated in the TREC Video Retrieval Workshop series in 2005 [23] and 2006 [24]. The test collection consists of news and other TV video clips in English, Chinese and Arabic with additional textual data and semantic concept classes. TRECVID’s high-level feature detection task entails detecting high-level semantic classes from the database (test set) using some given examples (development set). We utilized PicSOM’s multimodal data mining methodology to extract such classes successfully. The results of the semantic feature detection were then utilized in the content-based video search task as well.

**A. TRECVID video data**

The TRECVID data contains about 790 news videos divided into a total of almost 100 000 video clips. Each video clip has one or many keyframes, which are representative still images taken from the video. Also the audio channel was extracted as a child of the video clip. TRECVID provided textual data acquired by using automatic speech recognition software and machine translation from Chinese (Mandarin) and Arabic to English. Some video clips also had closed-captioning text which could be utilized in a similar manner.

In the PicSOM system the videos and the parts extracted from these were arranged as hierarchical trees as exemplified in Fig. 8, with the main video clip as the parent object and the different parts as its children. In this way the relevance assessments can be transferred between the related objects by the PicSOM algorithm as described in Section III-B. Different
features were extracted from each media type, and Self-Organizing Maps were trained from these as is shown with some examples in the figure.

The LSCOM-Lite [25] set of 39 semantic classes were provided with the TRECVID development data. These are each a set of video clips in the training set that belong to a given semantic class, for example videos depicting “an exterior of a building”.

The set of used features was selected for each concept separately. For this purpose, we applied the well-known Sequential Forward Selection (SFS) scheme [26]. As the optimization criterion we used the average precision at 2000 returned items with two-fold cross validation on the development set. The selection process typically resulted in 4–7 parallel features.

B. Concept detection results

The results of our experiments in TRECVID 2005 and 2006 as compared to the evaluated systems were quite satisfactory (clearly above average in most cases) and are reported in more detail in our TRECVID reports [23], [24]. Semantic classes which can be expected to be visually homogenous, such as maps, waterscape/waterfront, sports (e.g. green football fields), weather (weather maps and graphical bulletins), worked the best. Also for some semantic classes the text features played a strong role, for example the class police/security having prominent words, such as “armed”, “forces” and “resistance”. The worst results were in more ambiguous classes, such as prisoner, depicting “a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.”. However in this case the results were overall bad in the competing systems as well.

C. Feature map correlations

Once again one can find interesting correlations by looking at the LSCOM-Lite semantic classes mapped to the feature map grids. Examples from the maps class in TRECVID 2005 were projected to a SOM generated from the MPEG-7 Color Structure data of the keyframe images (see Fig. 9). In the same figure, to the right, is the distribution of the sports class from TRECVID 2006 mapped on the MPEG-7 Edge Histogram SOM. Color Structure shows two main clusters, with some smaller ones in between, indicating good correlation with the semantic class. The MPEG-7 Edge Histogram map shows two very sharp clusters, indicating that this is a good feature for finding keyframes of sports events, which also explains the good concept detection results we got in this case.

In Fig. 10 we have zoomed in on a 16×16 area of the upper cluster in the MPEG-7 Edge Histogram SOM. The map units are once again represented by their visual labels. In this figure we can see a very high concentration of correct images, such as scenes from soccer matches, tennis courts, a golf court etc.

These examples demonstrate the usefulness of the PicSOM approach for very large sets of complex multimedia data. The SOM-based methodology manages to overcome this complexity in data and even takes advantage of information from the different modalities. The results were not always as clear-cut as in the case with the “Pockets Full of Memories” experiments, where the data was less complex. However, the distributions on the feature maps were still useful for data mining purposes, such as visual browsing, knowledge discovery and retrieval of similar objects in the database.
VI. CONCLUSIONS AND FUTURE VIEWS

We have shown how object hierarchies combined with the mapping properties of the Self-Organizing Map can be used in solving the problems of the semantic gap and multimodal fusion in multimedia data mining. The hierarchical structuring combines the relevance values from related objects of different modalities, while the kernel smoothing of SOM-mapped class distributions provides an elegant way of emphasizing discriminative features while attenuating others.

In the experiments we demonstrated how the SOM distributions can be used for knowledge discovery and data mining, i.e. finding interesting associations and semantic relations in a multimedia database. The TRECVID evaluations also show how these properties can be used in large-scale real-world retrieval tasks with very promising results. The topology-preserving property of the SOM mapping provides a uniform formalism for all statistical features that can be extracted from any media domain. This formalism in turn suits well for both interactive browsing and automatic content-based information retrieval and data mining.

The development of data mining and information retrieval systems for multimodal data will inevitably benefit from mimicking aspects of the information processes in the human brain. One such aspect is the a priori knowledge, such as previous experiences and familiar contexts, that humans implicitly utilize when searching for relevant information. In our initial experiments on incorporating analogous functionality in a retrieval system, the relevance history of each database object from past queries was stored. The information was then later used with SOMs as a semantic similarity measure between the objects [27]. The results were promising, inspiring us to further explore this avenue, especially in applying it to multimedia retrieval scenarios where huge amounts of user action history are available for data mining and knowledge discovery.

ACKNOWLEDGMENTS

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.

REFERENCES