Analysis of Semantic Information Available in an Image Collection Augmented with Auxiliary Data*

Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Timo Honkela

Adaptive Informatics Research Centre,
Helsinki University of Technology
P.O.BOX 5400, FI-02015 TKK, Finland
{mats.sjoberg,ville.viitaniemi,jorma.laaksonen,timo.honkela}@hut.fi

Abstract. An art installation was on display in the Centre Pompidou National Museum of Modern Art in Paris, where visitors could contribute with their own personal objects, adding keyword descriptions and quantified semantic features such as *age* or *hardness*. The data was projected in real-time onto a Self-Organizing Map (SOM) which was shown in the gallery. In this paper we analyze the same data by extracting visual features from the images and organize the image collection with multiple SOMs. We show how this mapping facilitates the emergence of semantic associations between visual, textual and metadata modalities by studying the distributions of the different feature vectors on the SOMs.

1 Introduction

In this paper we conduct an analysis on an image collection that has been augmented with descriptive features and keywords. The image collection consists of a database collected during an interactive museum installation. This installation, "Pockets Full of Memories" was on display in the Centre Pompidou National Museum of Modern Art, Paris, France from April 10 to September 3, 2001 [1]. The visitors contributed over 3300 objects digitally scanning and describing them. This information was stored in a database and organized by the Self-Organizing Map (SOM) algorithm [2] that positioned objects of similar descriptions near each other in a two-dimensional map. The map of objects was updated online and projected on a wall in the gallery [1].

The self-organizing algorithm was the basic method used to create the "wall of objects". The SOM organized the input items into an ordered display, a planar map. In this exhibition, the input features consisted of attributes and keywords given by the exhibition visitors together with the objects. The attribute values and keywords were transformed into numerical form that could

* This work was 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.*

Please use the following format when citing this chapter:
Sjoberg, Mats, Viitaniemi, Ville, Laaksonen, Jorma, Honkela, Timo, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 600–608

serve as inputs to the SOM algorithm. Close to each item on the map there were items that had been given similar attribute values, items that had similar keywords, or both. Thus, all the items with a particular keyword were not necessarily next to each other if the other features varied. The ordering of the final map is a consequence of all the inputs. The phenomenon is called emergence: the order of the objects is not determined beforehand, but emerges through the audience contributions. The classification system is merely created through the large number of local interactions on the map, rather than specified by hand.

In the exhibition, even if the visual qualities of two different images were very similar, their respective owners may have evaluated their attributes very differently based on their subjective points of view. In this paper, we take a step further: we aim at correlating the visual features from the images themselves and the given metadata in order to extract semantic information. In the following, we describe the image segmentation process, how class distributions are analyzed and what features were used. After that, we present the experiment results and draw some conclusions.

2 Methods for Finding Semantic Associations

In this section, we describe the methodological steps needed in finding semantic associations, i.e. associations between the automatically computed multimodal features and the separate metadata. We outline the method for image segmentation and then provide details of the semantic association process. Finally we shortly present the different feature extraction methods used. The described processing stages have been implemented in our PicSOM[2] content-based image retrieval (CBIR) system [3].

2.1 Segmentation

The purpose of image segmentation is to partition images into segments that can be analyzed separately. Individual segments are often easier to analyze and interpret than the image as a whole, e.g. when the segments correspond to distinct objects in the physical world. Analysis of the individual segments also facilitates the interpretation of the whole image in terms of its constituent segments and their relationships. In the current application, the image segmentation algorithm is used to separate the objects of interest from image backgrounds. The subsequent image analysis is thus more accurately focused on the properties of the relevant objects, not on the properties of the different backgrounds against which the objects happen to be scanned.

The segmentation proceeds in two steps. First, a generic color image segmentation partitions the images into eight regions. The generic segmentation method first radically oversegments the image. This is achieved by applying

[2] `http://www.cis.hut.fi/picsom`

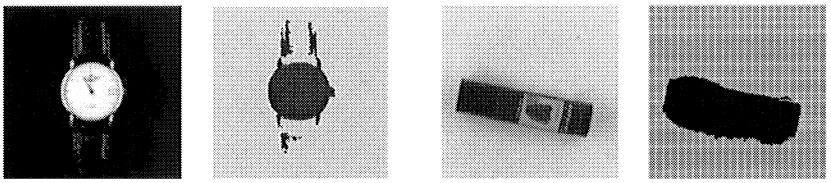

Fig. 1. Two images from the collection and their segmentations. In the left "watch" image the background is dark whereas it is light in the right "lipsalve" image.

the k-means algorithm [4] to the CIE L*a*b* color space [5] representations of the image pixels. After this process, the resulting regions are merged until eight regions are left. The merging criterion is based on the combination of the average color of the regions and a multi-scale edge strength measure.

The second step of segmentation exploits the special image structure with centrally located objects against nearly constant-colored backgrounds. Here a straightforward set of inference rules is applied to tag some of the regions as background. The remaining regions are interpreted to represent the objects of interest. The tagging begins by identifying the likely background color by considering those regions that are homogenous in color and form a significant part of the peripheral image area. Having the background color identified, similarly colored homogeneous regions are tagged as background if they are adjacent to either the image borders or already tagged background regions.

The image segmentation results are good in most cases. Some exceptions result from the background color seamlessly penetrating the object. An example can be seen in Fig. 1 where the dark background color can not be separated from the dark areas of the band of the wrist watch. Also objects that stand out poorly from the background and disconnected objects (e.g. transparent plastic bags and thin necklaces) often result in deficient segmentations. Fortunately, even when the segmentation results are strictly speaking erroneous, they are still often adequate for the subsequent image analysis as similar objects are segmented in a qualitatively similar manner. For instance, as the wrist bands of watches are often regarded partly as background, the image analysis concentrates on the clock-faces for most wrist watches. For the current experiments the segmentations are at least adequate in focusing the image analysis to the relevant parts of the images. It does not seem likely that the experiment results would be significantly compromised by faulty segmentations.

2.2 Analysis of Class Distributions

In a typical application of the Self-Organizing Map (SOM) in data mining, visualization or information retrieval, a SOM is trained in a fully unsupervised mode, using a large batch of training data vectors. Yet, the data often contains some semantically related object groupings or classes, and sets of objects belonging to such user-defined classes are known. Such a set of vectors can be

mapped on a trained SOM by finding the best matching unit for each vector in the set. These "hits" over the map units form a discrete probability distribution over the two-dimensional SOM surface which characterizes the object class. Qualitatively different distributions can be obtained from the same data by using different feature extraction techniques, leading to different numerical representations of the data items.

The mapping of a semantic class on a specific SOM gives insight into how well the corresponding feature can cluster the vectors of that class. The sparse value fields on the maps are low-pass filtered to spread the information. This also helps visual inspection as the marked areas become larger and more uniform.

In our study, we map the user-given keywords on the SOM surfaces along with the images themselves to see which keywords, numerical attributes and visual features correlate the best. This in turn will reveal the semantic characteristics of the objects from their visual appearances and associated metadata.

2.3 Extraction of Multimodal Features

We used in total three different image features, the user-provided attribute values and keywords. A 64×64-sized SOM was trained for each of these five features. In the following, the features are described in some detail.

Visual Features: The visual features were 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 standard MPEG-7 [6] *Edge Histogram* feature measures the distribution of edge directions within the object and thus describes the texture and local shapes. *Zernike moments* [7] describe the overall shape of the object's segmentation mask. The *color moment* feature characterizes the color distribution within the object with the three first central moments.

Value Features: For each image, eight values were given by the owner of the object, quantifying its properties with regard to specific attributes as shown in Fig. 2. The property 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. We have scaled the resulting values to the range $[-1, 1]$ and collected them as components of an 8-dimensional *values* vector.

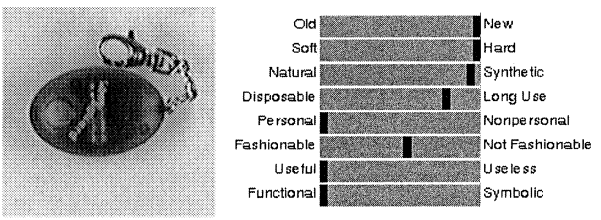

Fig. 2. An "attackalarm" and the property values given by its owner.

Keyword Features: The keyword text data was primarily used as such to annotate the images and to define per-keyword image classes. In addition, we also created an *n-gram* statistical feature made of character triplets extracted from the keywords.

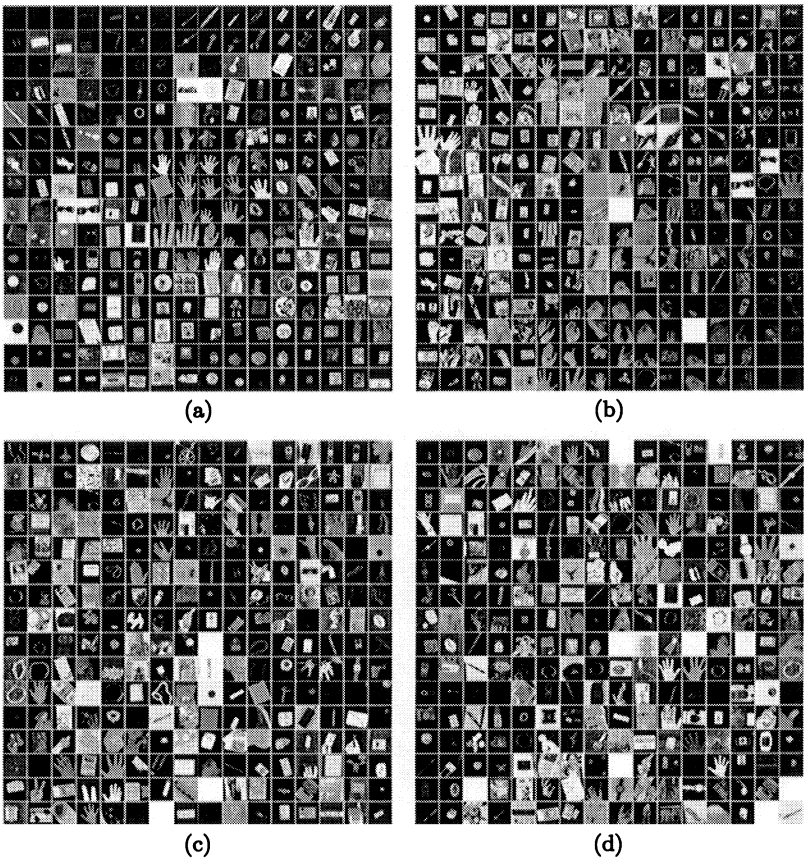

Fig. 3. Organization of images on SOM surfaces. The SOMs in the upper row have been trained with visual features MPEG-7 *Edge Histogram* (a) and *color moments* (b). In the lower row, the SOM (c) is based on the *values* feature and the SOM (d) on the *n-gram* feature.

3 Experiments and Results

3.1 Creation of Visually Organized Maps

Fig. 3 shows four SOMs created with different features. Each SOM unit is represented by a visual label which is the most similar image of the database in that feature space. The SOM surface in Fig. 3(a) is organized according to the *Edge Histogram* feature. Objects with similar shapes and orientations form clusters. Within the clusters the object shapes change continuously, thus retaining to topographical ordering of the shape feature space. In subfigure (b) another visual feature, *color moments*, has been used, and a color-based organization is evident.

The SOM of Fig. 3(c) 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. The organization produced by the *n-gram* feature in the bottom-right SOM (d) is not that evident, but one can see that e.g. shoes appear in nearby locations.

3.2 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 $[-1, 1]$ are 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. 4 we have the distributions of three different semantic classes mapped onto the *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* 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 in Fig. 3(a) indicate 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, the *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 *Edge Histogram* does not discriminate well as the distribution is relatively disperse.

In Fig. 5 the class *disposable* has been mapped on three different SOMs: *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 (not shown here), we notice that the positively-marked upper-right corner of the map shows mostly rectangular objects like candy boxes and pieces of paper like bus tickets. A similar analysis of the visual labels of the *color moments* SOM in Fig. 3(b) shows mostly items with white or light colors.

3.3 Matching of English and French Words

One interesting aspect of the keyword collection is that it consists of both English and French words. One typical class of objects in the database is referred to as *pen* or *pencil* in English and *crayon* in French. In the collection, there are 11 images keyworded as *pen*, 9 as *pencil* and 33 as *crayon*. The distributions of these sets on the *Zernike moment* SOM are plotted in Fig. 6. One can see that the areas of densest object distributions are mostly located in the bottom part of the SOM surface. This result supports our working hypothesis that meaningful relationships between both intra-lingual synonyms and inter-lingual word translations can emerge based on auxiliary, non-textual data modalities.

4 Conclusions

The self-organization of objects is an effective method for detecting inherent structures, patterns and clusters in complex collections of data. With it, one is able to automatically find different kinds of associations between the items when different data modalities and features are considered in parallel. We are ascertained that these phenomena will prove to be useful in semantic analysis of multimodal data collections. Potential application areas include emergent semantic representations useful in the contexts of the semantic web, machine translation, visual data mining, and creation of pictorial dictionaries.

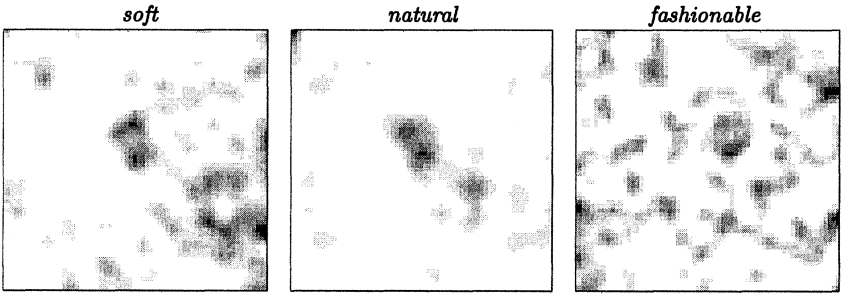

Fig. 4. The class models of *soft*, *natural* and *disposable* on the *Edge Histogram* SOM.

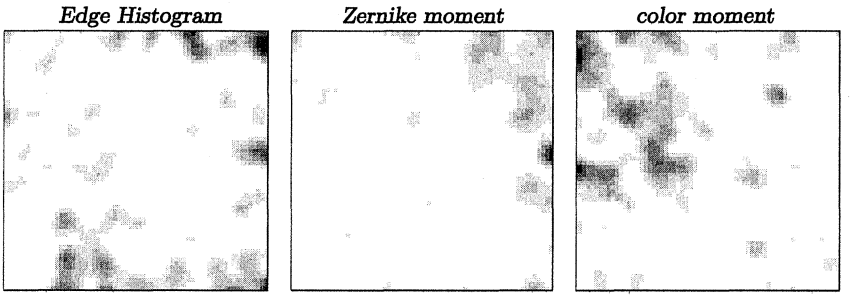

Fig. 5. The class models of *disposable* objects on the MPEG-7 *Edge Histogram*, *Zernike moment* and *color moment* SOMs.

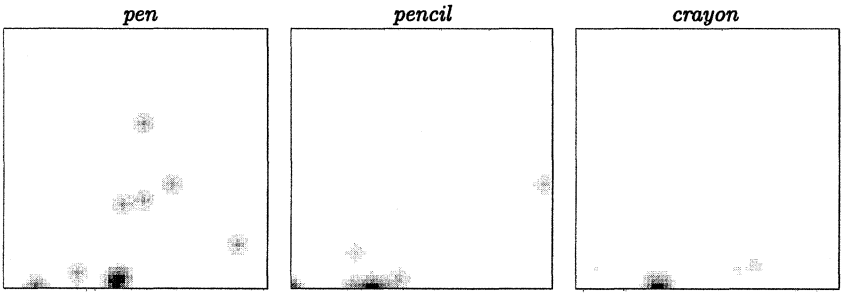

Fig. 6. Distributions of *pen*, *pencil* and *crayon* classes on the *Zernike moment* SOM.

The described procedure also demonstrates the use of automatic image segmentation to focus the processing on relevant parts of the images. In the light of the presented example, it is evident that segmentation does not need to be exactly correct in order to be helpful in processing visually-grounded semantic information. In general, autonomous machine learning from large multimodal databases in a statistical manner seems to provide an efficient and robust method for modeling grounded semantic relationships.

References

1. Legrady, G., Honkela, T.: Pockets full of memories: an interactive museum installation. Visual Communication **1** (2002) 163–169
2. Kohonen, T.: Self-Organizing Maps. Third edn. Volume 30 of Springer Series in Information Sciences. Springer-Verlag (2001)
3. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing **13** (2002) 841–853
4. Schalkoff, R.J.: Pattern Recognition: Statistical, Structural and Neural Approaches. John Wiley & Sons, Ltd. (1992)

5. CIE: Supplement No. 2 to CIE publication No. 15 Colorimetry (E-1.3.1) 1971: Official recommendations on uniform color spaces, color-difference equations, and metric color terms (1976)
6. ISO/IEC: Information technology - Multimedia content description interface - Part 3: Visual (2002) 15938-3:2002(E).
7. Khotanzad, A., Hong, Y.H.: Invariant image recognition by Zernike moments. IEEE Transaction on Pattern Analysis and Machine Intelligence **12** (1990) 489–497