Designing Semantic Image Retrieval Systems

Dipl.-Inf. Marcel Spehr

18.12.2014

Visual images constitute one of the most common formats of electronic data that need to be efficiently accessed. In the context of this thesis, access can be understood as the retrieval of a specific subset of images from a larger set. The best method to identify this subset is problem dependent.

At the heart of this thesis lies the quest for good design principles for the creation of semantic image retrieval systems. To state the problem more concisely:

How does one find the most suitable toolchain to solve a semantic image retrieval task?

The thesis is meant as a construction manual for prospective system designers and guide them in the composition of suitable techniques. In order to accomplish this task, the problem was broken down into its main modules: 1) *semantics*, 2) *visual features*, and 3) *pattern recognition*. Modules 2) and 3) were discussed regarding their relation to 1) in detail. From this, selection as well as construction strategies were laid out. Thus, it is the first work that treats the image retrieval problem in its entirety from a stringent semantic perspective. Adherence to the proposed principles will allow a well-educated computer scientist to construct a retrieval set-up that suits specific semantic needs.

Image retrieval systems serve one purpose. Guess the user's query intention as good as possible and quickly deliver images to his or her satisfaction. Even though it is straightforward to state this, it is in most cases still unclear how to solve this task. This thesis makes and furthermore substantiates the following claims:

1. The vague query *"deliver semantically similar images to a given image"* is ill-posed and in general without further constraints unsolvable.
2. If the semantic and the image domain are reduced to a very specific subset, useful contributions can be achieved.

The contributions of this thesis are

1. An extensive discussion on
   a) how to restrict the aforementioned domains by introducing the concept of a *Semantic Coordinate System* (inspired by linguistic theory) and
   b) appropriate algorithms to process visual information. A taxonomy of feature extraction algorithms is introduced, that enables application driven feature selection and construction.

2. A novel image feature is developed from scratch and its usefulness demonstrated. This work was published as [SGF11].

3. A case study for image retrieval according to their painting style is done. Problem specific image features are selected and constructed. Several machine-learning algorithms are employed to verify their effectiveness. This work was published as [SWF09].

4. A semi-automatic procedure for label-free image searches in high dimensional feature spaces is developed, which also serves useful purposes for feature optimization. This work was published as [SHHG13].

5. A web- and cloud-based software system is introduced that allows automatic parameter evaluation for the design of a retrieval system. This work was published as [SGG15].

This article will give a concise walk-through the complete thesis and shall serve—along with the graph in figure 1.—to outline the central ideas that underlie this work. This section is structured in six paragraphs that correspond to the six contributions listed above.

1.a) Making Image Semantics and Semantic Similarity Explicit Image data accumulate in many application areas and diverse questions are posed with respect to them. Depending on the context that generates the images, domain specific visual characteristics are exhibited. These characteristics in turn determine solution strategies for tasks like semantic image retrieval.

This thesis presents the first linguistically, scientifically sound analysis of visual semantics based on the set of semantic primitives of the *Natural Semantic Metalanguage*. This leads to the concepts of visual variables and a *semantic coordinate system* that allow images to have explicit semantic coordinates. Thereby, images become comparable regarding depicted semantic content and the semantic aspect of an image-based query can be made explicit. The classification of the semantic variables according to three classes—*nominal*, *ordinal*, and *quantitative*—then determines what kind of statements about similarity between images can be made at all. More precisely, it determines whether statements regarding the similarities between triples of images are possible. It becomes clear, that when semantic variables are combined, retrieval requests cannot be unambiguously served.

The classification according to this semantic scheme determines suitable tools to solve the query task, e.g. *classification* for *nominal* and *regression algorithms* for *ordinal* or *quantitative* semantic variable recognition. The mapping of images to coordinates in a semantic space leads to the concept of a *target similarity* that underlies each semantic retrieval request. Three strategies—$\Phi^* d^*$, $\Phi^* d^{\bigstar}$, and $\Phi^{\bigstar} d^*$—are introduced that serve to approximate this target similarity by computational means. These strategies are implemented in a series of instructive case studies that extend over the whole thesis.

1.b) Visual Image Feature Selection and Construction Once the type of the semantic variable is determined, one can engage in the problem of approximating the target similarity. There are three strategies available, one can employ: $\Phi^* d^*$ (select feature space and distance), $\Phi^* d^{\bigstar}$ (select feature space and construct distance), and $\Phi^{\bigstar} d^*$ (construct feature space and select distance). A comprehensive toy example details how each strategy may be implemented.

All strategies encompass the selection respectively the construction of appropriate feature spaces. Pixel-wise basic symbols represented by raw colour or brightness values are usually no elements of meaning. Hence, *semantically opaque* images must be mapped into a feature space. This process is also called feature extraction. There exists an overwhelming amount of literature on feature extraction algorithms. To give both novices and experts a guideline, a taxonomy of techniques is presented that arranges candidate algorithms along three degrees of freedom. The categorization of well-established techniques along three methodical axes provides for a modular and well-sorted development kit for the selection and construction of problem specific algorithms. According to 1) *what* kinds of visual information are characteristic for the image classes at hand, 2) *where* the visual support in the image domain resides, and 3) the adequate *granularity* of the feature space that is to be retained, established methods can be selected and novel methods constructed.

2. Custom Image Feature Development: The *Sum-of-Superellipses* model If none of the available features seems to suffice the problem, strategy $\Phi^{\bigstar}d^*$ must be employed. The process of designing a custom-built feature from scratch is presented. This includes: the observation of commonalities and differences of images depicting values of a specific semantic variable, the formation of a mathematical hypothesis for harvesting these, and an experimental study to substantiate the hypothesis using a supervised classification technique.

The *Sum-of-Superellipses* is a novel empirically derived low parameter model for amplitude spectra of natural images. Its superiority to older models lies in its ability to represent characteristic properties in a closed formula with a small set of intuitive parameters. It also proved adequate for describing local patches of natural images and thereby reproducing their scale invariance property. It is even possible to reconstruct images using only 6 fitted parameter values to a recognizable level.

The described development of the model serves as an example on how the strategy $\Phi^{\bigstar}d^*$ is implemented to approximate a target similarity. The similarity discussed in the thesis addressed the resemblance between images depicting the same scene category (e.g. the amplitude spectra of images depicting coasts are similar to each other and differ from amplitude spectra of forest images). The dimensions of the feature space Φ are the model parameters. The mapping ϕ is constituted by the fitting procedure. The *SoS* can thus be used as a module in the proposed design process to construct feature extraction algorithms (e.g. one could adapt the spatial sampling process to salient points instead of a regular grid, or multiple corresponding model parameters could be quantized and summarized in histograms).

3. Case Study: Retrieval of Paintings Two scenarios regarding the retrieval of paintings are discussed. First, a method for automatically clustering images according to the overall visual appearance or *look*, much as untrained observers do, is presented. For this, a target similarity has to be approximated without training data. Because the appearance of paintings is complex and spans many aspects ranging from colour content to semantics a large number of feature spaces is constructed (cf. strategy $\Phi^{\bigstar}d^*$) and selected (cf. strategy Φ^*d^*), each of which is insufficient to capture appearance on its own. However, taken together they can be used to parse a database of images into visually meaningful groups. It is shown that such an *appearance-based* clustering is affected by—yet is not the same thing as a clustering based on—distinct artistic periods or styles. Much like human observers, the system confounds style and content when assessing the similarity of two images. Some of the clusters have a very intuitive visual quality (e.g. dark portraits or hazy landscapes). However, other clusters are more heterogeneous, containing images whose principal shared attribute is not belonging to one of the other well-defined classes. Interestingly, in previous experiments, some

subjects reported forming *miscellaneous* groups to classify images that did not belong with the others. Thus, these failures may reflect a key aspect of the data.

Second, the ability of the features for classifying paintings into the historical art periods—both in an unsupervised and a supervised fashion—is tested. Performance in both cases is well above chance with clear variations in discrimination across art periods: paintings from the Gothic period, for example, are easy to separate (a result which agrees well with the perceptual data found in user studies). In contrast, Baroque and Romanticism are hard to recognize.

4. Feature Optimization Without Prior Semantic Image Knowledge: The *Refkep* Image Browser If no prior knowledge about the semantic relation between images is given by labelled training data, a human must be taken into the loop. For this, a novel semi-automatic approach to solve the image retrieval task in the presence of abstract features and an unspecific target similarity is presented.

Interactive feedback is used to estimate the layout of semantic classes in (possibly many) feature spaces. The proposed system even works when the target class cannot be made explicit. It suffices, that the user has a mental model that enables him to label images as class member. In addition to its image browsing capabilities, *Refkep* has additional benefits. Using the user feedback, it refines a mathematical model that can afterwards be used to select appropriate feature spaces (cf. strategy $\Phi^* d^*$) and even to construct distance measures (cf. strategy $\Phi^* d^{\bigstar}$) for further retrieval purposes.

The system has an extremely simple interface which yet is powerful enough to convey meaning about the underlying distribution in feature space of the images and thus helps to bridge the *semantic gap*. The described modular pipeline is easily adaptable to diverse application domains. It distinguishes itself by allowing the definition of almost arbitrary similarity measures from which a metric feature space can be deduced. A multi-resolution hierarchy on the data points is employed to achieve scalability and interactivity for the browsing procedure. A novel method incorporates user interaction data as *Relevance Feedback* into the computation of the *Mixture-of-Gaussian* model that describes the target class. This model can be henceforth used to 1) generate new suggestions for the user, 2) select suitable features, and 3) construct a similarity measure that approximates the user intended target similarity.

A case study, in which works of Giuseppe Arcimboldo had to be discovered among 20.000 paintings of other artists, proved the system to be effective.

5. Automatic Image Feature Benchmarking using *Wifbs* Ideally, prior knowledge in the form of semantically labelled images is available and can be transferred to novel images. In that case, the target similarity approximation task essentially becomes one of intensive benchmarking and selecting the best distances as well as feature extraction, data processing and classification algorithms (cf. strategy $\Phi^* d^*$). The thesis presents a system to efficiently select and parametrize the algorithmic means to best achieve this objective. Each module of the solution toolchain can be thoroughly evaluated. Thereby, *Wifbs* complements *Refkep* in enabling system designers to streamline a toolchain to specific needs. *Wifbs* relieves the user from concerning himself with technical details like process parallelization, data and parameter administration or result presentation. Its easy extensibility ensures maximal flexibility for a wide range of applications.

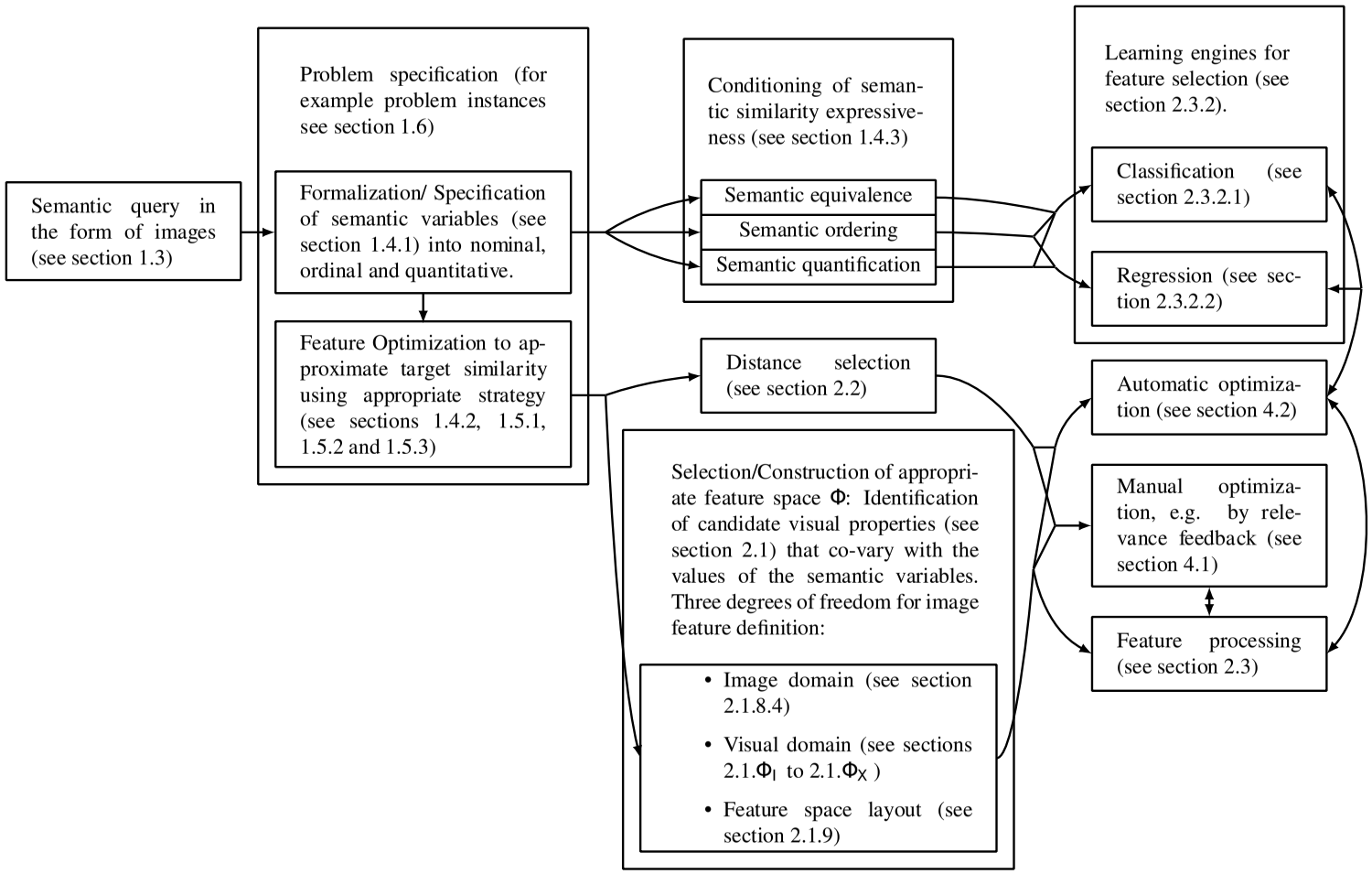

Figure 1. : The figure above illustrates the workflow for the design process of a retrieval system. It also shows this thesis' chapter structure. Given a vague semantic query, first the intended semantic attributes must be formalized and specified into i) nominal, ii) ordinal or iii) quantifiable. i) allows only the usage of classification engines. ii) and iii) additionally allow the usage of regression engines. Once the intended semantic variable is defined, one has to identify candidate distances and visual features. The latter can vary in examined visual domain, distribution in image domain and feature space layout. Candidates must afterwards be evaluated a) if training data is available automatically for optimal classification and regression results or b) manually if not. A feature processing can be included in the optimization loop.