# Hybrid Models for Combination of Visual and Textual Features in Context-based Image Retrieval

*Leszek Kaliciak*

A thesis submitted in partial fulfilment
of the requirements of
The Robert Gordon University
for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with
The University of Aberdeen and The Open University

July 2013

## Abstract

Visual Information Retrieval poses a challenge to intelligent information search systems. This is due to the semantic gap, the difference between human perception (information needs) and the machine representation of multimedia objects. Most existing image retrieval systems are monomodal, as they utilize only visual or only textual information about images.

The semantic gap can be reduced by improving existing visual representations, making them suitable for a large-scale generic image retrieval. The best up-to-date candidates for a large-scale Content-based Image Retrieval are models based on the "Bag of Visual Words" framework. Existing approaches, however, produce high dimensional and thus expensive representations for data storage and computation. Because the standard "Bag of Visual Words" framework disregards the relationships between the histogram bins, the model can be further enhanced by exploiting the correlations between the "visual words".

Even the improved visual features will find it hard to capture an abstract semantic meaning of some queries, e.g. "straight road in the USA". Textual features, on the other hand, would struggle with such queries as "church with more than two towers" as in many cases the information about the number of towers would be missing. Thus, both visual and textual features represent complementary yet correlated aspects of the same information object, an image. Existing hybrid approaches for the combination of visual and textual features do not take these inherent relationships into account and thus the combinations' performance improvement is limited.

Visual and textual features can be also combined in the context of relevance feedback. The relevance feedback can help us narrow down and "correct" the search. The feedback mechanism would produce subsets of visual query and feedback representations as well as subsets of textual query and textual feedback representations. A meaningful feature combination in the context of relevance feedback should take the inherent inter (visual-textual) and intra (visual-visual, textual-textual) relationships into account.

In this work, we propose a principled framework for the semantic gap reduction in large scale generic image retrieval. The proposed framework comprises development and enhancement of novel visual features, a hybrid model for the visual and textual features combination, and a hybrid model for the combination of features in the context of relevance feedback, with both fixed and adaptive weighting schemes (importance of a query and its context). Apart from the experimental evaluation of our models, theoretical validations of some interesting discoveries on feature fusion strategies were also performed. The proposed models were incorporated into our prototype system with an interactive user interface.

# Acknowledgements

I would like to thank my supervisors, Prof. Dawei Song, Dr. Nirmalie Wiratunga, and Dr. Jeff Pan, for their guidance and support during my PhD study. I am especially grateful to Mr. Ben Horsburgh for his help with the implementation of some of the models and invaluable exchange of research ideas. Special thanks also go to Dr. Jun Wang, Dr. Peng Zhang, and Mr. Lei Wang for the collaboration and fruitful discussions, and to Prof. John McCall, Dr. Eyad Elyan, Dr. Andrei Petrovsky, Dr. Virginia Dawod, for various reasons.

Many thanks to The Robert Gordon University for granting me full studentship, and to my colleagues and friends: Carlos, Aakash, Ania, Gosia, Chamaka, Costas, Olivier, Jean-Claude, Malcolm, Peter, Sadiq, Amina, Richard, Noura, Nuka, Yanghui, Aleksandra, Alex, Pierre, Jackie, Sunitha, Saad and all the others that I have forgotten to mention. I am also indebted to The Aberdeen University and The Open University for collaboration and provision of the workspace and access to their research facilities. In addition, my gratitute goes to The Tianjin University in China, and all the staff I had a pleasure to meet there, for the hospitality and collaboration.

Last but not least, I thank my parents Edward and Krystyna, my sisters Ania and Iwona, and my friends Mariusz and Elwira, for their encouragement and support.

*This research was funded in part by the EPSRC Renaissance project (Grant No:EP/F014708/2), EPSRC AutoAdapt project (Grant No:EP/F035705/1) and the EU's Marie Curie Actions-IRSES QUONTEXT project (Grant No:247590).*

# Declarations

I declare that all of the work in this thesis was conducted by the author except where otherwise indicated.

Parts of the work outlined in this thesis have appeared in the following publications.

**Chapters 2, 3, 7**

- Kaliciak, L., Song, D., Wiratunga, N., Pan, J., (2010) Novel Local Features with Hybrid Sampling Technique for Image Retrieval. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (**CIKM2010**), pp. 1557-1560, Toronto, Canada.

- Kaliciak, L., Song, D., Wiratunga N., Pan, J., (2012) Improving Content Based Image Retrieval by Identifying Least and Most Correlated Visual Words. *The 8th Asia Information Retrieval Societies Conference* (**AIRS2012**), Springer, LNCS 7675, Tianjin, China.

**Chapters 2, 4, 7**

- Wang, J., Song, D., Kaliciak, L., (2010) Tensor Product of Correlated Text and Visual Features: A Quantum Theory Inspired Image Retrieval Framework. *AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes* (**QI2010**), pp. 109-116, Washington DC, USA.

- Kaliciak, L., Wang, J., Song, D., Zhang, P., Hou, Y., (2011) Contextual Image Annotation via Projection and Quantum theory Inspired Measurement for Integrating Text and Visual Features. *The 5th International Symposium on Quantum Interactions* (**QI2011**), Springer, pp. 217-222, Aberdeen, UK.

- Wang, J., Song, D., Kaliciak, L., (2010) RGU at ImageCLEF2010 Wikipedia Retrieval Task. *CLEF (Notebook Papers/LABs/Workshops) 2010*, Padua, Italy.

**Chapters 2, 5, 7**

- Kaliciak, L., Song, D., Wiratunga, N., Pan, J., (2013) Combining Visual and Textual Systems within the Context of User Feedback. *The 19th International Conference on Multimedia Modeling* (**MMM2013**), Springer, LNCS 7732, pp. 445-455, Huangshan, China.

**Chapter 8**

- Kaliciak, L., Horsburgh, B., Song, D., Wiratunga N., Pan, J., (2012) Enhancing Music Information Retrieval by Incorporating Image-Based Local Features. *The 8th Asia Information Retrieval Societies Conference* (**AIRS2012**), Springer, LNCS 7675, Tianjin, China.

# Contents

# List of Figures

# List of Tables

# Mathematical Symbols

The following table shows some of the mathematical symbols widely utilized in this thesis.

Table 0.1: Mathematical Symbols

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| $E$ | mean | $\sigma$ | standard deviation |
| $s$ | skewness | $tf$ | term frequency (text) |
| $idf$ | inverse document frequency | $corr$ | correlation |
| $vt$ | visual term ("visual word") | $T$ | transposition operator |
| $f$ | visual term frequency | $s(.,.)$ | similarity measurement |
| $d^v$ | visual feature vector | $d^t$ | textual feature vector |
| $\|.\rangle$ | ket, Dirac notation | $\langle.\|$ | bra, Dirac notation |
| $\delta$ | Kronecker delta | $\otimes$ | tensor product |
| $\oplus$ | vector concatenation | $L(d)$ | subspace generated by document $d$ |
| $\langle.\|.\rangle$ | inner product | $\|\|.\|\|$ | vector norm |
| $tr$ | trace operator | $P$ | projector onto a subspace |
| $H$ | Hilbert space | $str$ | strength of the relationship |
| $Q_m$ | modified query vector | $Q_0$ | original query vector |
| $D_j$ | related documents vector | $D_k$ | non-related documents vector |
| $D_r$ | set of related documents | $D_{nr}$ | set of non-related documents |

# Chapter 1

# Introduction

The aim of this thesis is to significantly reduce the semantic gap, the difference between user information needs and machine representation of images, by combining visual and textual features and exploiting the correlation and complementarity of both feature spaces. We also want to theoretically investigate the interactions between similarity measurements and operators utilized in existing hybrid models to shed more light on the combination methods.

The semantic gap can be also reduced by improving existing visual representations, making them suitable for a large-scale generic image retrieval. The best up-to-date candidates for a large-scale Content-based Image Retrieval are models based on the "Bag of Visual Words" framework. Existing approaches, however, produce high dimensional and thus expensive representations in terms of data storage and computation. The "Bag of Visual Words" model can be further enhanced by exploiting the correlations between the "visual words".

Even the improved visual features will find it hard to capture an abstract semantic meaning of some queries, e.g. "straight road in the USA". Textual features, on the other hand, would struggle with such queries as "church with more than two towers" as in many cases the information about the number of towers would be missing. Thus, both visual and textual features represent complementary yet correlated aspects of the same information object, an image. Existing hybrid approaches do not take these inherent relationships into account and thus the combinations' performance improvement is limited.

Visual and textual features can be also combined in the context of relevance feedback. The relevance feedback can help us narrow down and "correct" the search. The feedback mechanism would produce subsets of visual query and feedback representations as well as subsets of textual query and textual feedback representations. A meaningful feature combination in the context of relevance feedback should take the inherent inter (visual-textual) and intra (visual-visual, textual-textual) relationships into account.

In this work, we propose a principled framework for the semantic gap reduction in large-scale generic image retrieval. The proposed framework comprises development and enhancement of novel visual features, a hybrid model for the visual and textual features combination, and a hybrid

model for the combination of features in the context of relevance feedback, with both fixed and adaptive weighting schemes (importance of a query and its context). Apart from the experimental evaluation of our models, theoretical validations of some interesting discoveries on feature fusion strategies were performed. Based on the experimental results, researchers hinted at the potential interchangeability of specific fusion strategies. We show that specific fusion schemes are indeed interchangeable and this duality is caused by the interactions between similarity measurements and fusion operators.

The proposed models were incorporated into our prototype system with an interactive user interface.

This chapter presents the fundamental concepts of Content-based Image Retrieval (CBIR), the motivation behind the research with research objectives, and organization of the thesis.

## 1.1 Content-based Image Retrieval

Content-based Image Retrieval (CBIR) is the application of computer vision methods to the image retrieval problems. Most existing image retrieval search engines utilize only the textual information attached to images (metadata, tags). CBIR tries to model an actual visual content of images. The visual features, similarly to textual ones, represent images in a vector form. The similarity between two images can then be measured as a distance between these visual representations, for example. Thus, CBIR is in most cases based on the standard Vector Space Model.

### 1.1.1 Colour Spaces

Colour spaces are abstract mathematical models that represent colours as tuples of numbers, thus mapping the colours to a reference colour space. Many visual features utilize the colour information, thus they are defined in a specific colour space. The choice of the colour space may affect the image retrieval performance, because some colour spaces better resemble the human perception of colours.

Thus, the two most common colour spaces are RGB (red, green, blue; cubic coordinates) and HSV (hue, saturation, value). As HSV (cylindrical representation of RGB colours) better model human colour perception, it is desirable to first convert images from RGB to HSV colour space (del Bimbo 1999).

Colour spaces can be utilized in both low and high level visual features.

### 1.1.2 Low Level Visual Features

Low level visual features capture colour, texture, and shape properties of an image. When the features are extracted from the whole image, we call them global. If one tries to capture the local properties of an image (image sampling, segmentation), we extract local features. Thus, global features are often referred to as low level visual features.

Colour-based methods include colour histograms, colour moments, and colour correlograms. Texture-based methods consist of approaches incorporating co-occurrence matrices, Gabor wavelets, and fractal dimension. Shape may be captured by statistical moments or Fourier descriptors.

Sometimes combinations of features can enhance the retrieval. One of the intuitive choices is to combine colour with texture. The easiest way to do this is to simply concatenate the features or compute the co-occurrence matrices for individual colour channels (most existing approaches convert images to grayscale format first).

#### Colour

This feature is one of the most widely used. One of the advantages of using colour feature is its insensitivity to variations in image size and orientation. The computational inexpensiveness is also very important.

The most common approach is to represent colours in a form of histogram (Bui, Lenz & Kruse 2005), usually computed as three independent colour distributions. First, an image is split into individual colour channels (grayscale representations of individual colours). Thus, in case of RGB colour space we have Red, Blue and Green channels, while HSB splits into Hue, Saturation and Brightness etc. The next step is to discretize image colours and count how many pixels belong to each bin. Because histograms do not take the spatial information into account, two images perceived by humans as different may have similar representations. Colour space may play an important role in visual information retrieval by colour.

Another representation of colour features is the three colour moments extracted from each individual colour channel: mean, standard deviation and skewness. This approach assumes that the distribution of colour can be treated as probability distribution. The first moment can be interpreted as an average colour value, second as a square root of the variance of the distribution, and third as the measure of asymmetry in the distribution.

Colour moments usually perform better than colour histogram (Stricker & Orengo 1995). However, the drawback of the colour moments is their sensitivity to illumination changes.

Colour correlograms try to capture the spatial information of colour in colour histograms. The easiest way to capture this is to divide and image into a small, uniform number of sub-images (i.e. 6), generate the colour histograms for each image patch, and concatenate thus obtained histograms.

#### Texture

Texture is a visual pattern, a property of almost all surfaces. Texture can be defined in terms of coarseness, contrast, directionality, line-likeness, regularity, and roughness (Tamura, Mori & Yamawaki 1978).

It can be also represented as a co-occurrence matrix (Gotlieb & Kreyszig 1990). The matrix describes the way certain grayscale pixel intensities occur in relation to other grayscale pixel inten-

sities. It counts the number of such patterns. Another, more complex definition of co-occurrence matrix takes into account all eight directions plus the distances between pixels. This approach to texture description, despite being one of the very first, is still popular. Once the matrices are computed, we extract some meaningful statistics which can be later compared with one another.

Many methods based on wavelets have been applied to content-based systems. One can use the energy of the wavelet coefficients as a texture descriptor (de Ves, Ruedin, Acevedo, Benavent & Seijas 2007).

Fractals have also found their application to texture description. Specifically, fractal dimension with its box counting algorithm can be used to characterize texture. (Dobrescu, Dobrescu & Ichim 2006)

#### Shape

Shape can be represented as a boundary or the entire region. If we want to focus on regional properties (colour, texture, etc.) we should choose the internal representation (region-based). When shape characteristics are important, the best option is the external representation (boundary-based). The shape descriptors should be invariant to scaling, translation, and rotation. The most widely used descriptors are statistical moments and Fourier descriptors. In order to obtain Fourier descriptors we represent the boundary in a form of a complex number and apply the discrete Fourier transform.

The statistical moments such as mean, variance, higher order moments can also successfully describe the required shape. There are plenty of different improved variations of the moments.

### 1.1.3 Higher Level Visual Features

Local features capture the information about local image properties. These approaches are usually based on the "Bag of Visual Words" framework (Lowe 1999), which was inspired by the "Bag of Words" models from text information retrieval. Thus, the local features were designed to better model human perception and have some ability to recognize objects.

#### Mid-Level Visual Features

The first step in the "Bag of Visual Words" framework for Content-based Image Retrieval is to localize the so-called interest points (point-like, region-like). There are various detectors that can perform this task (Mikolajczyk & Schmid 2004). We can divide them into two (often overlapping) groups:

- Corner detectors (point-like features)
- Blob detectors (regions)

The regions around interest points should contain rich local information about the image. The recent state-of-the-art detectors make these regions invariant to scale, affine transformations, and illumination/brightness variation. There are also other sampling techniques, namely random and dense.

The next step is to describe the regions around the keypoints with one of the available descriptors. The common method is to utilize certain clustering techniques (e.g. k-means clustering, fuzzy clustering) to create the so-called codebook or visual vocabulary. The descriptors are classified into clusters (visual words), and the images are represented as the number of occurrences of each visual word, or a presence or absence of a feature (or another alternative). The most popular interest points detectors are: Harris-affine, Hessian-affine, Scale Invariant Feature Transform (SIFT), Maximally Stable Extremal Regions (MSER). Among descriptors we have SIFT (detector and descriptor), local jets (image derivatives), steerable filters, and generalized moment invariants.

#### High Level Visual Features

The high level visual features are often based on image segmentation or grouping of visual words into semantically meaningful groups. These approaches, however, are too data storage and computationally costly for a large scale generic image retrieval. Therefore, they are only utilized in domain specific tasks with a limited number of semantic concepts.

#### Enhancing "Bag of Visual Words" Framework

One of the drawbacks of the "Bag of Visual Words" framework is the assumption that the visual words (bins in histograms of visual words counts) are independent of each other. Actually, these dimensions are often correlated. This correlation can be exploited to further improve the aforementioned framework (Liu, Liu, Liu & Lu 2009).

### 1.1.4 Relevance Feedback

Relevance feedback is a mechanism which can guide the system in the direction of users information needs. It has the ability to narrow down and refine the search, thus improving the overall performance of a search engine. This is one of the tools that can help us reduce the semantic gap.

The relevance feedback can be collected directly or indirectly. When the relevance of retrieved images is judged explicitly, for example, we refer to it as explicit feedback. Implicit feedback collects the information about the interactions with the system based on the query history or click through data, for instance.

Models based on relevance feedback (for visual and textual features) often modify or expand the query so that it better resembles the current information need.

One of the first approaches which is, however, still widely utilized is the Rocchio algorithm (Rocchio 1971). It modifies the query based on the relevance feedback (positive and negative) so

that it moves closer to the centroid of relevant documents and further away from the centroid of irrelevant ones.

An automated variation of relevance feedback mechanism is called Pseudo Relevance Feedback. It assumes that the top $n$ documents are all relevant to the current query.

The problem with Pseudo Relevance Feedback models is their high dependency on the quality of initial results.

## 1.2 Hybrid Models for Visual and Textual Features Combination

Existing hybrid models usually combine visual and textual features in the following way

1. pre-filter the data collection by visual content and then re-rank the top images by text (Yanai 2003);
2. pre-filter the data collection by text and then re-rank the top images by visual content (Tjondronegoro, Zhang, Gu, Nguyen & Geva 2006);
3. pre-filter the data collection by visual (textual) content and then aggregate the scores of the textual (visual) representations of the top retrieved images (transmedia pseudo-relevance mechanism (Maillot, Chevallet & Lim 2007));
4. fuse the representations (early fusion (Rahman, Bhattacharya & Desai 2009));
5. fuse the scores or ranks (late fusion (Simpson, Rahman, Singhal, Demner-Fushman, Antani & Thoma 2010)).

The most widely used early fusion operation is vector concatenation. In case of a late fusion, the most common strategies rely on a sum of scores, their product, and the linear combination of scores.

Based on the experimental results, researchers have hinted at the potential interchangeability of specific fusion strategies.

## 1.3 Motivation Behind the Research

The main purpose of this research is to significantly reduce the semantic gap. This can be achieved by developing high level visual features - currently feasible in domain specific tasks only. Another approach would be to intelligently combine various sources of evidence (to exploit correlation and complementarity of the feature spaces) and improve the individual components (representations), adjust them to generic large-scale image retrieval task.

Existing hybrid models for the combination of visual and textual features are often based on the data fusion strategies and do not capture the complementarity and correlation between feature spaces. Moreover, there is a belief that specific fusion strategies may be interchangeable (similar

experimental performance). In addition to proposing novel hybrid models, we are the first to prove that specific fusion schemes are indeed equivalent. Later in the thesis we discuss the important implications of these observations.

Thus, we propose a principled framework to reduce the semantic gap by:

- development of novel visual features for a large scale generic image retrieval;
- enhancement of the Bag of Visual Words by exploitation of the correlation between visual words;
- combination of visual and textual features by utilizing features correlation and complementarity;
- combination of visual and textual features in the context of relevance feedback by exploiting the inter and intra relationships between the feature spaces;
- adaptation of weights associated with the importance of a query and its context (feedback sets) in the novel hybrid relevance feedback model;
- development of a prototype hybrid system with an interactive user interface;

A parallel aim of our research is the theoretical investigation of the existing hybrid models in terms of potential interchangeability between specific fusion strategies.

The following gaps have been identified and research questions asked:

1. **developing novel visual features for a large scale generic image retrieval**

   The visual features based on the Bag of Visual Words framework are currently the best candidate for large-scale generic image retrieval. They have some ability to recognize objects and generally outperform global models. Most of these methods, however, produce high dimensional vectors which leads to high data storage and computational costs. This, in turn, requires the application of dimensionality reduction techniques. Moreover, most existing approaches (in the retrieval of generic real-life images) are based on interest points detectors, but quite often more discriminant image patches are generated at random. The higher the number of random sample points, the better the performance. This, however, comes at a price of higher computational and data storage cost.

   Can we develop low-dimensional but effective local features which could be utilized in generic large-scale image retrieval? Can we also increase the discriminative power of sampling methods without increasing the computational and data storage cost?

2. **enhancing local features by exploiting the correlation between visual words**

   Methods that utilize information about correlations between visual words try to group semantically similar visual words together. They usually consider co-occurrences at one contextual level only and are computationally expensive and not scalable. Moreover, existing

approaches (query modification like frameworks) modify the current query, which leads to the normalization of histograms. This may not be desirable, since the (mid-level) semantic meaning of bins may be lost and the representations may become less discriminative due to the varied complexity of images. Some researchers report performance improvement after normalization, while others report the opposite. We believe that this may be also domain specific. The $tf \cdot idf$ weighting scheme may work better in case when the precise object matching is important. Our experiments, for example, showed that the normalization of histograms of visual words counts hampers the retrieval performance.

Can we completely avoid the modification (and thus renormalization) of histograms of visual words counts when trying to enhance the Bag of Visual Words representation based on the correlations between visual words? Can we develop a model that would make this enhancement feasible in large-scale generic image retrieval? Can we compare different notions of correlation at different contextual levels within the aforementioned framework? Can we introduce more intuitive notion of image-level correlation? We argue that the standard image level correlation is not intuitive enough. For example, if the frequencies of two pairs of visual words are $\{5, 10\}$ and $\{5, 100\}$ then the latter will be assigned higher correlation value. We would, however, expect the former pair to be at least equally correlated.

3. **combining visual and textual features by utilizing features correlation and complementarity**

   The ranking of documents is often computed by heuristically combining the feature spaces of different media types or combining the ranking scores computed independently from different feature spaces. All these combination methods treat the textual and visual features individually, and combine them in a rather heuristic manner. Therefore, this makes it difficult to capture the relationship between the features. Indeed, as both the textual and visual representations describe the same image, there are inherent correlations between them which should be incorporated into the retrieval process as a whole in a more principled way.

   It is commonly believed that the fusion strategies (utilized in the existing hybrid models) represent different combination methods and that the drawback of early fusion is the potential curse of dimensionality, while the drawback of a late fusion is its inability to capture the correlation between the features dimensions.

   Can we utilize the inherent correlations between visual and textual features in a hybrid model? Is it possible to show the meaningfulness and theoretical interpretation of some combination methods? It is relatively easy to combine the scores (for example) in such a way that the retrieval performance will be improved. It is more difficult, however, to show that the combination is meaningful and universal.

   Our tensor-based model requires images to have annotations. Existing auto-annotation methods based on segmentation models or grouping of visual words are computationally

expensive and not scalable to large data collections. Methods based on the clustering of visual and textual features, on the other hand, neglect the contextual information.

How can we develop a computationally cheap model that can capture the contextual information, and can be seamlessly integrated into the tensor-based framework?

4. **combining visual and textual features in the context of relevance feedback by exploiting the inter and intra feature relationships between the feature spaces**

   Existing hybrid relevance feedback models, similarly to the hybrid models, combine the features in a rather heuristic manner. They do not exploit the inter and intra feature relationships intrinsic to both feature spaces. The intra relationships correspond to individual feature spaces, while the inter relationships correspond to correlation and complementarity of both visual and textual feature spaces.

   How can we exploit the aforementioned inter and intra feature relationships in a hybrid relevance model for a large scale generic image retrieval?

5. **adaptation of weights associated with the importance of a query and its context (feedback sets)**

   A query can be more or less related to its context. Existing models which try to make the weights adaptive, require the manual setting and training of these weights. Moreover, there is a lack of approaches that would incorporate adaptive weighting schemes into hybrid relevance feedback models.

   Can we develop an approach, based on the relationship strength between the query and its context, that does not require the training phase? How can we incorporate the adaptive weighting scheme in a hybrid relevance feedback model?

6. **development of a prototype hybrid system with an interactive user interface**

   Interactive user interfaces can help us facilitate user-system interactions and fully exploit the implemented models. We need a hybrid relevance feedback prototype system that offers all the functionalities of some monomodal approaches (and more, i.e. a natural way of incorporating various degrees of relevance and exploratory search).

All these gaps are important to the hybrid modelling, which we choose as our tool for semantic gap reduction.**Thus, the key evaluable and verifiable objectives of this thesis are:**

1. **experimental evaluation of novel visual features for generic large-scale CBIR and enhancement of local visual features;**
2. **experimental evaluation of hybrid model for the combination of visual and textual features;**
3. **experimental evaluation of hybrid model for the combination of visual and textual features in the context of relevance feedback;**

4. **experimental evaluation of adaptation of weights associated with the importance of a query and its context (feedback images) in our hybrid relevance feedback model;**

5. **theoretical verification of the hypothesis on the interchangeability of specific fusion strategies;**

6. **theoretical verification of the hypothesis on the representation of query modification in a late fusion form;**

## 1.4 Contributions

Next, we describe our contributions.

### 1.4.1 Novel Visual Features (Objective 1)

We introduce novel global methods based on our novel edge detector, bilateral filtering, directional derivatives, and pixel intensities thresholding.

We propose a novel method based on the local features, incorporating an easy to implement descriptor and a hybrid sampling technique. We also compare different sampling methods: hybrid (a combination of random and detector-based sampling), purely random, purely detector-based and dense on three large data collections. The hybrid sampling produces more discriminative of image patches than the commonly used detector-based method. The proposed descriptor is easy to implement and produces low dimensional feature vectors which, in turn, reduces the computational and data storage cost. Empirical evaluation has been performed on three large image collections, namely ImageCLEF 2007, MIRFlickr 25000 and BGS datasets.

Our approach is easy to implement, not sophisticated, with low computational and data storage cost (mostly because our vectors are low dimensional), and the hybrid sampling technique can be used in other methods based on the "bag of visual words" to improve the retrieval performance. Moreover, the evaluation of the proposed method has been conducted on three different large data collections without changing the initial setup. In this way we avoided "fine-tuning" of the parameters to the specific data collection which makes the results more general and reliable.

### 1.4.2 Enhancement of Local Features (Objective 1)

We propose a new approach for identifying and utilizing the information about correlations between visual words. We implement and test various notions of correlation at different contextual levels (we refer to them as image-level and proximity based). To the best of our knowledge, this is the first time these two were compared within this type of framework in image retrieval. Our local features consist of low dimensional histograms, where bins representing visual words are highly correlated. We identify the most and the least correlated coefficients and use thus obtained

information, along with the visual terms' frequencies from the current query, to weight the similarity measure. Certain coefficients in the similarity measure corresponding to the most correlated terms are then increased, while the coefficients related to the least correlated pairs are deemphasized. The evaluation was performed on three large data collections, namely ImageCLEF 2007, MIRFlickr 25000 and BGS. The evaluation was performed within the Pseudo Relevance Feedback framework.

Experimental results show the superiority of two notions of correlation, which are image level correlations. For these two correlations, we report significant improvement in terms of Mean Average Precision on two data collections within PRF evaluation framework. Moreover, the addition of information about the least correlated visual words often further improves the performance. The proximity based notion of correlation does not show a significant improvement in the context of this model.

The proposed method is computationally and data storage cheap, utilizes correlation at different contextual levels, and avoids the normalization of histograms.

### 1.4.3 Novel Hybrid Model (Objective 2)

We propose a quantum theory inspired multimedia retrieval framework based on the tensor product of feature spaces, where similarity measurement between a query and a document corresponds to the quantum measurement. At the same time, the correlations between dimensions across different feature spaces can also be naturally incorporated in the framework. The tensor based model provides a formal and flexible way to expand the feature spaces, and seamlessly integrate different features, potentially enabling multi-modal and cross media search in a principled and unified framework.

Experimental results on a standard multimedia benchmarking collection show that the quantum-like measurement on a tensored space leads to remarkable performance improvements in terms of average precision over the use of individual feature spaces separately or concatenation of them.

### 1.4.4 Novel Hybrid Model and Image Auto-Annotation (Objective 2)

We introduce and test two novel approaches for making associations between tags and images. We also experiment with mid-level semantic image representations based on the "bag of visual words" model. This was a follow-up work on our tensor-based unified image retrieval framework. In order to prepare the data for the quantum-like measurement in the tensor space, we need to alleviate the problem regarding the unannotated images. The first proposed approach projects the unannotated images onto the subspaces generated by subsets of training images (containing given textual terms). We calculate the probability of an image being generated by the contextual factors related to the same topic. In this way, we should be able to capture the visual contextual properties of images, taking advantage of this extended vector space model framework. The other method performs quantum-like measurement on the density matrix of unannotated image, with respect to

the density matrix representing the probability distribution obtained from the subset of training images. These approaches can be seamlessly integrated into our unified framework for image retrieval.

### 1.4.5 Novel Hybrid Relevance Feedback Model (Objective 3)

We also extend our hybrid model for visual and textual feature combination within the context of relevance feedback. The approach is based on mathematical tools also used in quantum mechanics - the predicted mean value of the measurement and the tensor product of the density matrices, which represents a density matrix of the combined systems. It was designed to capture both intra-relationships between features' dimensions (visual and textual correlation matrices) and inter-relationships between visual and textual representations (tensor product). The model provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context by applying quantum-like measurement. It opens a door for a series of theoretically well-founded further exploration routes, e.g. by considering the interference among different features. It is easily scalable to large data collections as it is general and computationally cheap. The results of the experiment conducted on ImageCLEF data collection show the significant improvement over other baselines.

### 1.4.6 Dynamic Weighting of Query and its Context (Objective 4)

The aforementioned model for the visual and textual features combination utilizes fixed weights corresponding to the importance of the query and its context. However, the query can be more or less related to its context. Inspired by this observation, we incorporate adaptive weighting scheme into the hybrid CBIR relevance feedback model. Thus, each query is associated with a unique set of weights corresponding to the relationship strength between a visual query and its visual context as well as the textual query and its textual context. The higher the number of terms or visual terms (mid-level features) co-occurring between current query and the context, the stronger the relationship and vice versa. If the relationship between query and its context is weak, context becomes important. We adjust the probability of the original query terms, the adjustment will significantly modify the original query. If the aforementioned relationship (similarity) between query and its context is strong, however, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

We tested the enhanced model within the user simulation framework. For fair comparison purposes, the best performing sets of fixed weights were selected. We have shown that our enhanced model with adaptive weighting scheme can outperform the original one with fixed weights. Moreover, an addition of another visual feature (colour histogram, global feature) further improved both the hybrid CBIR relevance feedback model and the enhanced model's performance, whereas the

performance of the baselines did not change much.

Our contribution here is related to showing how to measure the relationship strength between query and its context, and how to incorporate the adaptive weighting scheme into the state-of-the-art existing model (hybrid, user feedback context) to further improve the retrieval. The proposed adaptive weighting approach is relatively easy to implement and does not require any training of features.

### 1.4.7 Duality of Fusion Strategies (Objective 5)

In this work, we theoretically investigate some interesting interactions between common similarity measurements and common operators related to early fusion strategy (e.g. concatenation of vector representations). We show that these interactions between certain similarity measurements and early fusion strategies result in combinations of representations at the decision level (late fusion strategy). In other words, we theoretically prove that certain combinations of early fusion strategies and certain similarity measurements are equivalent to particular combinations of measurements (i.e. relevance scores) computed on individual feature spaces.

Our findings are important from both theoretical and practical perspectives. First, we should be careful when comparing early and late fusion strategies as they may represent equivalent approaches. Second, specific combinations of individual relevance scores can have a sound theoretical interpretation which would be easier to analyze as an early fusion. Finally, knowing how to represent early fusion as a late one can help us avoid the curse of dimensionality.

Thus, an interesting open question arises - the most important, in our opinion, consequence of our observations. Does late fusion strategy, contrary to current belief, capture the relationships between feature spaces or does the interaction between the similarity measurement and early fusion operators decorrelate features? This question should make us look at feature combination methods from different perspective.

### 1.4.8 Query Modification as a Late Fusion (Objective 6)

Query modification strategies utilize relevance feedback to modify the query in order to narrow down the search. We show that query modification can be represented as a late fusion strategy. This observation has a few important implications. First, some complex combinations of measurements performed on individual feature spaces may be regarded as a query modification technique. Second, query modification represented as a late fusion does not require an actual modification of query representation. The actual query modification would often lead to query renormalization. While renormalization in text IR is often desirable, renormalization of mid-level visual representations may even hamper the retrieval performance. Third, it is often easier to implement and work with relevance scores. Finally, knowing how to represent query modification as a late fusion can help us develop a family of hybrid relevance models (combinations of scores computed on lower dimensional feature spaces as opposed to high dimensional hybrid representations).

### 1.4.9 Prototype Hybrid System with Interactive User Interface

We present an interactive user interface which has been synchronized with our prototype system. The user interface is the communication platform between the system and the user. In order to show a working demo of the prototype system, we have integrated it with the interactive user interface. Thus, we present a unified working framework (demo of our prototype system) for content-based image retrieval comprising: novel visual features for generic image retrieval and their combinations with existing visual features (if selected), combination of visual and textual features in the first round retrieval, combination of visual and textual features in the context of relevance feedback (search refinement), interactive user interface. To the best of our knowledge, this is the first prototype system that utilizes hybrid model for combination of visual and textual features as well as a hybrid relevance feedback model, and allows for a full interaction with the system.

### 1.4.10 Other Contributions

Other contributions include a proposal of a family of hybrid models for visual and textual features combination, as well as a family of hybrid relevance feedback models. We also provide a theoretical model for the combination of visual and textual features in the context of user feedback with various degrees of relevance (discrete and continuous). Thus, a user would drag and drop feedback images in a natural way onto a relevance bar. We also introduce a number of variations of the experimentally tested models.

#### 1.4.10.1 Application of Novel Local Features to Music Genre Classification

We introduce a novel approach to MIR. Having represented the music tracks in the form of two dimensional images, we apply the "bag of visual words" method from visual IR in order to classify the songs into 19 genres. The motivation behind this work was the hypothesis that 2D images of music tracks (spectrograms) perceived as similar would correspond to the same music genres (perhaps even similar music tracks). Conversely, we can treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This would point to an interesting interchangeability between visual and music information retrieval.

We obtained classification accuracy of 46% (with a 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.

The main advantages of our method are: it is a more intuitive, easy way to automatic music classification, has classification accuracy comparable with state-of-the-art and is a promising new research direction.

Some of the work presented in this thesis have been published and presented at various conferences. For the list of publications see the "Publications" appendix at the end of the thesis.

## 1.5 Organization of the Thesis

The thesis is organized as follows:

- The literature survey of relevant works is presented in Chapter 2, along with identified problems and gaps.
- Chapter 3 presents our novel visual features for large scale generic image retrieval. We also introduce an approach for the enhancement of the "Bag of Visual Words" framework based on the correlations between visual words.
- Chapter 4 describes our tensor-based hybrid model for the combination of visual and textual features, projection-based approaches for extracting image-tag associations, and investigation on the interactions between similarity measurements and early fusion operators with theoretical validations.
- A novel hybrid relevance feedback model, which exploits inter and intra feature relationships between visual and textual feature spaces, is presented in Chapter 5. Here, we also present an enhanced hybrid relevance feedback model with dynamic weighting of query and its context depending on the relationship strength between them. Moreover, some variations of the original model are presented in this chapter. One is a hybrid relevance feedback model with discrete and continuous degrees of relevance. Another is a variant based on orthogonal projection. Chapter 5 also includes observations on how the query modification techniques can be represented as specific combinations of relevance scores. The hypothesis is supported by theoretical proofs.

  Further, the generalization of the original model is presented that allows us to combine multiple visual features and textual features. We also show how the observation on the query modification represented as a late fusion can lead to generation of a family of hybrid relevance feedback models.
- The experimental results (novel visual features, enhancement of local features, hybrid model, hybrid relevance feedback model, adaptive weighting scheme) are presented in Chapter 6.
- Chapter 7 describes our prototype demo system with an interactive user interface.
- Chapter 8 presents an application of our novel local features to music genre classification and food recognition task.
- Finally, in the last chapter, we conclude the thesis and highlight our key contributions and future research directions.

# Chapter 2

# Literature Survey

This chapter presents existing research related to our project. Thus, we start from an overview of existing models which try to capture the visual content of images. We divide the visual features into two groups: low-level and higher-level visual features. Related work identifies the Bag of Visual Words framework as the (current) main candidate for generic large-scale content based image retrieval. The framework assumes that the visual words are independent of each other whereas in fact, they are often correlated. These correlations can be utilized in order to enhance the Bag of Visual Words framework. Thus, we also review the literature related to the exploitation of the relationships between visual words (local features).

Works related to the combination of visual and textual features follows the literature survey on visual features. We review research on various data fusion strategies and pre-filtering approaches. We also list some works related to the applications of mathematical tools from quantum mechanics to information retrieval, as some of our models utilize such tools. Moreover, because one of our hybrid models require the images to be associated with text, we review approaches related to image auto-annotation.

Only limited research has been done on hybrid relevance feedback models in the literature. Therefore, the review of related work in this area will be focused on monomodal relevance models (one feature only). Similarly, adaptive weighting schemes for balancing the importance of query and its context represent a novel and underexplored research area. A few existing adaptive weighting schemes were utilized in monomodal relevance models.

Some of our models were implemented in the prototype system with interactive user interface. Thus, we also review the works on existing state-of-the-art systems with interactive user interfaces.

The identified gaps are described in detail in the core part of this chapter and in the chapter summary.

## 2.1 Content-based Image Retrieval

Content-based Image Retrieval (CBIR) is the application of computer vision methods to the image retrieval problems. Thus, the content of images in the data collection may not only be represented by the textual descriptions (tags) attached to these images but also by their visual content.

An example CBIR framework is shown in Figure 2.1.

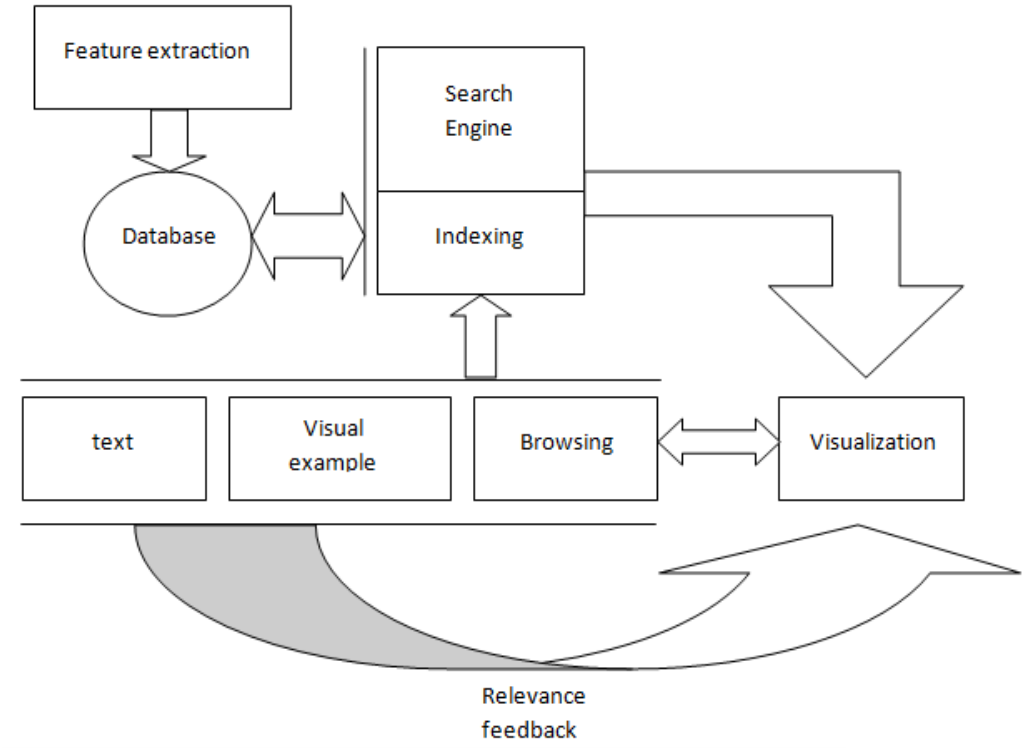


Figure 2.1: Content-based Image Retrieval System - Framework

Both text and visual content can be characterized in a vector form, hence the most common CBIR framework is based on the Vector Space Model. The similarity between two images can be measured as a distance between two vectors.

The CBIR system can often be queried by a visual example (an image representing the user's information needs) or text.

CBIR is highly correlated with many other research disciplines including digital image processing. CBIR incorporates techniques from low to higher level processing. Low level processes involve noise reduction, image sharpening etc., mid level processes deal with segmentation for instance, and higher level processes try to "make sense" of the recognized objects.

Although many content-based methods and techniques exist, CBIR is unable to imitate human perception properly. That is why attempts to bridge the semantic gap (the difference between human perception and computer representation of an image) by all possible means (hybrid approaches) have become so important. One of the approaches that can reduce the semantic gap is

user relevance feedback algorithm, usually combined with some learning techniques (genetic algorithms for example). Other ways to reduce the gap are based on classification algorithms (classifying images into classes instead of ranking them), using support vector machine or Bayesian classifier for instance. Ontological engineering can help to logically describe features carrying abstract meaning. Because of the subjectivity of human perception, it is important to maintain user profiles (short or long-term), each profile consisting of (for example) some weights that can simulate individual user perception by changing the similarity measure accordingly. In addition, tags, ratings, comments, etc. (the so-called "social media") and their analysis can also improve the retrieval.

Figure 3.2 presents an example CBIR system.

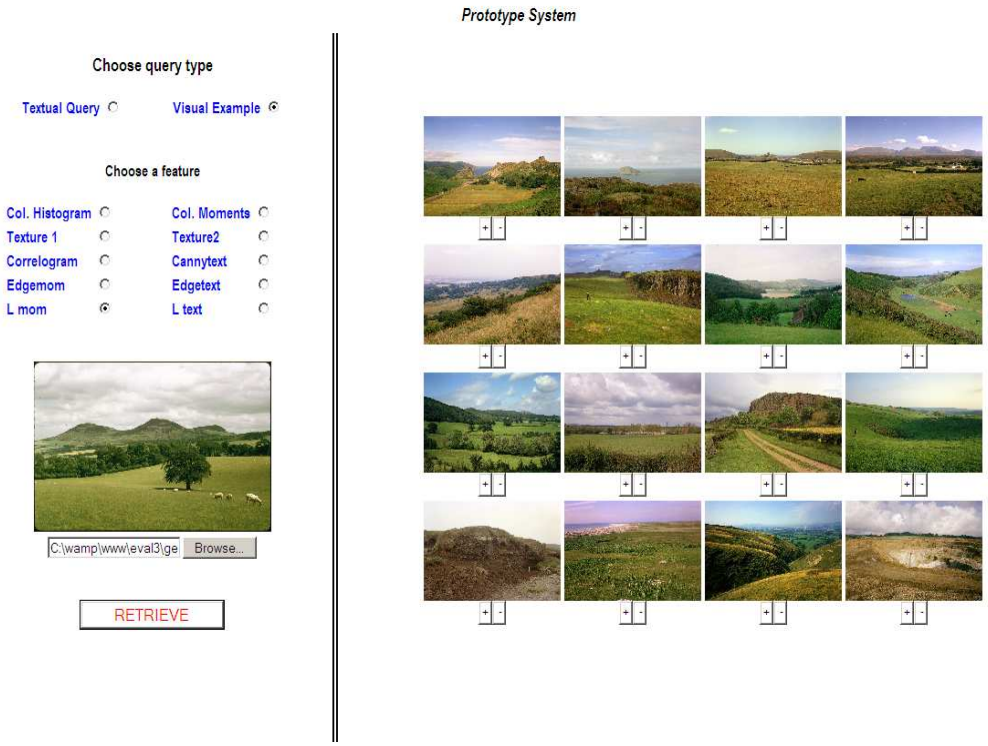


Figure 2.2: Content-based Image Retrieval System

## 2.2 Visual Content Representations

Visual features comprise colour, texture, shape, structure, and semantic primitives. When the features are extracted from the whole image, we call them global. If we try to capture the local properties of an image (image sampling, segmentation), we extract local features.

### 2.2.1 Low-level Visual Features

We will refer to global features as low-level visual features.

#### Colour

Colour histogram is one of the simplest but widely used visual features. According to (del Bimbo 1999) hue and saturation components help to retain light independent colour properties. (Bui et al. 2005) incorporate text, colour histogram, and texture into their prototype system. The addition of content significantly improved the performance of the system. The problem with colour histogram is that two visually different images may have similar colour distributions (colour histograms).

The colour moments were utilized in (Muhling, Ewerth, Stadelmann, Shi & Freisleben 2009) in HSV (Hue, Saturation, Value) colour space. It has been experimentally proven by (Stricker & Orengo 1995) that the three colour moments perform better than colour histogram. The drawback is that colour moments are sensitive to illumination changes.

#### Texture

For the analysis of texture from the psychological point of view the reader can be referred to (Landy & Graham 2004). (Tamura et al. 1978) define texture in the following terms: coarseness, contrast, directionality, line-likeness, regularity, and roughness. It was shown that the first three statistics perform best.

The co-occurrence matrix is one of the first approaches to texture modelling. (Gotlieb & Kreyszig 1990) have found experimentally that the best discrimination between different textures can be obtained by contrast, inverse difference moment, and entropy. The co-occurrence matrix approach was also incorporated in (Celebi & Alpkocak 2000) for texture representation and clustering. The statistics computed from matrices were entropy, contrast, homogeneity, and variance. (Muhling et al. 2009) implement the grayscale co-occurrence matrices constructed for eight orientations with following statistics: energy, entropy, contrast and homogeneity. Because construction of a co-occurrence matrix is computationally expensive, some authors proposed new methods to reduce the computational burden. (Mokji & Abu Bakar 2007) computed the matrix based on the Haar wavelet transform. They managed to reduce the construction cost by up to 62.5%.

In general, a lot of methods based on wavelets have been applied to content-based systems. One can use the energy of the wavelet coefficients as a texture descriptor. (de Ves et al. 2007) propose a novel texture descriptor based on a wavelet transform. The information from the moduli and orientations of wavelet coefficients is extracted in order to characterize textures in a database, working under certain assumptions.

Fractals also have found their application to texture description. Specifically, fractal dimension with its box counting algorithm can be used to characterize texture. (Dobrescu et al. 2006) intro-

duce a novel integrated feature vector for texture classification, consisting of statistics extracted from co-occurrence matrix and a histogram of fractal dimension.

### 2.2.2 Mid-level Visual Features

(Mikolajczyk & Schmid 2004) evaluated a few most widely used affine invariant detectors and descriptors, and the results were as follows:

- Best repeatability score (the same keypoints detected after transformations) was obtained by MSER, Harris-affine, and Hessian-affine detectors.
- MSER was the most accurate detector (low regions overlap error).
- Harris-affine had the highest number of points detected.
- Harris-affine and Hessian-affine obtained the most stable results (regardless of the scene type).

The "Bag of Visual Words" approach based on a SIFT detector and descriptor was first proposed by Lowe in (Lowe 1999). Other good sources of information about scale-space and local features are (Mikolajczyk & Schmid 2005) and (Lindeberg 1994). Figure 2.3 presents the "Bag of Visual Words" Framework (adapted from (Yang, Jiang, Hauptmann & Ngo 2007)).

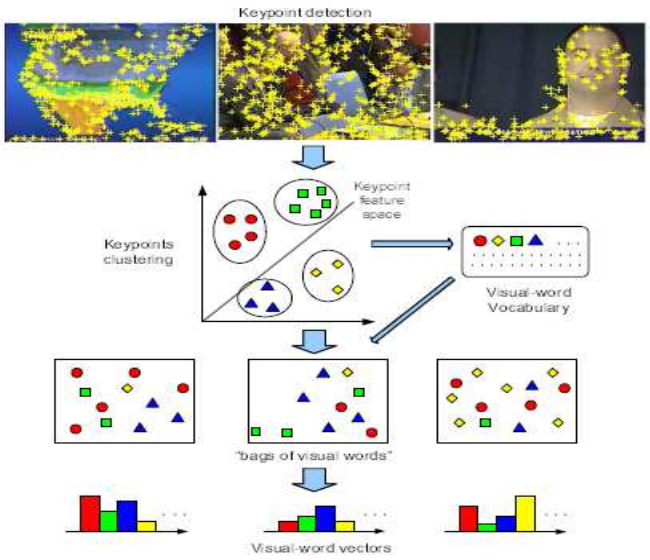


Figure 2.3: "Bag of Visual Words" Framework

(ao Lopes, de Avila, Peixoto, Oliveira & de Albuquerque Araújo 2009) utilize the local features to nude and pornographic images detection. Instead of using the well known SIFT descriptor, they implemented a Hue-SIFT method (colour histogram and local features based on SIFT) in order to take into account the colour information.

(Yang et al. 2007) incorporate and test some methods derived from the textual information retrieval domain into CBIR (Bag of Visual Words): term weighting, stop word removal, feature selection. They conducted experiments to test the influence of the vocabulary size (number of clusters) and spatial information on the retrieval performance. It turns out that the spatial information is quite important when working with small vocabularies but loses its significance as the size of the codebook increases. The spatial information they utilize, however, is the absolute spatial information. An image is first divided into a number of identical subimages, and the histogram of visual words counts is generated for each subimage.

The method proposed in (Nowak & Jurie 2007) learns the similarity measure and utilizes a random sampling technique to retrieve previously never seen objects. The algorithm is based on randomized binary trees (a combination of scale invariant feature transform and geometry) and randomly selected rectangular regions of random size. Authors try to find corresponding patches for each selected region. Although they obtained very good results, the method is rather suitable for object instance recognition, not generic image retrieval.

(Nowak, Jurie & Triggs 2006a) compare different sampling techniques (corner detectors, random sampling, dense sampling). The experimental results showed that the number of sampled patches significantly affected the retrieval performance. Surprisingly, for a large number of patches (which is desirable) the keypoints detected by corner/blob detector tend to lose their discriminative power. It is therefore better to use randomly selected patches. However, there is no comparison with combinations of different sampling techniques and the high number of sample points increases the cost in terms of data storage and computation.

The novel descriptor proposed in (Nowak, Jurie & Triggs 2006b) uses Gaussian filter banks to characterize local patches. The method seems to outperform SIFT but only on a specific small data collection - Wang database.

(Moosmann, Triggs & Jurie 2007) present a novel clustering method for generating visual codebook. Instead of using k-means clustering, the authors introduced extremely randomized clustering forests. The experimental results showed that the new approach produces more accurate results and faster training.

Most existing approaches use single visual vocabulary. Multiple vocabularies were incorporated in (Perronnin, Dance, Csurka & Bressan 2006), where an image is characterized by a set of histograms of visual words counts. The training stage was performed on a universal vocabulary consisting of images from all classes. Next, the general vocabulary was adapted to vocabularies with class-specific data. Current approaches find it hard to distinguish between similar categories like cats and dogs for example. By representing images as sets of histograms, it was possible to increase the discriminative power of local features. However, the main limitation of this method is its high computational cost.

In (Quack, Ferrari, Leibe & Van Gool 2007) a new approach regarding detection of frequent and distinctive feature configurations is presented. It is based on existing data mining techniques and its aim is to reduce the number of clutter features returned by low-level feature extraction. The

evaluation though was performed on very small data collections.

(Ohbuchi & Furuya 2008) present an interesting application of "Bag of Visual Words" to 3D models retrieval. The initial sampling of local patches was performed on images of the model rendered from multiple view orientations. As expected, the computational cost of the algorithm was very high.

In (Sivic & Zisserman 2003) the local features model was implemented to retrieve similar objects outlined by a user in videos. The algorithm was applied to only two popular movies which makes it hard to asses the general usefulness of the approach.

Finally, (Athanasakos, Stathopoulos & Jose 2010) undermines the surprisingly good retrieval performance reported so far by various researchers, claiming that the methods were finely-tuned to collection specific properties. That is why it is so important to test the models on different data collections.

The visual features based on the Bag of Visual Words framework are currently the best candidate for large-scale generic image retrieval. They have some ability to recognize objects and generally outperform global models. Most of these methods, however, produce high dimensional vectors. This, in turn, requires the application of dimensionality reduction techniques which leads to higher computational cost. Moreover, most existing approaches (in the retrieval of generic real-life images) are based on interest points detectors, but quite often more discriminant image patches are generated at random. The higher the number of random sample points, the better the performance. This, however, comes at a price of higher computational and data storage cost.

Based on the above discussion, the open research questions arise: Can we develop low-dimensional but effective local features which could be utilized in generic large-scale image retrieval? Can we also increase the discriminative power of sampling methods without increasing the computational and data storage cost?

We propose novel visual features based on the Bag of Visual Words framework for large scale generic image retrieval. The proposed visual features utilize an easy to implement low dimensional descriptor and a hybrid sampling technique. The novel hybrid sampling enhances the discriminative descriptor's power by capturing both background and foreground properties of images.

### 2.2.3 Enhancement of Local Features Based on Correlation Between Visual Words

An interactive image retrieval model with adaptive similarity measure is introduced in (Rui, Huang, Ortega & Mehrotra 1998). The weights for adjusting the similarity measure (with respect to the image content representation - global features) are calculated according to the consistency of the vectors' components representing images collected from user relevance feedback. First, the representations of images deemed relevant by the user are stacked to form a matrix. Next, if a column contains elements with similar values then this particular dimension is considered to be a good indicator of the user's information need and the weight is calculated as an inverse of standard deviation across this dimension.

(Liu, Liu, Liu & Lu 2009) exploit co-occurrence information in the spatial domain. The researchers make an assumption that the related visual words would appear in a certain neighbourhood. They utilize the equivalent of $tf \cdot idf$ weighting scheme from text retrieval. Having obtained the information about the relationships, they use it to update the current query by weighting all the coefficients in the histogram. This leads to the normalization process which may hamper the system's performance (Yang et al. 2007).

Another approach (Zhang, Huang, Lu, Wen & Tian 2010) tries to capture the spatial relationships between pairs of visual words by building a visual word tree. The tree is generated by clustering interest points that co-occurred within some spatial distance. Latent Semantic Analysis (LSA) is then applied to compute the importance of each visual word to the given query, and the most important ones become the so-called topic words. The $tf \cdot idf$ weighting scheme and the topic words are then utilized to re-rank the images. This approach, although quite efficient in comparison with others, is not applicable to real user evaluation because of the computational cost (high dimensional Scale Invariant Feature Transform descriptor and costly LSA).

The model proposed in (Yuan, Wu & Yang 2007) utilizes data mining techniques to discover spatially co-occurrent patterns of visual words. The authors report limitations of standard codebook generation techniques (related to synonymy and polysemy of visual words) and propose a novel approach, which constructs a higher-level visual phrase lexicon consisting of groups of co-located visual words.

Spatial correlations are also exploited in (Savarese, Winn & Criminisi 2006) where they are represented by correlograms. Experimental results show, that the joint models (Bag of Visual Words and correlogram) outperform standard appearance-only models. However, models based only on correlograms perform worse than standard Bag of Visual Words approach.

(Jamieson, Dickinson, Stevenson & Wachsmuth 2006) propose to group features that exist within a local neighbourhood, claiming that arrangements or structures of local features are more discriminative. Such groups of visual words are then associated with annotation words.

A trigram model is proposed in (Wu, Li, Li, Ma & Yu 2007) to help in image classification. The method captures spatial correlations between image patches. Comparison between unigram and trigram models shows that the latter one improves the classification accuracy.

Another model (Zheng, Wang & Gao 2006) defines a visual phrase-based image similarity. First, the number of occurrences of each visual word is counted. Then the occurrences of adjacent patch pairs formed by frequent visual words are counted and finally, the visual phrases are generated by selecting the adjacent patch pairs whose occurrences are higher than a threshold. The similarity between two images is measured by the cosine similarity with $tf \cdot idf$ weighting scheme adapted from text retrieval.

In general, methods that utilize information about correlations between visual words try to group semantically similar visual words together. They usually consider co-occurrences at one contextual level only, and are computationally expensive and not scalable. Moreover, existing approaches (query modification like frameworks) modify the current query, which leads to the

normalization of histograms. This may not be desirable, since the (mid-level) semantic meaning of bins may be lost and the representations may become less discriminative due to the varied complexity of images. Some researchers report performance improvement after normalization, while others report the opposite. We believe that this may be also domain specific. The $tf \cdot idf$ weighting scheme may work better in case when the precise object matching is important. Our experiments, for example, showed that the normalization of histograms of visual words counts hampers the retrieval performance.

The further research questions are: Can we completely avoid the modification (and thus renormalization) of histograms of visual words counts in the context of local features enhancement based on the correlations between visual words? Can we develop a model that would make this enhancement feasible in large-scale generic image retrieval? Can we compare different notions of correlation at different contextual levels within the aforementioned framework? Can we introduce more intuitive notion of image-level correlation? We argue that the standard image level correlation is not intuitive enough. For example, if the frequencies of two pairs of visual words are $\{5, 10\}$ and $\{5, 100\}$ then the latter will be assigned higher correlation value. We would, however, expect the former pair to be at least equally correlated.

We propose a new approach for identifying and utilizing the information about correlations between visual words. We implement and test various notions of correlation at different contextual levels (we refer to them as image-level and proximity based). To the best of our knowledge, this is the first time these two are compared within this type of framework in image retrieval. Our local features consist of low dimensional histograms, where bins representing visual words are highly correlated. We identify the most and the least correlated coefficients and use the thus obtained information, along with the visual terms' frequencies from the current query, to weight the similarity measure. Certain coefficients in the similarity measure corresponding to the most correlated terms are then increased, while the coefficients related to the least correlated pairs are deemphasized.

The proposed method is computationally and data storage cheap, utilizes correlation at different contextual levels, and avoids the renormalization of histograms.

## 2.3 Combining Visual and Textual Information

The combination of visual and textual features is often based on various data fusion strategies. Some experimental results also hinted at the potential interchangeability of fusion approaches.

### 2.3.1 Quantum Theory Inspired Framework for Information Retrieval

Here, we list some works related to the application of mathematical tools from quantum mechanics to information retrieval (IR), as some of our models utilize them.

A theoretical formalism for modelling information retrieval tasks was first proposed by van Rijsbergen in (van Rijsbergen 2004). This book has recently inspired some interesting research

in IR, which provide various useful formalization and mathematical tools. It has been shown that many, seemingly heterogeneous IR models, can be encompassed by the single mathematical formalism which generalizes the standard IR Vector Space Model.

The research that followed concentrated on three main areas (Song, Lalmas, van Rijsbergen, Frommholz, Piwowarski, Wang, Zhang, Zuccon, Bruza, Arafat, Azzopardi, Buccio, Huertas-Rosero, Hou, Melucci & Ruger 2010): Spaces, geometrical representation and characterization of context by the concept of subspaces of a Hilbert space; Interferences between documents, topics and user's cognitive status in contextual relevance measurement process; Frameworks and operational methods for contextual and multimodal information retrieval.

(Piwowarski, Frommholz, Lalmas & van Rijsbergen 2010) discuss how an interactive probabilistic framework inspired by quantum theory and the mathematical tools of Hilbert spaces can be used to tackle contemporary challenges in IR.

The most important discoveries in the area of quantum theory inspired applications to IR, economy, psychology, etc. are presented at the annual Quantum Interaction (QI) conferences.

### 2.3.2 Hybrid Models

Most approaches that combine visual and textual features in Content-based Image Retrieval systems are fusion methods that would:

1. pre-filter the data collection by visual content and then re-rank the top images by text (Yanai 2003);
2. pre-filter the data collection by text and then re-rank the top images by visual content (Tjondronegoro et al. 2006);
3. pre-filter the data collection by visual (textual) content and then aggregate the scores of the textual (visual) representations of the top retrieved images (transmedia pseudo-relevance mechanism (Maillot et al. 2007));
4. fuse the representations (early fusion (Rahman et al. 2009));
5. fuse the scores or ranks (late fusion (Simpson et al. 2010)).

Pre-filtering by text and re-ranking by visual content is usually a well performing method. However, the main drawback of this approach is that the images without the textual description will never be returned by the system. Moreover, this type of pre-filtering relies heavily on the textual features and the assumption that the images are correctly annotated.

The most common early fusion technique is concatenation of visual and textual representations. The main drawback of the early fusion approach, however, is the well known curse of dimensionality. In the next chapters we show, that the curse of dimensionality can often be avoided as the similarity between the fused representations may be characterized as the combinations of similarities computed on individual feature spaces.

In case of late fusion, the most widely used method is the arithmetic mean of the scores, their sum (referred to as CombSUM), or their weighted linear combination. One of the best performing systems on the ImageCLEF2007 data collection, XRCE (Mensink, Csurka, Perronnin, Sánchez & Verbeek 2010), utilizes both (for comparison purposes) early (concatenation of features) and late (an average of scores) fusion approaches. Another common combination method, referred to as CombPROD in the literature (Depeursinge & Muller 2010), is the square of the geometric mean of the scores - their product. It has been argued, that the major drawback of the late fusion approaches is their inability to capture the correlation between different modalities (Mensink, Verbeek & Csurka 2011). However, in the next chapters we show, that in some cases the late fusion can be represented as early fusion.

Other combination methods involve a combination of late fusion and image re-ranking (Clinchant, Ah-Pine & Csurka 2011) (because the first step is the pre-filtering of the collection by text, the model is called semantic combination).

The fusion approach that can be easily modified to incorporate the user feedback is based on the transmedia pseudo-relevance mechanism. This so-called inter-media feedback query expansion is based on textual query expansion in most of the papers ((Depeursinge & Muller 2010),(Chang & Chen 2009)). Typically, textual annotations from the top visually-ranked images (or from a mixed run) are used to expand a textual query.

The ranking of documents is often computed by heuristically combining the feature spaces of different media types or combining the ranking scores computed independently from different feature spaces. All these combination methods treat the textual and visual features individually, and combine them in a rather heuristic manner. Therefore, this makes it difficult to capture the relationship between the features. Indeed, as both the textual and visual representations describe the same image, there are inherent correlations between them which should be incorporated into the retrieval process as a whole in a more principled way.

It is also commonly believed that the fusion strategies (utilized in existing hybrid models) represent different combination methods and that the drawback of early fusion is the potential curse of dimensionality, while the drawback of a late fusion is its inability to capture the correlation between the features dimensions.

The research questions we are going to look at include: Is it possible to show the meaningfulness and theoretical interpretation of some combination methods? It is relatively easy to combine the scores (for example) in such a way that the retrieval performance will be improved. It is more difficult, however, to show that the combination is meaningful and universal. Can we utilize the inherent correlations between visual and textual features in a hybrid model?

Inspired by Quantum Theory, we propose a quantum theory inspired multimedia retrieval framework. It is based on the tensor product of feature spaces, where similarity measurement between a query and a document corresponds to the quantum measurement. At the same time, the correlations between dimensions across different feature spaces can also be naturally incorporated in the framework. The tensor based model provides a formal and flexible way to expand the

feature spaces, and seamlessly integrate different features, potentially enabling multi-modal and cross media search in a principled and unified framework.

### 2.3.3 Image Auto-annotation

Image annotation is a broad research area, therefore it would be difficult to refer to all interesting papers. In general, we can classify image annotation techniques into three groups (see (Duygulu, Barnard, Freitas & Forsyth 2002)): recognition as translation, statistical models and combined approaches.

The first category of image annotation models may be compared to machine translation. Models try to predict one representation given another. Thus, (Duygulu et al. 2002) first performs image segmentation and then classifies the regions into corresponding "blobs" by utilizing k-means clustering. Next, the corresponding word for each blob is found by choosing the word with the highest probability computed by the Expectation Maximization algorithm. However, due to the segmentation process, this approach can be computationally expensive, and the segmentation techniques do not always perform well.

Some methods utilize information about the correlations between "visual words" (for more information about "Bag of Visual Words" approach see 3.2) and try to group semantically similar visual words together. Such subsets of visual words can then be associated with textual terms. These approaches usually consider only co-occurrences at the local level, are computationally expensive and not scalable. Thus, (Jamieson et al. 2006) propose to group features that exist within a local neighbourhood, claiming that arrangements or structures of local features are more discriminative. Such groups of visual words are then associated with annotation words.

Approaches that belong to the second category of image annotation models, usually cluster image representations and text. In this way a joint probability distribution may be generated that links images and words. Finally, the labels for images that have high posterior probability may be predicted. For instance, (Li & Wang 2008) exploit statistical relationships between images and words without recognizing individual objects in images. This real-time annotation method, according to authors, can provide more than 98% of the images with at least one correct annotation out of the top 15 selected words. However, the high number of labels assigned to the given image may introduce a lot of noise in the form of, for example, contradictory meaning.

In general, existing auto-annotation methods based on segmentation models or grouping of visual words are computationally expensive and not scalable to large data collections. Methods based on the clustering of visual and textual features, on the other hand, neglect the contextual information.

Then we are concerned with the following research question: How to develop a computationally cheap model that can capture the contextual information, and can be seamlessly integrated into our feature combination framework?

We propose two models for making the tag-image associations. The first proposed approach projects the unannotated images onto the subspaces generated by subsets of training images (con-

taining given textual terms). We calculate the probability of an image being generated by the contextual factors related to the same topic. In this way, we should be able to capture the visual contextual properties of images, taking advantage of this extended vector space model framework. The other method performs quantum like measurement on the density matrix of unannotated image, with respect to the density matrix representing the probability distribution obtained from the subset of training images. These approaches can be seamlessly integrated into our unified framework for image retrieval.

### 2.3.4 Hybrid Relevance Feedback Models

Due to the scarcity of hybrid relevance feedback models, we are going to focus on monomodal relevance feedback models (one feature only), and relevance feedback models that can be modified to incorporate both visual and textual features.

Relevance feedback can be utilized to narrow down and "correct" the search.

#### Relevance Feedback

(Middleton, Shadbolt & De Roure 2004) use both explicit and implicit feedback to recommend online academic research papers. Users' research interests are inferred from relevance feedback and browsing history.

Another approach incorporating explicit feedback (Chang, Cheng, Lai, Wu, Chang & Wu 2001) learns complex image-query concepts to improve the retrieval in a small number of iterations. The user marks the images relevant to his/her query concept and the system infers his/her information needs by using two classification techniques.

(Chu & Park 2009) propose a novel framework for personalized recommendation incorporating demographic information about the user, relevance feedback, click through history and other implicit data.

In (Wilson & Srinivasan 2005) the relevance feedback is used to automatically adjust visual features' weights in order to better model user perception. This is done within a Bayesian network architecture and the feedback is implemented via diagnostic inference.

The system presented in (Felden & Linden 2007) asks the user to provide some personal information, gathers implicit feedback, and creates ontology-based profiles.

(Berger, Denk, Dittenbach, Pesenhofer & Merkl 2007) also collect explicit information about the user's age, gender, etc. and their interests and preferences. They investigate the correlation between tourism related photographs and tourist activities.

The model presented in (Shah-Hosseini & Knapp 2004) incorporates implicit feedback (transaction logs) and relevance feedback to generate semantic net consisting of semantic classes of images. The transactions from the previous search are clustered with fuzzy clustering algorithm to create the classes.

(Yu, Ma, Tresp, Xu, He, Zhang & Kriegel 2003) introduce user profiles generated from user relevance feedback and low level visual features. The visual features are used to recommend new images to the user and learn his preferences (support vector machines) from explicit feedback (like/dislike). Then, the thus generated user profile is combined with the society of profiles (using collaborative ensemble learning) to further improve the art image retrieval.

(Albanese, Chianese, D'Acierno, Moscato & Picariello 2010) propose a model combining implicit feedback (usage logs) with visual features and semantic descriptors, to improve the image recommendations (art image retrieval).

An approach introduced in (Teevan, Dumais & Horvitz 2005) uses relevance feedback and usage patterns (previous queries, web pages visited so far, e-mails and documents created and read) to re-rank web search results. The authors emphasize the importance of the diversity and representativeness of data collected.

The system described in (Müller, Pun & Squire 2004) analyses the log files to learn feature weights from user behavior. The visual features consist of four different groups including colour histogram and Gabor filters for texture representation.

The Argo system (Wang, Yu, Zhang, Cai & Ma 2009) generates a user profile by analyzing the user's shared photo collection and then recommends relevant ads. The authors discover user interests by finding visually similar images to the ones in the user's collection (using low-level visual features) and extracting tags from them. The tags are then used to create an ontology (knowledge representation in user profile).

(Nanas, Vavalis & De Roeck 2010) propose a novel framework for user profile construction based on a document's content. Instead of using support vector machines, they introduce a network-based model for profile representation. This way they were able to avoid problems with high vector dimensionality.

(Castellano, Fanelli, Mencar & Torsello 2007) incorporate fuzzy clustering in order to generate user profiles. The clusters represent similar users' preferences created from the web log files.

#### Rocchio Algorithm and Probabilistic Relevance Feedback Models

One of the first, simplest, yet well performing relevance feedback models is the Rocchio algorithm (Rocchio 1971). It modifies the original query in order to shift it closer to the centroid of relevant documents and further away from the centroid of irrelevant ones.

There are also probabilistic relevance feedback models (Rijsbergen 1979). However, the early probabilistic models did not in general perform as effectively as the conventional vector modification approaches (Salton & Buckley 1997). Only recently, the Rocchio algorithm has been rewritten in a probabilistic form (Zhang, Hou & Song 2009). Interestingly, the probabilistic form of the Rocchio algorithm resembles the Rocchio algorithm in the late fusion form. In the thesis we show how the Rocchio query modification approach can be represented as a late fusion strategy.

**Hybrid Models for the Combination of Features in the Relevance Feedback Context**

The aforementioned Rocchio algorithm can be modified to incorporate both visual and textual relevance feedback. Indeed, this was done in (Lu, Zhang, Wenyin & Hu 2003).

There are a few existing prototype image retrieval systems that integrate the features in the context of relevance feedback (Ortega-Binderberger, Mehrotra, Chakrabarti & Porkaew 1999), (Quack, Mönich, Thiele & Manjunath 2004), (Kherfi, Ziou & Bernardi 2004). These models utilize visual and textual features in a sequential manner, where only one modality is exploited at each relevance feedback iteration step.

Approaches that utilize both modalities simultaneously (in the context of relevance feedback) would usually combine the features in a linear manner, e.g. (Chen, Wenyin, Zhang, Li & Zhang 2001).

The model presented in (Sclaroff, La Cascia & Sethi 1999) utilizes the user feedback to modify the weights in the linear combination of visual and textual features.

The aforementioned hybrid relevance feedback models, similarly to the hybrid models, combine the features in a rather heuristic manner. They do not exploit the inter and intra feature relationships intrinsic to both feature spaces. The intra feature relationships correspond to individual features spaces, while the inter feature relationships correspond to correlation and complementarity of both visual and textual feature spaces. (Jin, He & Tao 2008) tries to exploit these relationships in a hybrid relevance model. In their approach, Manifold Ranking Algorithm and Similarity Propagation Algorithm are integrated to explore the multiple relationships of web images. However, they model the inter relationships via web hyperlinks. We, on the other hand, want to exploit the relationships between the visual and textual subspaces of a query and feedback images.

We aim to address the following research question: How can we exploit the aforementioned inter and intra feature relationships in a hybrid relevance feedback model for a large scale generic image retrieval?

We propose a model for visual and textual features' combination within the context of relevance feedback. The approach is based on mathematical tools also used in quantum mechanics - the predicted mean value of the measurement and the tensor product of the density matrices, which represents a density matrix of the combined systems. It was designed to capture both intra-relationships between features' dimensions (visual and textual correlation matrices) and inter-relationships between visual and textual representations (tensor product). The model provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context by applying quantum-like measurement. It opens a door for a series of theoretically well-founded further exploration routes, e.g. by considering the interference among different features. It is easily scalable to large data collections as it is general and computationally cheap.

### 2.3.5 Importance of Query and Its Context

Similarly to models which combine the features in the context of user feedback, there is not much research on the dynamic combination of features from the original query and features from its context (i.e. feedback images). We are going to mention two (text only) approaches that utilize adaptive weighting schemes.

(Wu, Xing, Li & Bi 2012) implement an adaptive data fusion method with dynamically adjustable weights. They investigate two methods for the weight updating, namely "performance square" updating and its mixture with linear regression analysis. Experiments conducted on the benchmark showed that both adaptive weights models outperformed the CombSUM fusion method. They combine evidence from different sources but do not incorporate any user feedback.

(Wang, Yang, Qi, Li & Zhao 2012) proposed an adaptive weighting approach to improve the current statistical context-sensitive retrieval model. They first investigate the so-called "potential for adaptability", the performance gap between the context-sensitive model with fixed weights and the one with adaptive weights, to show that the system can really benefit from having query-specific weights. They apply the support vector regression to build a weight-prediction model, which enables a more flexible combination of current query and its context.

In general, a query can be more or less related to its context. Existing models which try to make the weights adaptive, require the training of these weights. Moreover, there is a lack of approaches that would incorporate adaptive weighting schemes into hybrid relevance feedback models.

The following research questions arise: Can we develop an approach, based on the relationship strength between the query and its context, that does not require the training phase? How can we incorporate the adaptive weighting scheme in a hybrid relevance feedback model?

We incorporate an adaptive weighting scheme into our hybrid CBIR relevance feedback model. Thus, each query is associated with unique set of weights corresponding to the relationship strength between visual query and its visual context as well as the textual query and its textual context. The higher the number of terms or visual terms (mid-level features) co-occurring between current query and the context, the stronger the relationship and vice versa. If the relationship between a query and its context is weak, context becomes important. We adjust the probability of the original query terms, and the adjustment will significantly modify the original query. If the aforementioned relationship (similarity) between query and its context is strong, however, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

Thus, we show how to measure the relationship strength between query and its context, and how to incorporate the adaptive weighting scheme into the state-of-the-art existing model (hybrid, user feedback context) to further improve the retrieval. The proposed adaptive weighting approach is relatively easy to implement and does not require any training of features.

## 2.4 Prototype Hybrid Relevance Feedback Systems. Interactive User Interfaces

A few prototype systems utilize both visual and textual information in a hybrid relevance feedback model. These models, however, combine the features in an ad hoc manner. No inter or intra feature relationships are modelled within the proposed frameworks.

There are many monomodal (only visual content or only text) interactive systems. Some of them support deep interaction with the user, i.e. various degrees of relevance, exploratory search, positive and negative results, query history.

Thus, Fire (Deselaers, Keysers & Ney 2005) is a prototype system based purely on visual features. It presents both positive and negative results to the user. An importance of providing both negative and positive examples for relevance feedback was suggested in (Heesch, Yavlinsky & Rüger 2003), (Pickering & Rüger 2003), (Muller, Muller, Marchand-Maillet, Pun & Squire 2000).

(Liu, Zagorac, Uren, Song & Rüger 2009) developed an interface that provides such functionalities as querying by positive and negative visual examples, presenting positive and negative results, query history. The user can also specify the weight denoting the relevance of an image from the result list. Inputting a relevance weight by simply typing the number is not very intuitive though. The prototype system comprising the aforementioned user interface is based on a purely visual image retrieval.

Existing hybrid relevance feedback prototype systems do not support a user-system interaction at such high level as some monomodal approaches.

The system presented in (Chen et al. 2001) integrates both the visual and textual features by combining them linearly. The positive initial results are presented to the user, who can then give the binary feedback (relevant/irrelevant). The probabilistic model utilizing only textual information is employed to narrow down the search.

WebMARS (Ortega-Binderberger et al. 1999) combines the visual and textual features in the context of relevance feedback. The features (colour, text and keywords) are aggregated as weighted combinations. The model requires many arbitrarily chosen weights. The user is presented with positive feedback only and the given feedback is based on the binary relevance.

Cortina (Quack et al. 2004) also combines the features in a simple way as a linear combination. There is not much room for the user system interaction. No exploratory search is supported and the relevance judgment is binary.

The prototype system presented in (Sclaroff et al. 1999) utilizes a simple user interface and lets the users select relevant/irrelevant images for the search refinement. The relevance feedback is exploited to modify the weights in a linear combination of visual and textual features.

In general, existing hybrid relevance feedback prototype systems do not offer as much functionality as the monomodal systems in terms of interaction with the user.

Interactive user interfaces can help us facilitate user-system interactions and fully exploit the implemented models. We need a hybrid relevance feedback prototype system that offers all the

functionalities of some monomodal approaches (and more, i.e. a natural way of incorporating various degrees of relevance and exploratory search). We have developed such an interface and fully integrated it with our hybrid models.

Figures 2.4 and 2.5 present our prototype system at work.

## 2.5 Chapter Summary

From the literature survey of visual methods, we can see that existing state-of-the-art in Content-based Image Retrieval is based on the Bag of Visual Words framework (local features). The problem with local features, however, is their computational and data storage cost which is due to their high dimensionality. This is one of the reasons why the local features are not in commercial use in generic image retrieval. Moreover, most sampling methods are based on the interest points detectors which are good for object recognition tasks but not necessarily good for generic image retrieval (detected points belong to the foreground). Because of these issues, existing local features are not best suited for large-scale generic image retrieval.

The local features based on the Bag of Visual Words framework disregard the information about correlations between visual words. Existing models that exploit these relationships often utilize only one notion of correlation, and there are no comparisons between different notions of correlation. Moreover, the standard image level correlation can be enhanced to better reflect the dependencies between visual words. Existing approaches (query modification like frameworks) often modify the current query, which leads to the renormalization of histograms. This may not be desirable, since the (mid-level) semantic meaning of bins may be lost and the representations may become less discriminative due to the varied complexity of images.

The literature review on visual and textual features combination strategies shows that different fusion strategies are widely utilized for this purpose. Some experimental results also hinted at the potential interchangeability of various fusion methods. It is commonly believed, however, that the fusion strategies represent different combination methods and that the drawback of early fusion is the potential curse of dimensionality, while the drawback of a late fusion is its inability to capture the correlation between the features dimensions. In multimedia information retrieval, where a document may contain textual and visual content features, the ranking of documents is often computed by heuristically combining the feature spaces of different media types or combining the ranking scores computed independently from different feature spaces. All these combination methods treat the textual and visual features individually, and combine them in a rather heuristic manner. Therefore, this makes it difficult to capture the relationship between the features. Indeed, as both the textual and visual representations describe the same image, there are inherent correlations between them which should be incorporated into the retrieval process as a whole in a more principled way.

We propose a tensor-based model for visual and textual features combination that requires fully annotated data collections. Hence, the next part of the literature survey is on the image auto-

annotation methods. In general, we can classify image annotation techniques into three groups: recognition as translation, statistical models and combined approaches. Existing auto-annotation methods based on segmentation models or grouping of visual words are computationally expensive and not scalable to large data collections. Methods based on the clustering of visual and textual features, on the other hand, neglect the contextual information. We need a computationally cheap model that can capture the contextual information, and can be seamlessly integrated into the tensor-based framework. In our work, we utilize three different fast approaches for making associations between feature dimensions.

Because not much research on hybrid relevance feedback models exists, the literature survey here is focused on monomodal models (one feature). Existing hybrid relevance feedback models do not exploit the inter (visual-textual) and intra (visual-visual;textual-textual) feature correlations. Most of the monomodal models expand or modify the query. In the following chapters we will show that the query modification models can be represented as a late fusion strategy.

Relevance feedback models often utilize arbitrary weights corresponding to the original query and the feedback images (context). However, a query can be more or less related to its context. Based on the related work, we can see that a few existing models which try to make the weights adaptive, require the training of these weights. We propose to adapt the weights according to the relationship strength between query and its context. Moreover, we incorporate adaptive weighting in the hybrid relevance feedback model.

A brief overview of the interactive user interfaces inspires the interface design that facilitates our proposed hybrid model for combining textual and visual feature spaces in CBIR and in relevance feedback. Our prototype system with interactive user interface exploits combination of features, combination of features in the relevance feedback context, various degrees of relevance, positive and negative results, and exploratory search.

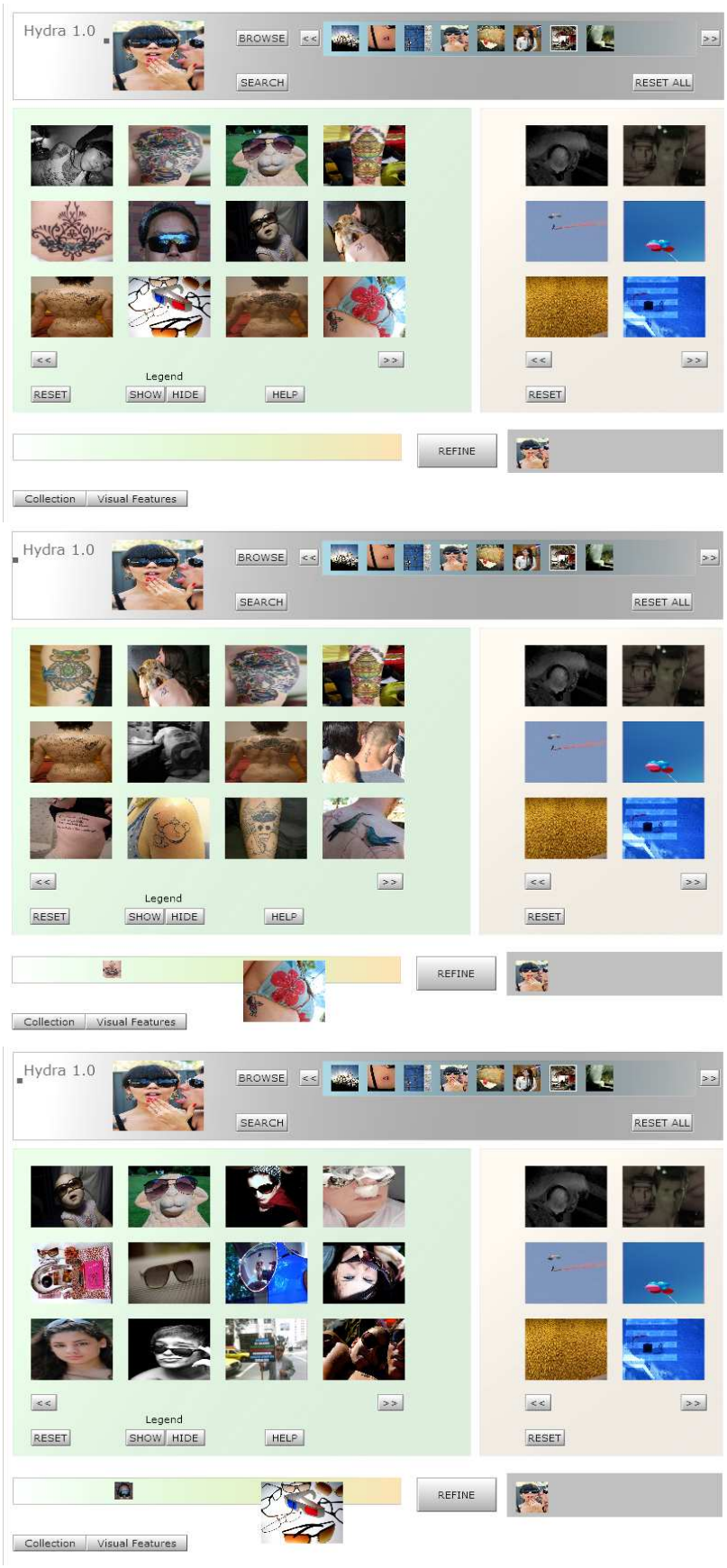


Figure 2.4: Narrowing down the search (combination of textual and visual features). Top: Users A and B query the system by visual example. The system identifies a few concepts and displays the results. Middle: User A specifies his interests by giving the feedback (tattoos) and refines the search. Bottom: User B specifies his interests by giving the feedback (sunglasses) and refines the search.

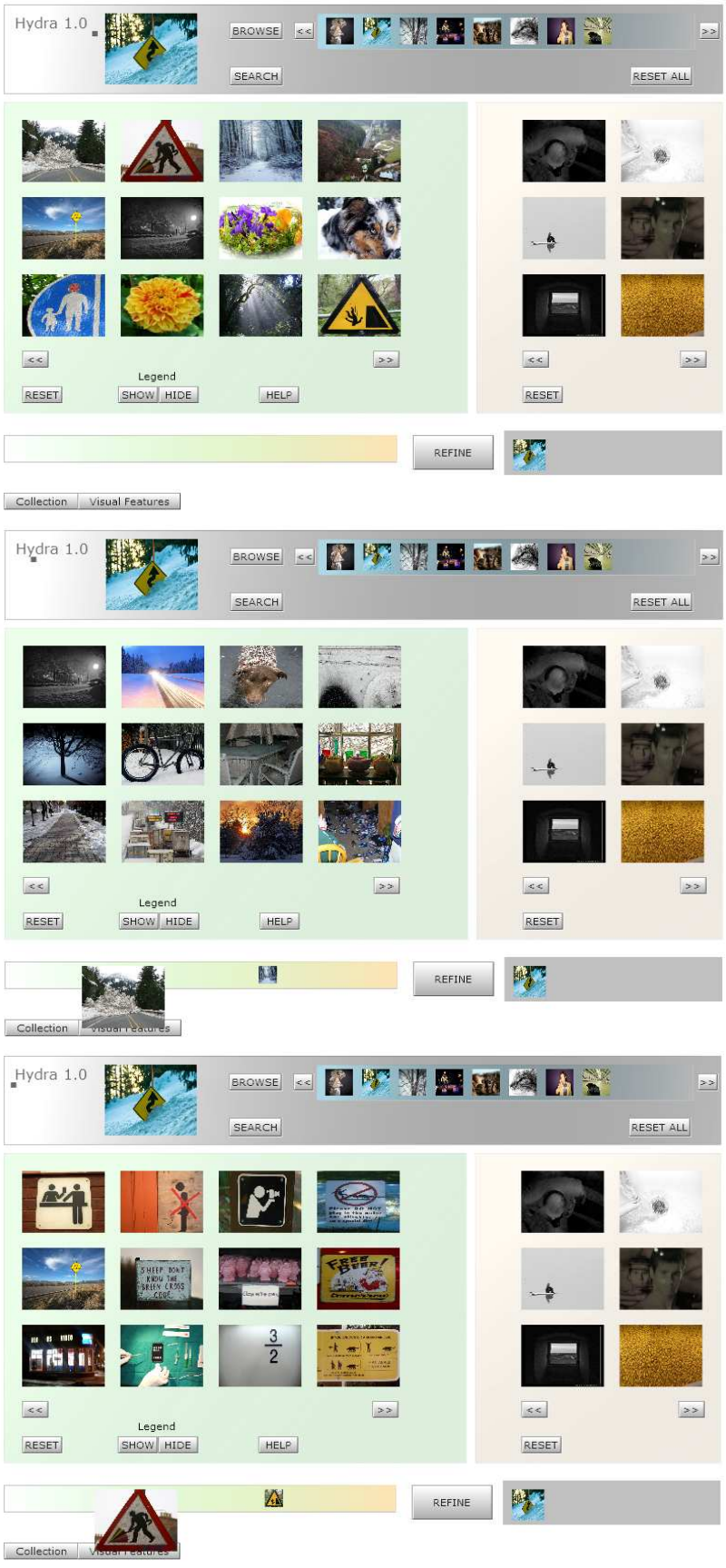


Figure 2.5: Narrowing down the search (combination of textual and visual features). Top: Users A and B query the system by visual example. The system identifies a few concepts and displays the results. Middle: User A specifies his interests by giving the feedback (snow) and refines the search. Bottom: User B specifies his interests by giving the feedback (signs) and refines the search.

# Chapter 3

# Novel Visual Features for Generic Image Retrieval

The semantic gap, the difference between human perception and machine representation of images can be reduced by improving the visual representations.

In this chapter, we present novel global and local features designed for generic image retrieval and enhancement of the local features based on the correlation between visual words. Large-scale generic image retrieval should employ visual features that are fast and effective. Existing state-of-the-art features (local features) are still computationally and data storage costly. Therefore, we aim at developing visual features that will be comparable with current state-of-the-art and will make the large-scale image retrieval more feasible and more effective.

Thus, this chapter presents our proposed methods. The experimental results are reported in Chapter 6.2.

## 3.1 Novel Global Methods

Global visual features, as the name suggests, characterize the visual content of images as a whole. Here, we introduce novel global features based on both Canny (a state-of-the-art edge detector) and the proposed, novel edge detector. These global features, apart from capturing different image properties, allow us to retrieve images of similar complexity. This observation can be utilized to develop better normalization techniques for the Bag of Visual Words framework. In the CBIR Bag of Visual Words framework, as opposed to text IR, normalization may even hamper the retrieval due to the different complexity of images (this may result in less discriminative representations).

### 3.1.1 Edge Detectors and Co-occurrence Matrix

The purpose of detecting edges in images is to capture important events and changes in image properties. This is also a method that allows us to reduce the amount of information in images,

keeping the most important structural properties only. Here, we present a novel edge detector. The aim was to develop a detector that would be efficient, fast and easy to implement.

The novel edge detection model consists of the following steps:

1. Bilateral filtering
   Some edge detection approaches utilize Gaussian smoothing to remove noise from images. Gaussian smoothing (blurring) averages the pixel intensities with respect to their neighbourhood. The problem with Gaussian smoothing, however, is that it also distorts the edges. Bilateral filtering alleviates this problem by utilizing kernels (masks) of different shapes. Thus, bilateral filtering is the kind of filtering that preserves the edges. Bilateral filtering at work can be seen in Figure 3.1.

Figure 3.1: Bilateral filtering. Left - original image. Right - filtered image

2. Directional image derivatives (four orientations) and their aggregation
   An intuitive idea is to expose the edges by shifting an image by one pixel distance in four different directions, then subtract the shifted images from the original one. Thus obtained images (directional derivatives) can then be aggregated. The shift orientations are depicted in Figure 3.2.

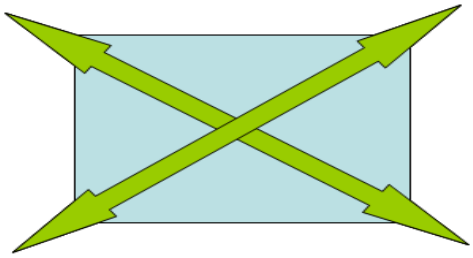

Figure 3.2: Image shift orientations

3. Pixel intensities thresholding

   The final step in our edge detection approach is the pixel intensities thresholding. We only keep the pixels that are below a certain threshold level.

There is no objective way to compare the edge detectors. We show examples of the novel detector at work (3.3, 3.4) for direct visual comparison with more sophisticated approaches, i.e. (Wickerhauser & Czaja 2004). Other examples of the novel edge detector are depicted in Figure 3.5 and 3.6.

Figure 3.3: Novel edge detector (top-right) versus the localized two-dimensional Fourier transform-based approach (bottom)

Having detected the edges in the images in the data collection, we need to represent them (their complexity, distribution) in a vector form. We utilize an eight orientation co-occurrence matrix for this purpose. The one orientation co-occurrence matrix is defined as

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases} 1, & if \quad I(p,q)=i \quad and \quad I(p+\Delta x, q+\Delta y)=j \\ 0, & otherwise \end{cases} \tag{3.1}$$

where $I(p,q)$ is an intensity of a pixel $(p,q)$.

The co-occurrence matrix describes the way certain grayscale pixel intensities occur in relation to other grayscale pixel intensities. It counts the number of such patterns. An enhanced version of co-occurrence matrix takes into account not only one, but more orientations. Moreover,

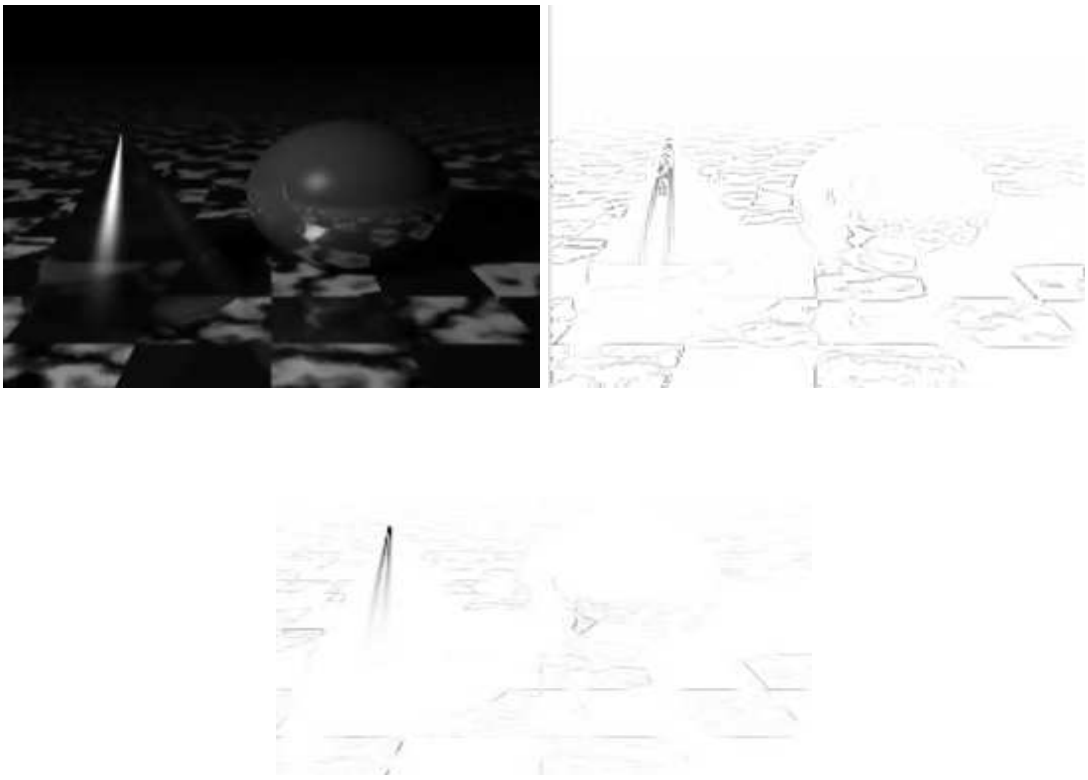

Figure 3.4: Novel edge detector (top-right) versus the localized two-dimensional Fourier transform-based approach (bottom)

various distances between pixels may be considered. These additional aspects would make the co-occurrence matrix partially invariant to rotation. The most discriminative statistics extracted from co-occurrence matrix are: contrast, inverse difference moment, entropy, energy, homogeneity, and variance.

Another implemented version of this model utilizes Canny edge detection technique. This enables us to test different edge detectors in the context of image retrieval.

### 3.1.2 Edge Detectors and Three Moments

We also exploit very simple statistics to capture the information about the detected edges. The aforementioned statistics are mean, standard deviation and skewness (the measure of asymmetry in the distribution).

Another implemented version of this model utilizes Canny edge detection technique.

## 3.2 Local Methods

Local features characterize local properties of images and have some ability to recognize objects. They were inspired by the Bag of Words from text retrieval and are considered mid-level visual

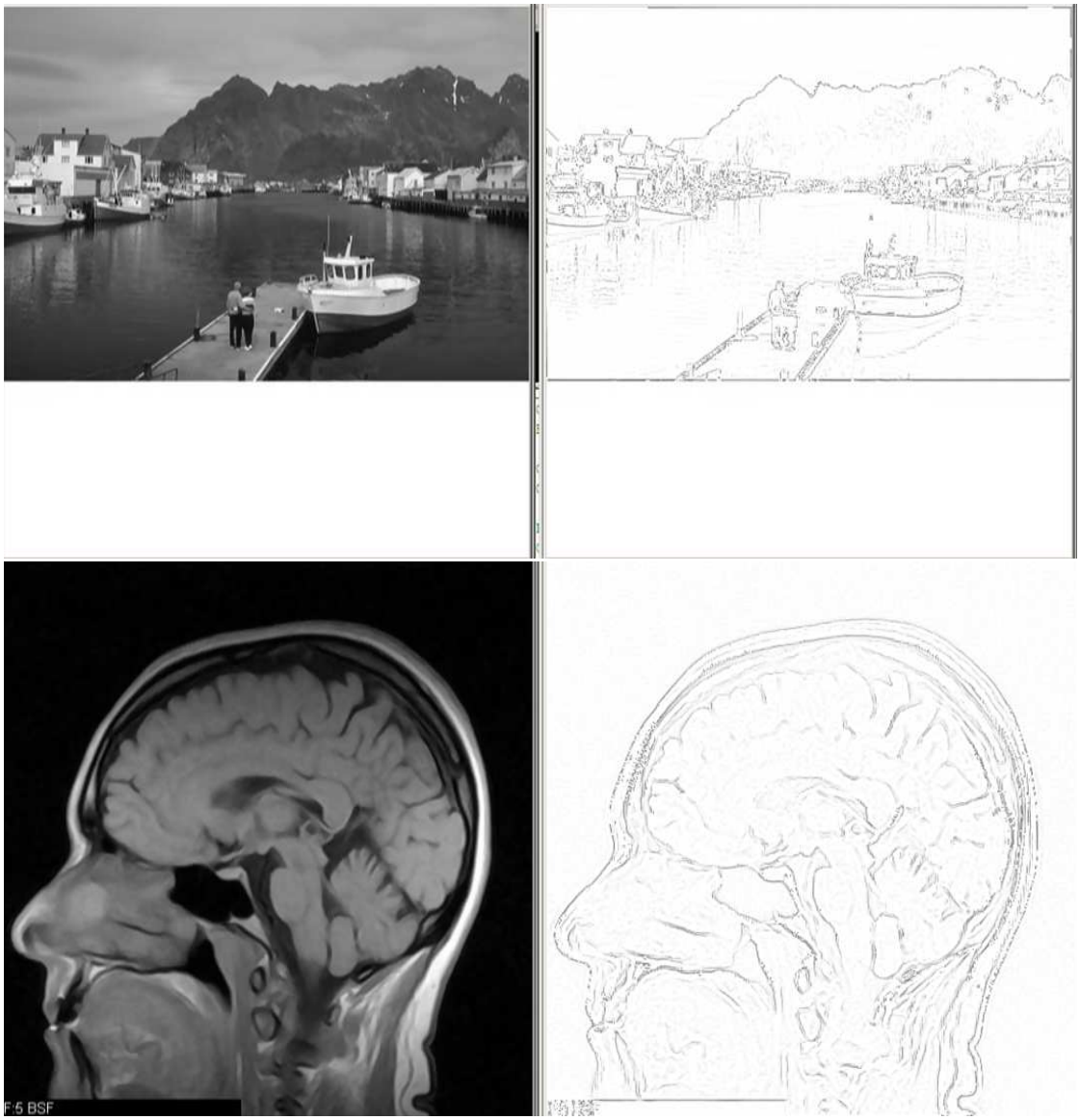

Figure 3.5: Novel edge detector at work

features. As high-level visual features are currently too computationally expensive (only applied to domain-specific tasks with a limited number of semantic concepts), local features represent up-to-date state-of-the-art in generic CBIR.

Even the standard Bag of Visual Words features are costly in terms of computation and data-storage as they produce high dimensional vectors. This is one of the reasons why the local features are not in commercial use yet. Thus, visual features for large-scale image retrieval need to address this problem.

Another problem is that local features with standard detector-based sampling are not particularly suited for generic image retrieval. Detector-based sampling concentrates on the foreground, which is beneficial for object recognition tasks. However, in generic image retrieval background

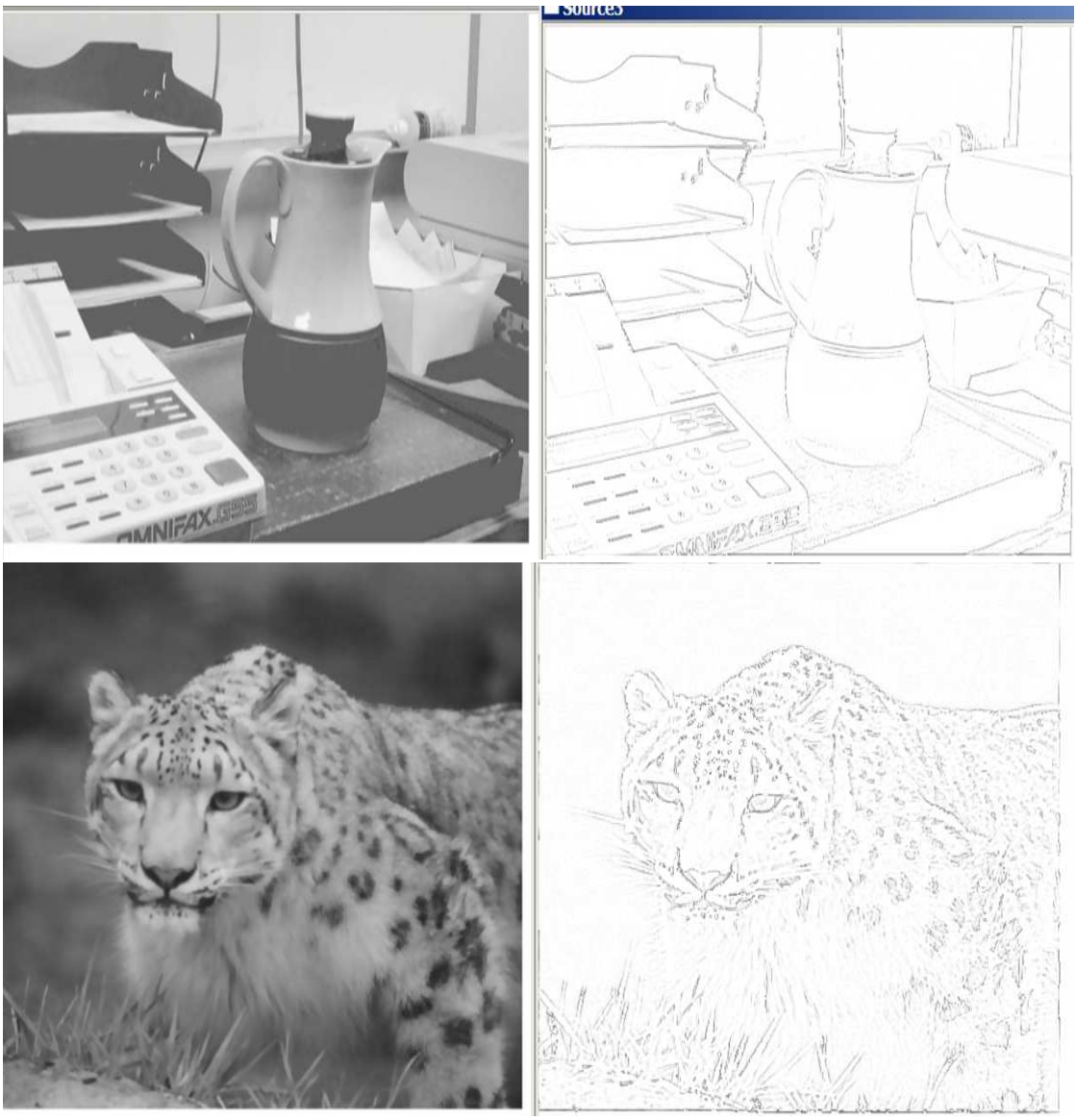


Figure 3.6: Novel edge detector at work

information is also very important as it often conveys the visual context[1] (different meaning of a foreground information depending on the background). Moreover, the set of sample points identified by a detector can get saturated, which would result in the detector's inability to produce more discriminative sample points. It has been shown, and our experiments confirm this finding, that random sampling works best if the number of sample points is high. This, however, comes at a price of high computational and data storage cost.

Our goal is to make the local features scalable to large data collections. We also make the sample points more discriminative without rising the computational and data storage cost.

Our model comprises the following stages:

[1] Visual context - background visual information.

- **Image sampling:** Apply Shi and Tomasi method (see (Shi & Tomasi 1994)) combined with a random points generator to sample an image. This will generate a sparse representation of an image in the form of a set of sub-images (local patches).

- **Description of local patches:** Characterize local patches in the form of co-occurrence matrix or colour moments. In the case of a co-occurrence matrix, extract the meaningful statistics - energy, entropy, contrast, and homogeneity. Compute the features for individual colour channels.

- **Feature vector construction:** Represent local patterns as colour moments or statistics calculated from the co-occurrence matrix.

- **Visual dictionary generation:** Apply clustering (in our work K-means) to the training set in order to obtain the codebook of visual words.

- **Histogram computation:** Create a histogram of visual word counts by calculating the Manhattan distance between image patches and cluster centroids and generate a vector representation for each image.

- **Similarity measurement:** Measure the distance between multidimensional vectors by using Minkowski's fractional similarity measure. It has been shown to perform well in CBIR (Liu, Song, Rüger, Hu & Uren 2008).

### Image Sampling

The sampling technique can have a significant influence on the retrieval performance. Examples of dense and detector-based sampling are depicted in Figure 3.7. Here, we introduce a hybrid sampling which combines Shi and Tomasi corner detection with a random number generator. The Shi and Tomasi method is based on the Harris corner detector (Figure 3.8; adapted from (Collins 2011)).

We apply the Shi and Tomasi detector but generate each second sample point at random. This method will take into account the properties of the foreground as well as the properties of the background which are especially important in the retrieval of images depicting natural scenes.

### Descriptors

We characterize each local patch in an image as an eight orientation co-occurrence matrix and three colour moments in HSV colour space.

The co-occurrence matrix describes the way certain gray-scale pixel intensities occur in relation to other grayscale pixel intensities. It counts the number of such patterns. The most discrimi-

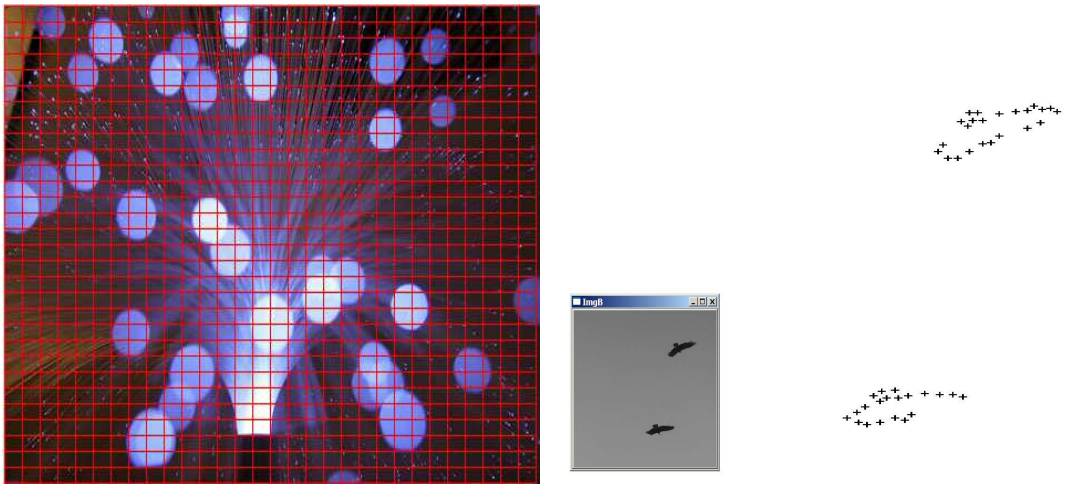

Figure 3.7: Dense and detector-based sampling

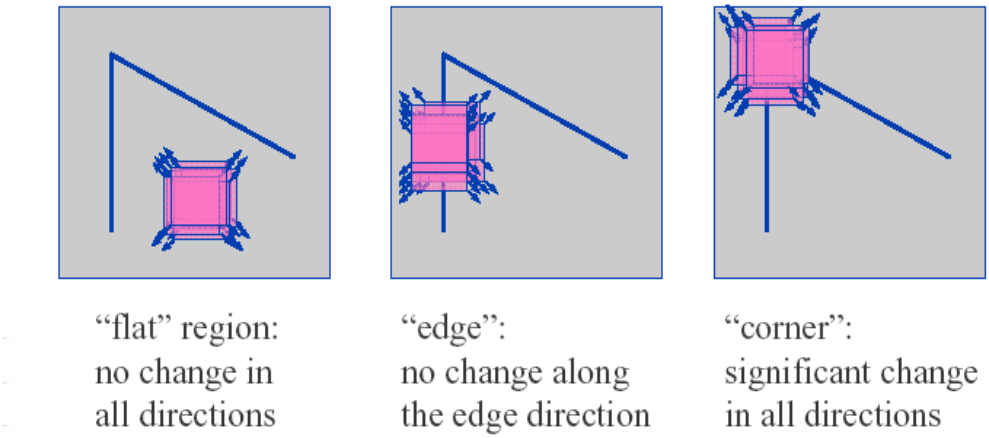


Figure 3.8: Harris corner detector

nating statistics extracted from the co-occurrence matrix are: contrast, inverse difference moment, entropy, energy, homogeneity, and variance.

The method based on three colour moments assumes that the distribution of colour can be treated as a probability distribution. Three statistics extracted from individual colour channels are mean, standard deviation and skewness

- Mean $E_i = \sum_{j=1}^{n} \frac{1}{N} p_{ij}$
- Standard Deviation $\sigma_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^2\right)}$
- Skewness $s_i = \sqrt[3]{\left(\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^3\right)}$

where $p_{ij}$ denotes the $ij$-th pixel intensity, and $N$ denote the number of pixels in an image.

The first moment can be interpreted as an average colour value, the second as the square root of the variance of the distribution, and the third as the measure of asymmetry in the distribution. Colour moments can also capture the textural properties of an image and are fairly insensitive to viewpoint changes. By computing them in HSV colour space we can make the statistics insensitive to illumination changes.

Figure 3.9 shows the novel local features at work.

Our approach is easy to implement, not sophisticated, with low computational and data storage cost (mostly because our vectors are low dimensional), and the hybrid sampling technique can be used in other methods based on the "Bag of Visual Words" to improve the retrieval performance.

## 3.3 Enhancing Local Features by Exploiting Correlation Between Visual Words

One of the drawbacks of the Bag of Visual Words framework is the assumption that visual words are independent of each other. This is usually not the case and the bins in the histograms of visual words counts are highly correlated. These correlations can be exploited to enhance the Bag of Visual Words framework.

Other models that try to utilize the relationships between visual words are often based on the query modification and do not compare different notions of correlation. However, query modification leads to representation renormalization which may hamper the retrieval. Some researchers report performance improvement after normalization, while others report the opposite. This may be due to the different complexity of images. We believe that this may be also domain specific. The $tf \cdot idf$ weighting scheme may work better in cases where the precise object matching is important. In our case, the normalization hampers the retrieval performance. However, we can completely avoid representation normalization by adapting the similarity measure based on the most and least correlated visual words. Moreover, we introduce and test different notions of correlation within the proposed framework. The introduced notions of correlation can also be utilized in various approaches which exploit co-occurrence matrices.

First, we generate a matrix of correlations between visual words for each top image returned in the first round retrieval. Second, we aggregate the matrices and identify the most and the least correlated coefficients. The thus obtained information, along with the visual words' frequencies from the current query, is then utilized to weight the similarity measure. Certain coefficients in the similarity measure corresponding to highly correlated terms are then increased, while the coefficients related to the least correlated visual words are deemphasized. The images returned in the first round retrieval are then re-ranked according to the modified similarity measure.

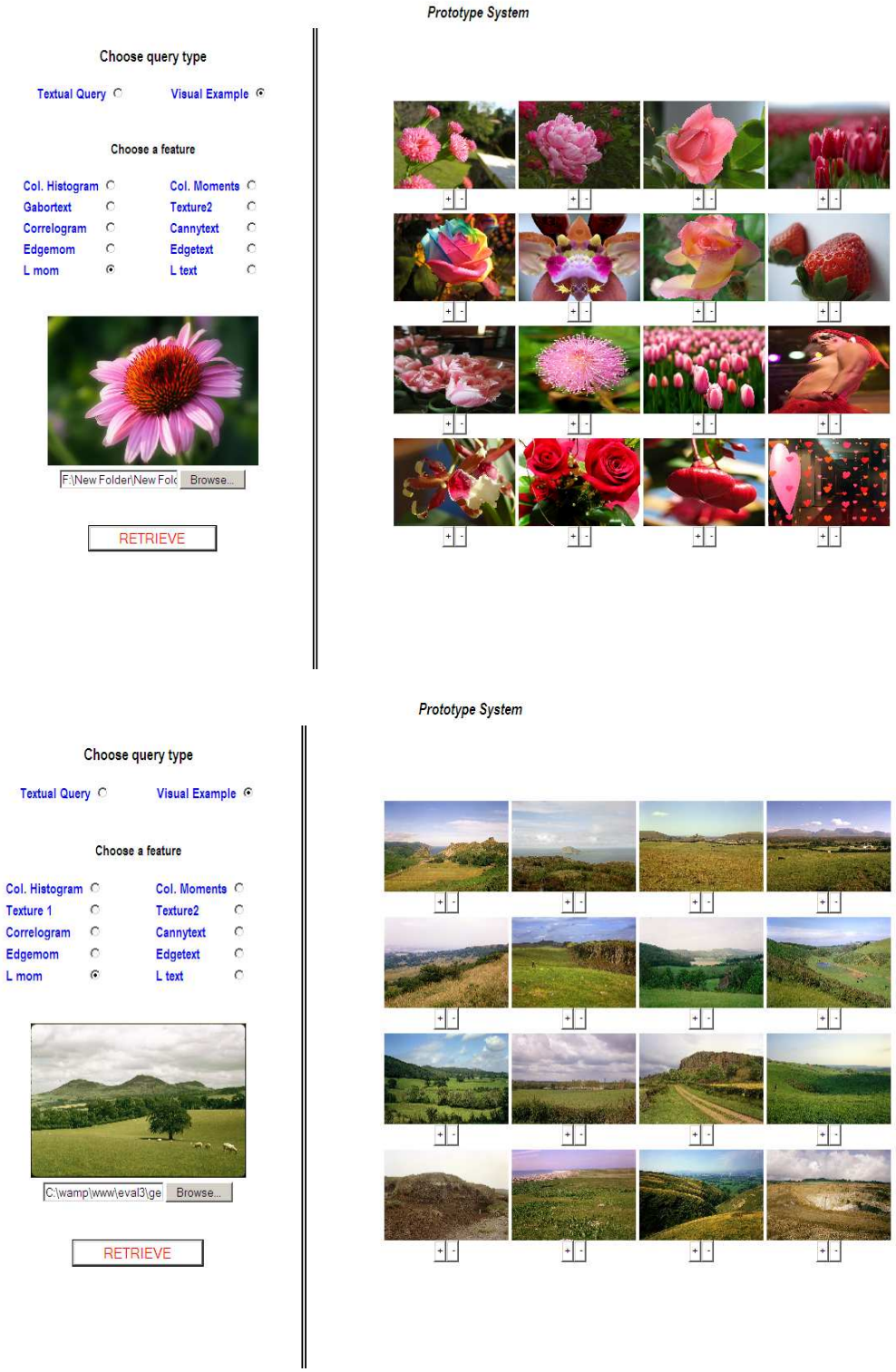


Figure 3.9: Novel local features at work

**Image-level Notions of Correlation**

Let us start by introducing image-level and proximity-based notions of correlation. In text retrieval document-level correlations seem to be stronger (see (Biancalana, Lapolla & Micarelli 2009)). A document may contain correlated terms not because of their proximity, but because they refer to the same topic.

Because our histograms of visual words' counts can be classified as a mid-level representation (the Bag of Visual Words reduces the semantic gap), we can introduce the correlations in a relatively intuitive way. Let us first focus on the correlations at the image level.

Correlation 1 can be regarded as the number of all pairs between the instances of different visual words (see Figure 3.10). Here, for instance, the squares denoted as A represent different instances of the same visual word (image patches) that appears within an individual image. When dealing with a set of images, we would aggregate the correlation matrices generated for each image. In case of Correlation 1, this would be equivalent to putting histograms of visual words counts as rows in a matrix and multiplying the transposition of this matrix by itself. This is an analogy to document-level correlation in text IR. Correlation 2 is a normalized version of Correlation 1, where the denominator is a total number of all possible pairs between occurrences of visual words (Figure 3.11).

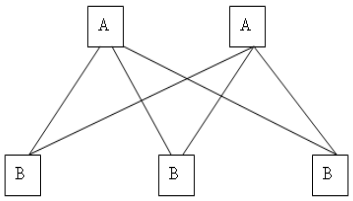


Figure 3.10: Interpretation of Correlation 1. This is the common document/image level correlation. Here, squares denote instances of visual words (image patches) and the links the relationships between them

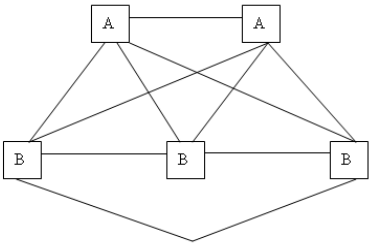


Figure 3.11: Normalization factor in Correlation 2. Here, squares denote instances of visual words (image patches) and the links the relationships between them

Correlation 3 can also be regarded as the number of pairs between the occurrences of different visual words, but this time the correspondence is as follows (see Figure 3.12).

Apart from the standard image/document level notion of correlation (correlation 1), we propose the following

1. $corr(vt_i, vt_j) = f_i \cdot f_j$

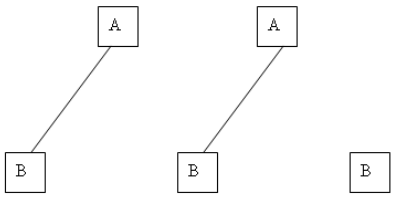


Figure 3.12: Interpretation of Correlation 3. Here, squares denote instances of visual words (image patches) and the links the relationships between them

2. $corr(vt_i, vt_j) = \frac{2 \cdot f_i \cdot f_j}{(f_i+f_j)\cdot(f_i+f_j-1)} = \frac{f_i \cdot f_j}{\binom{f_i+f_j}{2}}$

3. $corr(vt_i, vt_j) = \min(f_i, f_j)$

4. $corr(vt_i, vt_j) = \frac{f_i \cdot f_j}{\binom{f_i+f_j}{2}} + \min(f_i, f_j)$

where $vt_i$, $vt_j$ denote the $i$th and $j$th visual term respectively, and $f_i$, $f_j$ denote the frequencies (number of occurrences) of the terms. By calculating the correlations between all visual words in a particular image, we will obtain a matrix of correlations:

$$\begin{pmatrix} corr(vt_1, vt_1) & corr(vt_1, vt_2) & \dots & corr(vt_1, vt_n) \\ corr(vt_2, vt_1) & corr(vt_2, vt_2) & \dots & corr(vt_2, vt_n) \\ \vdots & \vdots & \dots & \vdots \\ corr(vt_n, vt_1) & corr(vt_n, vt_2) & \dots & corr(vt_n, vt_n) \end{pmatrix} \tag{3.2}$$

The matrix corresponding to the first notion of correlation can also be obtained by calculating the inner product of a transposed vector image representation and itself $h^T \cdot h$.

At first, there does not seem to be much difference between these three relationships. A closer look will show us the contradictions within our intuition of correlation.

Let us focus on Correlation 1. If the frequencies of two pairs of visual words are $\{5, 10\}$ and $\{5, 100\}$ then the latter will be assigned a higher correlation value. We would, however, expect the former pair to be at least equally correlated.

Normalization (Correlation 2) helps to overcome the above issue. However, if the frequencies are proportional, for example $\{10, 20\}$ and $\{40, 80\}$ then the former will score higher. But, intuitively, the latter is more correlated.

Correlation 3 seems to be intuitively right, but will ignore the additional information from the frequencies (see example for Correlation 1). Normalization of correlation 3 will produce similar side effects to Correlation 2. Therefore, we introduce Correlation 4, which does not seem to

contradict our intuition. Experimental results confirm the superiority of this notion of correlation in the user simulation.

**Proximity-based Notion of Correlation**

The above notions of correlation consider two instances of visual words to co-occur if they appear somewhere within an image (visual context - the whole image). Let us now introduce, by analogy to text retrieval, what we will refer to as proximity-based correlation. Two instances of visual words will be considered correlated if they appear together within a certain neighbourhood (visual context - "sliding window"). In case of dense sampling this is rather straightforward. When dealing with sparse sampling, however, we need to shift the window (square, circular) from one instance of a visual word to another. Figure 3.13 shows an example of proximity-based correlation. Here, the squares denote instances of various visual words. Now we can show how to

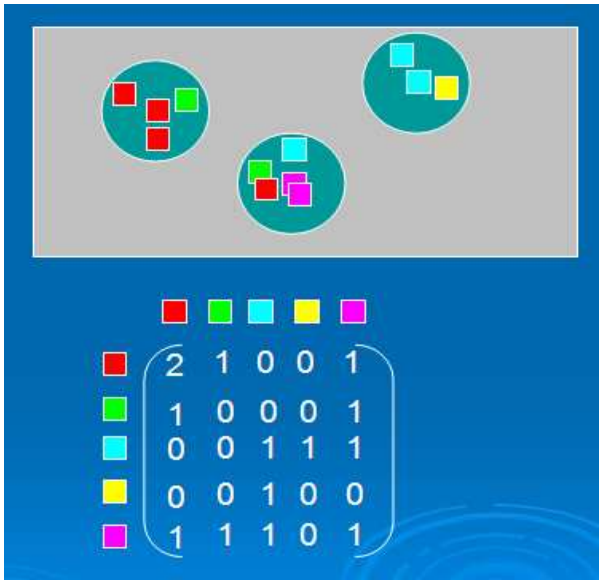


Figure 3.13: Proximity-based correlation. For the clarity of presentation, the matrix corresponds to only three instances of visual words (circles' centres)

incorporate the information about correlations into the Pseudo Relevance Feedback (PRF). PRF assumes that the top documents from the first round retrieval are all relevant to the query. Then, the additional information from the top documents is usually utilized to expand the query.

**Least and Most Dominant Visual Words**

Initially, the first round retrieval is performed. Then, for each image from the top returned images, the matrix of correlations will be created. We aggregate all the matrices in order to obtain the final matrix from which the most and least dominant correlations will be identified (in terms of values). Notice, that in case of Correlation 1, this approach would be equivalent to constructing

a matrix with rows corresponding to each image representation (from the top returned images)

$$M = \begin{pmatrix} f_1^1 & f_2^1 & \dots & f_n^1 \\ f_1^2 & f_2^2 & \dots & f_n^2 \\ \vdots & \vdots & \dots & \vdots \\ f_1^m & f_2^m & \dots & f_n^m \end{pmatrix} \tag{3.3}$$

and multiplying $M^T * M$, where $T$ denotes the transpose operation. The advantage of our method is that it does not restrict us to one notion of correlation and we can define it in a more intuitive way. Figure 3.14 shows an example of the aforementioned decomposition. It can be seen that the standard notion of image level correlation is an aggregation of a number of self-correlations corresponding to individual images.

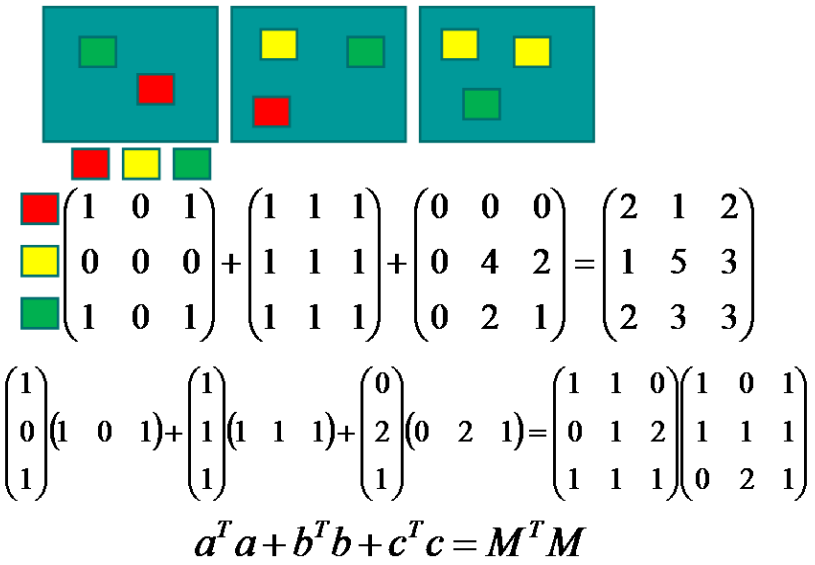


Figure 3.14: Decomposition of standard notion of image level correlation. Rectangles denote visual (visual words) or textual terms

We can identify a few most and least correlated visual words from the matrix of correlations. We can now utilize this information to modify the similarity measure. For this purpose, we are going to use Minkowski's fractional similarity measure (the method may be used with any measure from the Minkowski's family of distances).

**Adaptation of Similarity Measurement**

First, we must identify a certain number of most and least correlated visual terms by looking at the correlation matrix's elements' values above or below the diagonal (symmetric matrix). Let us assume that we have identified the dominant correlation $\{vt_k, vt_l\}$ and the least correlated pair

$\{vt_n, vt_m\}$. Let us now look at the query and extract the frequencies of visual terms corresponding to $vt_k, vt_l, vt_n, vt_m$. We can assume that $f_k \geq f_l$ and $f_n \geq f_m$, where $f$ denotes the frequency of a visual word taken from the query.

The similarity measure can be represented as

$$\begin{aligned} s(Q,I) &= \left( \sum_{i=1}^{N} \sqrt[1/p]{|f_{Q_i} - f_{I_i}|} \right)^p = \\ &= \left( \sqrt[1/p]{|f_{Q_1} - f_{I_1}|} \right)^p + \left( \sqrt[1/p]{|f_{Q_2} - f_{I_2}|} \right)^p \\ &+ \ldots + \left( \sqrt[1/p]{|f_{Q_l} - f_{I_l}|} \right)^p \\ &+ \ldots + \left( \sqrt[1/p]{|f_{Q_n} - f_{I_n}|} \right)^p \\ &+ \ldots + \left( \sqrt[1/p]{|f_{Q_N} - f_{I_N}|} \right)^p \end{aligned} \tag{3.4}$$

We can modify the similarity measure as follows

$$\begin{aligned} s(Q,I) &= \left( \sum_{i=1}^{N} \sqrt{|f_{Q_i} - f_{I_i}|} \right)^2 = \\ &= \left( \sqrt{|f_{Q_1} - f_{I_1}|} \right)^2 + \left( \sqrt{|f_{Q_2} - f_{I_2}|} \right)^2 \\ &+ \ldots + \left( \sqrt{\frac{f_{Q_k}}{f_{Q_l} \cdot \log_b corr\,(vt_k, vt_l)}\, |f_{Q_l} - f_{I_l}|} \right)^2 \\ &+ \ldots + \left( \sqrt{\frac{f_{Q_m}}{f_{Q_n} \cdot \log_b corr\,(vt_n, vt_m)}\, |f_{Q_n} - f_{I_n}|} \right)^2 \\ &+ \ldots + \left( \sqrt{|f_{Q_N} - f_{I_N}|} \right)^2 \end{aligned} \tag{3.5}$$

where $Q$ denotes the query representation, $I$ is an image representation from the data collection, and $corr\,(vt_k, vt_l)$ and $c\,(vt_n, vt_m)$ are the correlation values taken from the correlation matrix.

Thus, we increase the elements corresponding to the visual word in the query with the lower frequency value (dominant correlations), and decrease the elements corresponding to the visual word in the query with the higher frequency value (least correlated pairs). Having performed the similarity measure weighting, we re-rank the top images by calculating the new distance between the query and the images returned in the first round retrieval.

The proposed method is computationally and data storage cheap, utilizes notions of correlation at different contextual levels, and avoids the normalization of histograms.

## 3.4 Chapter Summary

In this chapter, we introduce our novel visual features for large scale generic image retrieval.

First, we develop global visual methods based on our novel edge detector and co-occurrence matrix. The purpose of detecting edges in images is to capture important events and changes in image properties. This is also a method that allows us to reduce the amount of information in images, keeping the most important structural properties only. Here, we introduce our novel edge detector. The aim was to develop a detector that would be efficient, fast and easy to implement. The 8-orientation co-occurrence matrix captures the distribution of edges in images. The statistics extracted from the matrix form visual representations of images. The best performing existing visual features combine global and local methods.

Novel local features (the so-called mid-level features) based on the "Bag of Visual Words" framework are presented next. Local features characterize local properties of images and have some ability to recognize objects. They were inspired by the "Bag of Words" framework from text retrieval and are considered mid-level visual features. Because high-level visual features are currently too computationally expensive (only applied to domain-specific tasks with a limited number of semantic concepts), local features represent up-to-date state-of-the-art in generic CBIR.

The standard Bag of Visual Words features are still, however, computational and data-storage costly as they produce high dimensional vectors. This is one of the reasons why the local features are not in commercial use yet. Thus, visual features for large-scale image retrieval need to address this problem.

Another problem is that local features with standard detector-based sampling are not particularly suited for generic image retrieval. Detector-based sampling concentrates on foreground, which is beneficial for object recognition tasks. However, in generic image retrieval background information is also very important as it often conveys the visual context (may change the meaning of the foreground information). Moreover, the set of sample points identified by a detector can get saturated, which would result in the detector's inability to produce more discriminative sample points. It has been shown, and our experiments confirm this finding, that random sampling works best if the number of sample points is high. This, however, comes at a price of high computational and data storage cost.

We propose easy to implement local features, not sophisticated, with low computational and data storage cost (mostly because our vectors are low dimensional). The utilized hybrid sampling technique can be used in other methods based on the "Bag of Visual Words" to improve the retrieval performance. It captures both foreground and background (visual context) properties of images, which are both important for generic image retrieval.

In this chapter, we also propose an enhancement of the "Bag of Visual Words" model by exploiting the correlations between visual words. The standard "Bag of Visual Words" framework assumes that visual words are independent of each other, while in fact they are often correlated.

Other models that try to utilize the relationships between visual words are often based on the

query modification and are storage and computationally expensive. However, query modification leads to representation renormalization which may hamper the retrieval. Some researchers report performance improvement after normalization, while others report the opposite. This may be due to the different complexity of images. We believe that this may be also domain specific. The $tf \cdot idf$ weighting scheme may work better in case when the precise object matching is important. In our case, the normalization hampers the retrieval performance. However, we can completely avoid representation normalization by adapting the similarity measure based on the most and least correlated visual words. Moreover, we introduce and test different notions of correlation within the proposed framework. The introduced notions of correlation can also be utilized in various approaches which exploit co-occurrence matrices.

Thus, we propose a novel approach to exploit the inter-relationships between the visual words. We introduce and test a few notions of correlation at different contextual levels. First, we generate a matrix of correlations between visual words for each top image returned in the first round retrieval. Second, we aggregate the matrices and identify the dominant and least correlated coefficients. The thus obtained information, along with the visual words' frequencies from the current query, is then utilized to weight the similarity measure. Certain coefficients in the similarity measure corresponding to highly correlated terms are then increased, while the coefficients related to least correlated visual words are deemphasized. The images returned in the first round retrieval are then re-ranked according to the modified similarity measure. Our hypothesis is that not only the most correlated visual words are important to the retrieval but also the least correlated ones.

# Chapter 4

# Combining Visual and Textual Systems

The semantic gap can be also reduced by combining various representations. Visual and textual features, for example, represent correlated and complementary aspects of the same information object, an image. This correlation and complementarity can be exploited to further improve image retrieval.

Existing methods often rank the images by heuristically combining feature spaces of different media types or combining the ranking scores computed independently from different feature spaces.

We propose a principled approach to feature combination inspired by Quantum Theory. Specifically, we propose a tensor product based model aiming to represent text and visual content features of an image as a non-separable composite system. The ranking scores of the images are then computed in the form of quantum-like measurement. In addition, the correlations between features of different media types are incorporated in the framework.

The tensor product requires images to have annotations. Originally, two relatively straightforward statistical approaches for making associations between dimensions of both feature spaces were utilized in the model. Later in this chapter, we alleviate the problem regarding unannotated images by projecting them onto subspaces representing visual context and by incorporating a quantum-like measurement. The proposed principled approach extends the traditional vector space model (VSM) and seamlessly integrates with our tensor-based framework.

The most common way of combining different features is to utilize fusion strategies, i.e. late and early fusion. Each fusion scheme is believed to have its inherent flaws and both are considered to represent different combination techniques. However, experimental results have hinted at the potential interchangeability between specific fusion methods (Lingenfelser, Wagner & André 2011). We have discovered that this duality is related to the interaction between specific similarity measurements and specific early fusion strategies. The last part of this chapter presents mathematical proofs which confirm some of the experimental findings. This interchangeability has many profound implications which will be discussed later in this chapter.

This chapter describes the fundamental concepts of a quantum theory inspired information

retrieval framework by which some of our models were inspired. It presents our hybrid model for visual and textual features combination, projection-based methods for the image-tag associations (a part of the hybrid model framework), and a theoretical investigation of the interactions between similarity measurements and early fusion operators. The aforementioned theoretical investigation leads to the discovery of the interchangeability of specific fusion strategies, which can drastically reduce the computational cost of our tensor-based model.

The experimental results are reported in Chapter 6.4.

## 4.1 Quantum Theory Inspired Image Retrieval Framework

Quantum Mechanics (QM) utilizes mathematical tools that can be very useful in CBIR. These tools can naturally expand the standard Vector Space Model which is usually employed in information retrieval. Moreover, they offer an opportunity to unify different models and formalize CBIR. There are also analogies between QM and IR. For more information regarding the foundations of quantum theory and the quantum theory inspired information retrieval the reader is referred to (Griffiths 2003) and (van Rijsbergen 2004).

#### Fundamental Concepts of Quantum Representation and Measurement

Let us expand the domain of our models from vector to Hilbert spaces. A simple Euclidean vector space with a standard inner product can be considered a Hilbert space.

Traditionally multimedia documents are represented as vectors in Euclidean vector space. For example, a document can be represented by its textual feature, $d^t = (tf_1, tf_2, \cdots, tf_n)^T$, where $tf_n$ is the frequency of term $t_n$ appearing in the document $d$, and $tf_n$ equals to zero when term $t_n$ does not appear in $d$. The visual feature representation for the document can assume the same form, e.g. $d^v = (f_1, f_2, \cdots, f_m)^T$, where $v$ denotes the type of visual feature and $f_i$ refers to the feature value of the $i$th-dimension in the feature space.

In quantum information retrieval, a document can be represented as a superposition state (in each individual feature space). In text feature space, $H_t$: $|d\rangle^t = \sum_i w_{t_i}|t_i\rangle$, where $\sum_i w_{t_i}^2 = 1$. Because the amplitude $w_{t_i}$ for each state $|t_i\rangle$ should be proportional to the probability that the document is about the term $t_i$, we can define it as the normalized term frequency of $t_i$, e.g. $w_{t_i} = tf_i/\sqrt{\sum_j^n tf_j^2}$. Note that the amplitude $w_{t_i}$ can be defined as any other traditional term weighting scheme, e.g. TF-IDF. The only restriction here is to make sure that the sum of $w_{t_i}^2$ should be equal to one. Similarly, a document can also be described as a superposition state in a content feature space $H_v$: $|d\rangle^v = \sum_i w_{f_i}|f_i\rangle$.

The density matrix (probability distribution) of a document is given by

$$\rho_d = |d\rangle\langle d| = \sum_i \alpha_i^2 |t_i\rangle\langle t_i| \tag{4.1}$$

where $|t_i\rangle$

$$\langle t_i|t_j\rangle = \delta_{ij} \tag{4.2}$$

form an orthonormal basis in a Hilbert space, and $\alpha_i$ represents an amplitude corresponding to term $t_i$. If $|t_i\rangle$ are not orthogonal, we can always change the basis. Thus defined density matrix is a diagonal matrix with trace 1, whose entry corresponds to the probability that the document is about the term $t_i$.

If we want to measure probability that a document is about text or a 'visual feature, i.e. the probability that the superposition document collapses to a certain state, we can apply the vector product $P(t_i|d) = |\langle t_i|d\rangle^t|^2 = w_{t_i}^2$, which can be seen as a projection onto the space spanned by $|t_i\rangle$: $P(t_i|d) = \langle t_i|\rho_d|t_i\rangle = w_{t_i}^2$.

An observable is a property of the system state that can be determined by a sequence of physical operations. If we treat a query as an observable, then the density of a query is

$$O = \rho_q = \sum_i q_i^2 |t_i\rangle\langle t_i| \tag{4.3}$$

where $q_i$ represents an amplitude corresponding to terms $t_i$ in the query.

Now, in order to perform a measurement on a document (compute the relevance of a document to the given query), we need to prepare the document density matrix in terms of the eigenstate of the observable:

$$\rho_d' = U\rho_d U' \tag{4.4}$$

$\rho'$ and $\rho$ define the same density matrix if and only if there is a unitary matrix U such that $U'U = I$

$$|t_i'\rangle\sqrt{w_i'} = \sum_j U_{ij}|t_i\rangle\sqrt{w_i} \tag{4.5}$$

According to the quantum measurement :

$$\begin{aligned}\langle O\rangle &= tr(U\rho_d U'UOU') \\ &= tr(U\rho_d U'U\rho_q U')\end{aligned} \tag{4.6}$$

For simplicity, we assume that $|t_i\rangle$ are orthogonal. Then

$$\begin{aligned}\langle O\rangle &= tr(\sum_i \alpha_i^2 |t_i\rangle\langle t_i|O) \\ &= tr(\sum_i \alpha_i^2 |t_i\rangle\langle t_i| \sum_j q_j^2 |t_j\rangle\langle t_j|) \\ &= \sum_i \alpha_i^2 q_i^2\end{aligned} \tag{4.7}$$

When all the images are associated with labels, we can utilize the tensor product of visual and textual features to get a unified system[1].

$$|d\rangle^{tv} = |d\rangle^{t} \otimes |d\rangle^{v} = \sum_{ij} \gamma_{ij} |t_i\rangle \otimes |f_j\rangle \tag{4.8}$$

where $\gamma$ is an amplitude of the composite, non-separable system, $|d\rangle^t$ denotes an image representation in a textual space, and $|d\rangle^v$ denotes an image representation in a visual space.

When considering a superposed multimedia document, the density matrix of its sub-systems (textual and visual, respectively) can be presented as $\rho_{d^t} = \sum_i \alpha_i^2 |t_i\rangle\langle t_i|$, and $\rho_{d^v} = \sum_i \beta_i^2 |f_i\rangle\langle f_i|$. When $|d\rangle^t$ and $|d\rangle^v$ are independent, the density matrix for the composite system is:

$$\rho_{d^{tv}} = \sum_{ij} \alpha_i^2 \beta_j^2 |t_i f_j\rangle \otimes \langle t_i f_j| = \rho_{d^t} \otimes \rho_{d^v} \tag{4.9}$$

In most situations, however, the two systems are not independent, i.e. $\rho_{d^{tv}} = \sum_{ij} \gamma_{ij}^2 |t_i f_j\rangle \otimes \langle t_i f_j|$. We can always separate the correlation term:

$$\rho_{d^{tv}} = \rho_{d^t} \otimes \rho_{d^v} + \rho_{correlation} \tag{4.10}$$

**Correlation Between Feature Dimensions. Mutual Information Matrix**

Now, in order to discover the visual and textual feature's semantic connection, we can utilize the mutual information matrix. The mutual information of two random variables represents a quantity that measures the mutual dependence between two variables. Here, mutual information matrix describes the dependency between the visual and textual features. It gives us the means to associate visual feature dimensions with textual terms.

Each element of the matrix represents the mutual information of feature on dimension $f_i$ and text on $t_i$, which can be defined as follows:

$$vt_{ij} = MI(f_i, t_j) = \log_2 \frac{P(f_i, t_j)}{P(f_i)P(t_j)} \tag{4.11}$$

Here are the definitions of each probability:

- $P(f_i, t_j)$ is the probability that a document contains textual term $t_j$ and elements (visual features) in the bin $i$, $P(f_i, t_j) = \frac{N_{Pixel}(f_i, t_j, c)}{N_{Pixel}(c)}$.
- $P(t_j)$ is the probability that textual term $t_j$ appears in the collection. We use geometric distribution $P(t_j) = 1 - (1-p)^k$ to represent the term distribution, which fits the term

[1]The tensor space opens a door to linking and expanding individual feature spaces as non-separable systems and allowing the correlations existing between them to be naturally incorporated in the unified theoretical framework.

occurrence distribution of our test collection. Here, $p$ is a shape parameter and $p = 0.5$, $k$ is the frequency of term $t_j$ occurring in the collection.

- $P(f_i)$ is the probability that an element (visual features) falls into feature bin $i$, $P(f_i) = \frac{N_{Pixel}(f_i,c)}{N_{Pixel}(c)}$.

Another way to find the correlations between the features' dimensions is presented here. Let $d_i$ denote the i-th image representation. Then the correlation matrix can be computed as

$$\begin{pmatrix} tf_1^1 & tf_1^2 & \dots & tf_1^k \\ tf_2^1 & tf_2^2 & \dots & tf_2^k \\ \vdots & \vdots & \dots & \vdots \\ tf_n^1 & tf_n^2 & \dots & tf_n^k \end{pmatrix} \cdot \begin{pmatrix} f_1^1 & f_2^1 & \dots & f_m^1 \\ f_1^2 & f_2^2 & \dots & f_m^2 \\ \vdots & \vdots & \dots & \vdots \\ f_1^k & f_2^k & \dots & f_m^k \end{pmatrix} = M \tag{4.12}$$

When the correlation matrix is summed up with respect to the word $t$, according to the image's visual feature, an association score between the image and the word can be derived.

$$score(d,t) = \sum_i P(f_i|d) \cdot MI(f_i,t) \tag{4.13}$$

Let us suppose that the document has a feature vector $v = (f_1, f_2, ...f_n)$. Then the expected association score for each word will be $C = v \cdot vt$, which in turn can be used to build density correlation between textual and visual features

$$|d\rangle^t_{expand} = \sum_i c_i|t_i\rangle, \;\; C = v \cdot vt \tag{4.14}$$

$$\rho^d_{correlation} = \sum_{ij} c_i\beta_j|t_i f_j\rangle\langle t_i f_j| \tag{4.15}$$

In practice, we can only choose top $n$ highly scored words to create the correlation density matrix, in order to reduce the computational cost.

**Quantum-like Measurement**

Based on quantum measurement, we score a document according to the observable's expectation on the document. With orthogonal assumption of textual basis $|t_i\rangle$ and visual feature basis $|f_i\rangle$, we have:

$$\begin{aligned} s(d,q) &= tr(\sum_{ij}(t_i^d \cdot f_j^d)^2 |t_i f_j\rangle\langle t_i f_j| \\ &\cdot \quad (t_i^q \cdot f_j^q)^2 |t_i f_j\rangle\langle t_i f_j|) = \\ &= \sum_{ij}(t_i^d \cdot f_j^d)^2 (t_i^q \cdot f_j^q)^2 \end{aligned} \tag{4.16}$$

This shows the same result of transition probability, which is explained as the probability that a system in state $d$ will be found in state $q$ (Aharonov, Albert & Au 1981), and it is computed as $P(q|d) = |\langle q|d\rangle|^2$. When this classical quantum view is applied to a retrieval model, $|\langle q|d\rangle|^2$ can be explained as the probability that a document can be observed containing the information described by the query.

Let us take the superposed document and query as an example:

$$|d\rangle = \sum_{ij} \gamma_{ij}^d |t_i f_j\rangle,\ |q\rangle = \sum_{ij} \gamma_{ij}^q |t_i f_j\rangle \tag{4.17}$$

Then the transition probability between them is:

$$\begin{aligned} s(d,q) &= P(d \to q) \\ &= |\langle d|q\rangle|^2 \\ &= \sum_{i,j}(\gamma_{ij}^d)^2 (\gamma_{ij}^q)^2 \end{aligned} \tag{4.18}$$

In such case, the measurement on the document density matrix is the same as the inner product of two states, which equals to the cosine similarity of two flattened tensors, where the document and query are represented in a tensor form.

Our original method for associating dimensions of the textual feature space with visual features' dimensions, was based on maximum feature likelihood. Within the subset of images containing a textual term, the feature dimension on which most images have the highest feature values was detected by choosing the feature dimension having the highest average feature value. Thereafter, the dimension of a visual feature was associated with a textual term. However, this rather straightforward method can bring the textual noise to the images and was introduced to build and test the unified image retrieval framework. The next subsection describes the projection-based approaches for image-tag associations.

## 4.2 Projection-based Approaches for Image-Tag Associations

Many images do not have textual labels. Therefore, in order to prepare the data for the quantum-like measurement in the tensor space, we need to associate textual terms with images.

The idea behind the projection-based methods is that dimensions of context define subspaces to which vectors of the information objects are projected (Melucci 2005), see also (Di Buccio, Melucci & Song 2011). Thus, we first build a density matrix [2] from the subsets of images containing the textual term $t_i$. This matrix represents a probability distribution and incorporates information about the occurrence of some contextual factors (corresponding to basis vectors). It can be characterized in terms of co-occurrences between visual terms (e.g. visual words). Let $d_i^v$ denote the vector representation of the i-th image. Then the co-occurrence matrix $A$ can be computed as

$$A = \sum_i |d_i^v\rangle\langle d_i^v| \tag{4.19}$$

Here, we assume that the correlations at the image-level may be stronger than the correlations based on the proximity between visual terms (instances of visual words are considered correlated if they appear together within a certain neighbourhood). An image may contain correlated terms (pixels, visual words) not because of their proximity, but because they refer to the same topic (image represents the context). This assumption was inspired by (Biancalana et al. 2009), where the page-based (text) correlations performed best.

#### Measurement Based on Orthogonal Projection

The symmetric correlation matrix $A$ can then be decomposed to estimate the basis, which would represent the "relevance" context:

$$A = U \cdot F \cdot U^T = \sum_i f_i |u_i\rangle\langle u_i| \tag{4.20}$$

where $U$ is a unitary, orthogonal matrix, $f_i$ is an element of $F$ and $u_i$ are eigenvectors of $A$. Vectors $u_i$ form an orthogonal basis of the subspace (as projector) representing the influence of each contextual factor. The projector onto this subspace (denoted as $B$) is equal to $P_B = \sum_i |u_i\rangle\langle u_i|$. $P_B$ can be considered as the semantic subspace characterizing the term $t_i$. Now, each unannotated image $d_i$ can be projected onto this subspace, and the probability of relevance context of $d_i$ may be calculated as

$$Pr\left[L(B)|L(d_i)\right] = \langle d_i|P_B|d_i\rangle \tag{4.21}$$

where $L(d_i)$ denotes a subspace generated by $d_i$. Thus, the images are annotated with respect to the probability that they were generated by a context represented by $P_B$. An unannotated image

[2] The co-occurrence matrix is Hermitian and can be constructed in such a way that the trace would be unitary. Therefore the density and co-occurrence matrix will be used interchangeably in this thesis.

can then be associated with a textual term corresponding to the semantic subspace with the highest probability of projection.

**Alternative Measurement**

Here, we introduce a variation of the projection method based on the quantum measurement. The proposed approach performs quantum like measurement on the density matrix $A$ representing the probability distribution obtained from the subset of training images (containing given tags), and the density matrix $D$ of an unannotated image $d_i$. Therefore

$$P_i = tr(D_i \cdot A) \tag{4.22}$$

where $D_i = |d_i\rangle\langle d_i|$.

## 4.3 On the Interaction Between Certain Early Fusion Schemes and Similarity Measurements

Fusion strategies are widely utilized in Content-based Image Retrieval to combine different systems. The most common fusion schemes are early (fusion of representations) and late (fusion of relevance scores) fusion methods. Based on the experimental results, researchers have hinted at the potential interchangeability of specific fusion schemes (e.g. (Lingenfelser et al. 2011)). We theoretically prove that this interchangeability is related to the interaction between early fusion operators and similarity measurements.

Here, we are going to investigate some interesting interactions between common similarity measurements and common operators related to early fusion strategy (i.e. concatenation of vector representations). We are going to show that these interactions between certain similarity measurements and early fusion strategies result in combinations of representations at the decision level (late fusion strategy). In other words, we theoretically prove that certain combinations of early fusion strategies and certain similarity measurements are equivalent to particular combinations of measurements (i.e. relevance scores) computed on individual feature spaces.

Our findings are important from both theoretical and practical perspectives. First, we should be careful when comparing early and late fusion strategies as they may represent equivalent approaches. Second, specific combinations of individual relevance scores can have a sound theoretical interpretation which would be easier to analyze as an early fusion. Finally, knowing how to represent early fusion as a late one can help us avoid the curse of dimensionality.

### 4.3.1 Inner Product

We can start by making a few simple observations. Let us employ a standard inner product as the similarity measurement. It is easy to check that

$$\left\langle d_1^v \oplus d_1^t | d_2^v \oplus d_2^t \right\rangle = \langle d_1^v | d_2^v \rangle + \left\langle d_1^t | d_2^t \right\rangle \tag{4.23}$$

where $\langle \cdot | \cdot \rangle$ denotes an inner product, $\oplus$ is the direct product (concatenation of vectors) and $d_i^v, d_i^t$ are the visual and textual image representations of the $i$th image, for example.

From the above equation we can see, that concatenation of vectors is equivalent to addition of measurements (scores) performed on individual feature spaces.

To clarify, concatenation ($\oplus$) of two $n$ and $m$ dimensional vectors produces a new $n + m$ dimensional vector, for example

$$(a, b) \oplus (c, d, e) = (a, b, c, d, e) \tag{4.24}$$

The tensor product ($\otimes$) of two $n$ and $m$ dimensional vectors generates an $n \cdot m$ dimensional vector or an $n$ by $m$ dimensional matrix. For example

$$(a, b) \otimes (c, d, e) = (ac, ad, ae, bc, bd, be) \tag{4.25}$$

or

$$(a, b) \otimes (c, d, e) = \begin{pmatrix} ac & ad & ae \\ bc & bd & be \end{pmatrix} \tag{4.26}$$

Sometimes the addition of vectors is utilized as an aggregation function. Then

$$\begin{aligned} \left\langle d_1^v + d_1^t | d_2^v + d_2^t \right\rangle = \langle d_1^v | d_2^v \rangle + \left\langle d_1^v | d_2^t \right\rangle + \left\langle d_1^t | d_2^v \right\rangle + \left\langle d_1^t | d_2^t \right\rangle = \\ \left\langle d_1^v \oplus d_1^t | d_2^v \oplus d_2^t \right\rangle + \left\langle d_1^v \oplus d_1^t | d_2^t \oplus d_2^v \right\rangle \end{aligned} \tag{4.27}$$

This would probably be the least desirable type of combination as we would be comparing vectors from very different feature spaces ($\left\langle d_1^v | d_2^t \right\rangle$, $\left\langle d_1^t | d_2^v \right\rangle$).

It has been shown that the tensor product can be useful when combining the representations as it takes into account all of the combinations of vectors' dimensions and gives good discrimination in terms of similarity measurements (Li & Cunningham 2008). Assuming that the systems were prepared independently, we have

$$\left\langle d_1^v \otimes d_1^t | d_2^v \otimes d_2^t \right\rangle = \langle d_1^v | d_2^v \rangle \cdot \left\langle d_1^t | d_2^t \right\rangle \tag{4.28}$$

where $\otimes$ denotes the tensor operator.

From the above equation it turns out that the inner product of the tensor products is a product

of the measurements (scores) performed on individual feature spaces. One of the implications of this observation is that there is no need for performing the tensor operation. This also applies to the case when the Euclidean metric is used

$$\begin{aligned} s\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \\ \sqrt{\langle d_1^v \otimes d_1^t - d_2^v \otimes d_2^t | d_1^v \otimes d_1^t - d_2^v \otimes d_2^t \rangle} = \\ \sqrt{\langle d_1^v \otimes d_1^t | d_1^v \otimes d_1^t \rangle - 2\langle d_1^v \otimes d_1^t | d_2^v \otimes d_2^t \rangle + \langle d_2^v \otimes d_2^t | d_2^v \otimes d_2^t \rangle} = \\ \sqrt{\langle d_1^v | d_1^v \rangle \langle d_1^t | d_1^t \rangle - 2\langle d_1^v | d_2^v \rangle \langle d_1^t | d_2^t \rangle + \langle d_2^v | d_2^v \rangle \langle d_2^t | d_2^v \rangle} = \\ \sqrt{\|d_1^v\|^2 \cdot \|d_1^t\|^2 + \|d_2^v\|^2 \cdot \|d_2^t\|^2 - 2\langle d_1^v | d_2^v \rangle \langle d_1^t | d_2^t \rangle} \end{aligned} \tag{4.29}$$

### 4.3.2 Cosine Similarity

One of the best performing similarity measures in CBIR is the cosine similarity ($s_c$)

$$s_c\left(d_1, d_2\right) = \frac{\langle d_1 | d_2 \rangle}{\|d_1\| \cdot \|d_2\|} \tag{4.30}$$

If the features are normalized (for the clarity of the formulas - not a necessary assumption) then

$$\begin{aligned} \left\|d_1^v \otimes d_1^t\right\| = \sqrt{\langle d_1^v \otimes d_1^t | d_1^v \otimes d_1^t \rangle} = \\ \sqrt{\langle d_1^v | d_1^v \rangle \cdot \langle d_1^t | d_1^t \rangle} = \|d_1^v\| \cdot \left\|d_1^t\right\| = 1 = \left\|d_2^v \otimes d_2^t\right\| \end{aligned} \tag{4.31}$$

and

$$\begin{aligned} \left\|d_1^v \oplus d_1^t\right\| = \sqrt{\langle d_1^v \oplus d_1^t | d_1^v \oplus d_1^t \rangle} = \sqrt{\langle d_1^v | d_1^v \rangle + \langle d_1^t | d_1^t \rangle} = \\ \sqrt{\|d_1^v\|^2 + \|d_1^t\|^2} = \sqrt{2} = \left\|d_2^v \oplus d_2^t\right\| \end{aligned} \tag{4.32}$$

Therefore, we get

$$s_c\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = s_c\left(d_1^v, d_2^v\right) \cdot s_c\left(d_1^t, d_2^t\right) \tag{4.33}$$

$$s_c\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \frac{1}{2}\left(s_c\left(d_1^v, d_2^v\right) + s_c\left(d_1^t, d_2^t\right)\right) \tag{4.34}$$

Here, the square root of the similarity between the tensored representations is the geometric mean of the scores computed independently and the similarity between the concatenated representations is the arithmetic mean of individual scores.

Let us assume, that a model incorporates cosine similarity as a measurement used in combining the sub-systems (i.e. visual features or visual and textual features). Then, the concatenation or tensor operation produces the same effect as incorporation of the CombSUM or CombPROD late fusion methods, respectively.

### 4.3.3 Cosine Similarity, Weighted Combinations

If we utilize weighted combinations (with $r_1$, $r_2$ denoting the weights, the importance of visual and textual representations, for example), then we get[3]

$$\begin{aligned} s_c\left(r_1 d_1^v \oplus r_2 d_1^t, r_1 d_2^v \oplus r_2 d_2^t\right) = \\ \frac{1}{r_1^2 + r_2^2}\left(r_1^2 s_c\left(d_1^v, d_2^v\right) + r_2^2 s_c\left(d_1^t, d_2^t\right)\right) \end{aligned} \tag{4.35}$$

*Proof.* Because

$$\begin{aligned} \left\|\left(r_1 d^v\right) \oplus \left(r_2 d^t\right)\right\| = \\ \sqrt{\left\langle \left(r_1 d^v\right) \oplus \left(r_2 d^t\right) \mid \left(r_1 d^v\right) \oplus \left(r_2 d^t\right) \right\rangle} = \\ \sqrt{\left\langle r_1 d^v | r_1 d^v \right\rangle + \left\langle r_2 d^t | r_2 d^t \right\rangle} = \sqrt{r_1^2 \left\langle d^v | d^v \right\rangle + r_2^2 \left\langle d^t | d^t \right\rangle} = \\ \sqrt{r_1^2 \left\| d^v \right\|^2 + r_2^2 \left\| d^t \right\|^2} = \sqrt{r_1^2 + r_2^2} \end{aligned} \tag{4.36}$$

we get

$$\begin{aligned} s_c\left(r_1 d_1^v \oplus r_2 d_1^t, r_1 d_2^v \oplus r_2 d_2^t\right) = \\ \frac{\left\langle r_1 d_1^v \oplus r_2 d_1^t | r_1 d_2^v \oplus r_2 d_2^t \right\rangle}{r_1^2 + r_2^2} = \frac{\left\langle r_1 d_1^v | r_1 d_2^v \right\rangle + \left\langle r_2 d_1^t | r_2 d_2^t \right\rangle}{r_1^2 + r_2^2} = \\ \frac{1}{r_1^2 + r_2^2}\left(r_1^2 \frac{\left\langle d_1^v | d_2^v \right\rangle}{\left\| d_1^v \right\| \left\| d_2^v \right\|} + r_2^2 \frac{\left\langle d_1^t | d_2^t \right\rangle}{\left\| d_1^t \right\| \left\| d_2^t \right\|}\right) = \\ \frac{1}{r_1^2 + r_2^2}\left(r_1^2 s_c\left(d_1^v, d_2^v\right) + r_2^2 s_c\left(d_1^t, d_2^t\right)\right) \end{aligned} \tag{4.37}$$

□

### 4.3.4 Euclidean Metric

We can also find the relationships for the Euclidean metric

$$s_e\left(d_1, d_2\right) = \sqrt{\left\langle d_1 - d_2 | d_1 - d_2 \right\rangle}. \tag{4.38}$$

Thus

$$s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right)} \tag{4.39}$$

[3] Similar observations can be made for other similarity measurements. Here, we only present the weighted combinations for the cosine similarity.

and

$$s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right) - \frac{1}{2} s_e^2\left(d_1^v, d_2^v\right) s_e^2\left(d_1^t, d_2^t\right)} \tag{4.40}$$

*Proof.* (1) From the fact, that

$$s_e\left(d_1, d_2\right) = \sqrt{\left\|d_1\right\|^2 + \left\|d_2\right\|^2 - 2\left\langle d_1 | d_2 \right\rangle} \tag{4.41}$$

and

$$\left\|d_1 \oplus d_2\right\| = \sqrt{2} \tag{4.42}$$

we can show, that

$$\begin{aligned} s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\ \sqrt{\left\|d_1^v \oplus d_1^t\right\|^2 + \left\|d_2^v \oplus d_2^t\right\|^2 - 2\left\langle d_1^v \oplus d_1^t | d_2^v \oplus d_2^t \right\rangle} = \\ \sqrt{4 - 2\left(\left\langle d_1^v | d_2^v \right\rangle + \left\langle d_1^t | d_2^t \right\rangle\right)} = \\ \sqrt{2 - 2\left\langle d_1^v | d_2^v \right\rangle + 2 - 2\left\langle d_1^t | d_1^t \right\rangle} = \\ \sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right)} \end{aligned} \tag{4.43}$$

□

(2) Notice, that

$$\begin{aligned} s_e^2\left(d_1^v, d_2^v\right) \cdot s_e^2\left(d_1^t, d_2^t\right) = \\ \left(2 - 2\left\langle d_1^v | d_2^v \right\rangle\right) \cdot \left(2 - 2\left\langle d_1^t | d_2^t \right\rangle\right) = \\ 2\left(2 - 2\left\langle d_1^v | d_2^v \right\rangle\right) + \\ 2\left(2 - 2\left\langle d_1^t | d_2^t \right\rangle\right) - 2\left(2 - 2\left\langle d_1^v | d_2^v \right\rangle \left\langle d_1^t | d_2^t \right\rangle\right) = \\ 2s_e^2\left(d_1^v, d_2^v\right) + 2s_e^2\left(d_1^t, d_2^t\right) - 2s_e^2\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) \end{aligned} \tag{4.44}$$

□

### 4.3.5 Bhattacharya Similarity

Similarly, for the Bhattacharya similarity

$$s_b\left(d_1, d_2\right) = -ln\left(\sum_i \sqrt{\left(d_1\right)_i \cdot \left(d_2\right)_i}\right) \tag{4.45}$$

we get

$$s_b\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = s_b\left(d_1^v, d_2^v\right) + s_b\left(d_1^t, d_2^t\right) \tag{4.46}$$

and

$$s_b\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = -ln\left(e^{-s_b\left(d_1^v, d_2^v\right)} + e^{-s_b\left(d_1^t, d_2^t\right)}\right) \tag{4.47}$$

*Proof.* Let us denote

$$\sqrt{d} = \left(\sqrt{d_1}, \sqrt{d_2}, \ldots, \sqrt{d_n}\right) \tag{4.48}$$

Then

$$\begin{aligned}
s_b\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \\
-ln\left(\sum_k \sqrt{\left(d_1^v \otimes d_1^t\right)_k \cdot \left(d_2^v \otimes d_2^t\right)_k}\right) = \\
-ln\left(\left\langle \sqrt{d_1^v \otimes d_1^t} | \sqrt{d_2^v \otimes d_2^t} \right\rangle\right) = \\
-ln\left(\left\langle \sqrt{d_1^v} \otimes \sqrt{d_1^t} | \sqrt{d_2^v} \otimes \sqrt{d_2^t} \right\rangle\right) = \\
-ln\left(\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle \cdot \left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle\right) = \\
-\left(ln\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle + ln\left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle\right) = \\
-\left(ln \sum_i \sqrt{\left(d_1^v\right)_i \cdot \left(d_2^v\right)_i} + ln \sum_j \sqrt{\left(d_1^t\right)_j \cdot \left(d_2^t\right)_j}\right) = \\
s_b\left(d_1^v, d_2^v\right) + s_b\left(d_1^t, d_2^t\right)
\end{aligned} \tag{4.49}$$

□

For the concatenation, we have

$$\begin{aligned} s_b\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\ -ln\left(\sum_k \sqrt{\left(d_1^v \oplus d_1^t\right)_k \cdot \left(d_2^v \oplus d_2^t\right)_k}\right) = \\ -ln\left(\left\langle \sqrt{d_1^v \oplus d_1^t} | \sqrt{d_2^v \oplus d_2^t} \right\rangle\right) = \\ -ln\left(\left\langle \sqrt{d_1^v} \oplus \sqrt{d_1^t} | \sqrt{d_2^v} \oplus \sqrt{d_2^t} \right\rangle\right) = \\ -ln\left(\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle + \left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle\right) = \\ -ln\left(e^{ln\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle} + e^{ln\left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle}\right) = \\ -ln\left(e^{-s_b\left(d_1^v, d_2^v\right)} + e^{-s_b\left(d_1^t, d_2^t\right)}\right) \end{aligned} \tag{4.50}$$

□

### 4.3.6 Combination of Euclidean Metric and Cosine Similarity

Sometimes it might be beneficial to utilize different similarity measures for different feature spaces (Chen et al. 2001) (i.e. Euclidean metric for visual features and cosine similarity for textual space). Interestingly, we can fuse the scores in such a way, that their combination would correspond to (for example) measuring the Euclidean distance between the concatenated or tensored representations

$$\begin{aligned} s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\ \sqrt{s_e^2\left(d_1^v, d_2^v\right) - 2s_c\left(d_1^t, d_2^t\right) + 2} \end{aligned} \tag{4.51}$$

$$\begin{aligned} s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \\ \sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2s_c\left(d_1^t, d_2^t\right) + 2} \end{aligned} \tag{4.52}$$

*Proof.* Stems from the fact that

$$\begin{aligned} s_e^2\left(d_1^t, d_2^t\right) = \\ 2 - 2\left\langle d_1^t | d_2^t \right\rangle = 2 - 2\frac{\left\langle d_1^t | d_2^t \right\rangle}{\|d_1^t\| \, \|d_2^t\|} = 2 - 2s_c\left(d_1^t, d_2^t\right) \end{aligned} \tag{4.53}$$

and (4.39),(4.40). □

### 4.3.7 Minkowski Family of Distances

The Minkowski family of distances includes the widely utilized Manhattan and Euclidean metrics. Manhattan distance, for example, was utilized in (Liu 2010) to query the CBIR system by multiple

visual examples. In the aforementioned paper, individual scores corresponding to visual examples were aggregated. It is interesting to know, that if another researcher concatenated the representations corresponding to visual examples and utilized Manhattan metric, then the influence of these fusion methods on the retrieval performance would be exactly the same.

The Minkowski family of distances is represented by the formula

$$s_p\left(d_1, d_2\right) = \left(\sum_{i=1}^{n} \left|d_1^i - d_2^i\right|^p\right)^{\frac{1}{p}} \tag{4.54}$$

where $p \in \mathbb{N}$.

For the fractional values of $p \in (0, 1)$, the formula is not a metric in the mathematical sense. However, it has been shown (Liu et al. 2008), that the similarity measure with fractional values of $p$ works well in CBIR.

We are going to show, that

$$s_p\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \left(s_p^p\left(d_1^v, d_2^v\right) + s_p^p\left(d_1^t, d_2^t\right)\right)^{\frac{1}{p}} \tag{4.55}$$

*Proof.*

$$\begin{aligned} s_p\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\ \left\|d_1^v \oplus d_1^t - d_2^v \oplus d_2^t\right\|_p = \\ \left\|\left(d_1^v - d_2^v\right) \oplus \left(d_1^t - d_2^t\right)\right\|_p = \\ \left(\left\|d_1^v - d_2^v\right\|_p^p + \left\|d_1^t - d_2^t\right\|_p^p\right)^{\frac{1}{p}} = \\ \left(s_p^p\left(d_1^v, d_2^v\right) + s_p^p\left(d_1^t, d_2^t\right)\right)^{\frac{1}{p}} \end{aligned} \tag{4.56}$$

where

$$\left\|d\right\|_p = \left(d_1^p + d_2^p + \ldots d_n^p\right)^{\frac{1}{p}} \tag{4.57}$$

Here, the representations do not have to be normalized. □

Hence, in these cases the early and late fusion approaches are interchangeable. The fusion of representations is then, in fact, the fusion of similarities computed independently on visual and textual feature spaces. This is, in our opinion, an interesting finding.

### 4.3.8 Example

Let us give a quick, trivial example. We can assume, that we have two images. Each image has both visual and textual representations. Because we want to compute the similarity between the images (i.e. Minkowski $p = \frac{1}{4}$), we need to fuse the visual and textual features. We can utilize

one of the equivalent fusion strategies (late and early fusion). Thus, for concatenation operation

$$d_1^v = (1, 3, 4)) \tag{4.58}$$

$$d_2^v = (0, 3, 5)) \tag{4.59}$$

$$d_1^t = (12, 1, 4, 2)) \tag{4.60}$$

$$d_2^t = (11, 0, 3, 1)) \tag{4.61}$$

Let us calculate

$$s_{p=\frac{1}{4}}\left(d_1^v, d_2^v\right) = \left(|1-0|^{\frac{1}{4}} + |3-3|^{\frac{1}{4}} + |4-5|^{\frac{1}{4}}\right)^4 = 2^4 = 16 \tag{4.62}$$

$$s_{p=\frac{1}{4}}\left(d_1^t, d_2^t\right) = \left(|12-11|^{\frac{1}{4}} + |1-0|^{\frac{1}{4}} + |4-3|^{\frac{1}{4}} + |2-1|^{\frac{1}{4}}\right)^4 = 4^4 = 256 \tag{4.63}$$

Therefore, the right-hand side of the equation becomes

$$R = \left(s^{\frac{1}{4}}_{p=\frac{1}{4}}\left(d_1^v, d_2^v\right) + s^{\frac{1}{4}}_{p=\frac{1}{4}}\left(d_1^t, d_2^t\right)\right)^4 = \left(16^{\frac{1}{4}} + 256^{\frac{1}{4}}\right)^4 = (2+4)^4 = 1296 \tag{4.64}$$

For the left-hand side, we have

$$L = s_{p=\frac{1}{4}}\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = s_{p=\frac{1}{4}}\left((1, 3, 4, 12, 1, 4, 2), (0, 3, 5, 11, 0, 3, 1)\right) = (1+0+1+1+1+1+1)^4 = 1296 \tag{4.65}$$

Thus, $R = L$.

What this mean is that both fusion strategies will lead to the same results and are thus interchangeable.

## 4.4 Chapter Summary

In this chapter, we propose a novel tensor-based hybrid model for the combination of visual and textual feature spaces. As the model was inspired by mathematical tools used in quantum theory, we first introduced fundamental concepts related to quantum theory inspired framework for information retrieval.

The aforementioned tools can naturally expand the standard Vector Space Model which is usually employed in information retrieval. Moreover, they offer an opportunity to unify different

models and formalize CBIR. There are also many analogies between QM and IR.

In existing hybrid models, the ranking of documents is often computed by heuristically combining the feature spaces of different media types or combining the ranking scores computed independently from different feature spaces. All these combination methods treat the textual and visual features individually, and combine them in a rather heuristic manner. Therefore, this makes it difficult to capture the relationship between the features. Indeed, as both the textual and visual representations describe the same image, there are inherent correlations between them which should be incorporated into the retrieval process as a whole in a more principled way.

Thus, we propose a quantum theory inspired multimedia retrieval framework based on the tensor product of feature spaces, where similarity measurement between a query and a document corresponds to the quantum measurement. At the same time, the correlations between dimensions across different feature spaces can also be naturally incorporated in the framework. The tensor based model provides a formal and flexible way to expand the feature spaces, and seamlessly integrate different features, potentially enabling multi-modal and cross media search in a principled and unified framework.

Our tensor-based model requires images to have annotations. Existing auto-annotation methods based on segmentation models or grouping of visual words are computationally expensive and not scalable to large data collections. Methods based on the clustering of visual and textual features, on the other hand, neglect the contextual information.

We introduce two different approaches for making image-tag associations. The first proposed approach projects the unannotated images onto the subspaces generated by subsets of training images (containing given textual terms). We calculate the probability of an image being generated by the contextual factors related to the same topic. In this way, we should be able to capture the visual contextual properties of images, taking advantage of this extended vector space model framework. The other method performs quantum like measurement on the density matrix of an unannotated image, with respect to the density matrix representing the probability distribution obtained from the subset of training images. These approaches can be seamlessly integrated into our unified framework for image retrieval.

The last part of this chapter is devoted to the theoretical investigation of the interchangeability of the fusion strategies.

Fusion approaches are often utilized in existing hybrid models. Recently, researchers hinted at the potential interchangeability of concrete fusion strategies. We theoretically prove that fusion schemes are, in concrete cases, interchangeable. This fact has many profound consequences. First, concrete early and late fusion strategies utilized in existing hybrid models may represent equivalent approaches. Second, concrete combinations of individual relevance scores can have a sound theoretical interpretation which would be easier to analyze as an early fusion. Finally, knowing how to represent early fusion as a late one can help us avoid the curse of dimensionality.

For future work we are going to search for other combinations of various operators and similarity measures that could interact in such a way as to represent late fusion. Moreover, the obser-

vations on the duality of fusion strategies may have additional interesting applications, namely the clustering of tensored or concatenated vectors.

# Chapter 5

# Combining Visual and Textual Systems in the Context of Relevance Feedback

The previous chapters presented two approaches to semantic gap reduction. The first was the development of novel visual features, their adaptation to large-scale generic image retrieval, and their enhancement based on the correlation between histograms dimensions. The second approach showed how to exploit the novel and enhanced representations in combination with other visual and textual features to further improve the retrieval effectiveness.

In this chapter, we discuss the third approach to semantic gap reduction, namely the combination of different features (i.e. textual and visual) in the context of relevance feedback[1]. Relevance feedback can narrow down the search and put the query in the right context. In the hybrid relevance feedback model, apart from the visual and textual query representations, we would have additional information in the form of sets of feedback images. The sets of feedback images would themselves consist of subsets (in Hilbert space - subspaces) of visual and textual feedback representations. Moreover, there are some inherent inter and intra feature relationships between visual and textual feature spaces.

Up to now, there is not much research on hybrid relevance feedback models. Existing models do not exploit the aforementioned relationships between the feature spaces.

In this chapter, we present a model for the combination of visual and textual sub-systems within the user feedback context. The model was inspired by the measurement utilized in quantum mechanics (QM) and the tensor product of co-occurrence (density) matrices, which represents a density matrix of the composite system in QM. It provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context. The proposed approach takes into account both intra (via co-occurrence matrices) and inter (via tensor operator) feature relationships between features' dimensions. It is also computationally cheap and scalable to large data collections.

[1]The context in this thesis, in reference to hybrid models, denotes images obtained from the relevance feedback.

Thus, the following chapter presents a novel hybrid relevance feedback model and its variations based on the orthogonal projection, image re-ranking, and discrete and continuous degrees of relevance. We also introduce a generalization of the hybrid relevance feedback model to multiple features, and dynamic weighting of a query and its context in our hybrid relevance feedback model. We also show how the query modification techniques can be represented as a combination of relevance scores.

The experimental results are reported in Chapter 6.6.

## 5.1 Hybrid Relevance Feedback Model

Modern retrieval systems allow the users to interact with the system in order to narrow down the search. This interaction takes the form of implicit or explicit feedback. The representations of the images in the feedback set are often aggregated or concatenated (or co-occurrence matrices may be aggregated to represent i.e. probability distribution matrix). The information extracted from the feedback set is utilized to expand the query or re-rank the top images returned in the first round of the retrieval.

Here, we are going to introduce our model for visual and textual systems' combination within the context of a user feedback. The model is defined on a Hilbert space (a real space[2] with a standard inner product) which can be thought of as a natural extension of the standard vector space model, with its useful notions of subspaces and projections. It was inspired by the mathematical tools utilized in Quantum Mechanics (QM) and is based on the expectation value, predicted mean value of the measurement (e.g. similarity or relevance measurement)

$$\langle A \rangle = tr\left(\rho A\right) \tag{5.1}$$

where $tr$ denotes the trace operator, $\rho$ represents a density matrix of the system (e.g. the context of a search query) and $A$ is an observable (e.g. an image). A good source of information on the analogies between Quantum Mechanics and Information Retrieval (IR) is (van Rijsbergen 2004). This book has recently inspired some interesting research in IR, which provide various useful formalization and mathematical tools.

The aforementioned type of measurement (similarity measurement in our case) is very useful as it allows us to directly utilize and combine term (visual or textual) co-occurrence matrices. The notion of co-occurrence, on the other hand, can be exploited in order to model the subspaces of feedback images (visual and textual).

Thus, an observable $A$ can be also represented as a density matrix (corresponding to the query or an image in the collection). Moreover, the co-occurrence matrices can be treated as density matrices (probability distribution) because they are Hermitian and positive-definite. Therefore, the density matrix of the system ($\rho$) utilized in the hybrid CBIR relevance feedback model can be

[2]Hilbert space is usually defined as a complex space with an inner product.

operationalized as the image-level co-occurrence matrices derived from the feedback images that capture the context of the query.

The trace operator acting on density matrices can be reformulated as (real matrices)

$$tr\left(\rho A\right) = \left\langle A \middle| \rho^T \right\rangle \tag{5.2}$$

where $T$ denotes the matrix transposition operation, and $\langle \cdot | \cdot \rangle$ represents an inner product. Moreover, for symmetric matrices $\left\langle A \middle| \rho^T \right\rangle = \langle A|\rho\rangle = \langle \rho|A\rangle$.

In the hybrid CBIR relevance feedback model, the tensor operator $\otimes$ is utilized to combine the density matrices corresponding to visual and textual feature spaces. In quantum mechanics, the tensor product of density matrices of different systems represents a density matrix of the combined system (Jacobs 2011). The tensor product of two matrices is just another matrix (with all the possible combinations of elements)

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes \begin{pmatrix} e & f & g \\ h & i & j \\ k & l & m \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} e & \begin{pmatrix} a & b \\ c & d \end{pmatrix} f & \begin{pmatrix} a & b \\ c & d \end{pmatrix} g \\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} h & \begin{pmatrix} a & b \\ c & d \end{pmatrix} i & \begin{pmatrix} a & b \\ c & d \end{pmatrix} j \\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} k & \begin{pmatrix} a & b \\ c & d \end{pmatrix} l & \begin{pmatrix} a & b \\ c & d \end{pmatrix} m \end{pmatrix} = \begin{pmatrix} ae & be & af & bf & ag & bg \\ ce & de & cf & df & cg & dg \\ ah & bh & ai & bi & aj & bj \\ ch & dh & ci & di & cj & dj \\ ak & bk & al & bl & am & bm \\ ck & dk & cl & dl & cm & dm \end{pmatrix} \tag{5.3}$$

The tensor product captures all the combinations of elements of the multiplied matrices.

The measurement (here, similarity measurement in the context of user feedback) is represented by

$$tr\left(\left(M_1 \otimes M_2\right) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right) \tag{5.4}$$

where $M_1$, $M_2$ represent density matrices (co-occurrence matrices) of the query and images in the feedback set corresponding to visual and textual spaces respectively, and $a$ and $b$ denote vectors representing visual and textual information for an image from the data collection. This measurement would be performed on all the images in the collection, thus re-scoring the dataset based on

the user feedback.

Thus, the intra correlations are captured by correlation matrices corresponding to individual feature spaces, and the composite system and inter correlations are modelled as the tensor product $M_1 \otimes M_2$.

Taking into account the properties of an inner product and tensor operator, we get

$$\begin{aligned} tr\left((M_1 \otimes M_2) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right) = \\ tr\left(\left(M_1 \cdot \left(a^T \cdot a\right)\right) \otimes \left(M_2 \cdot \left(b^T \cdot b\right)\right)\right) = \\ tr\left(M_1 \cdot \left(a^T \cdot a\right)\right) \cdot tr\left(M_2 \cdot \left(b^T \cdot b\right)\right) = \\ \left\langle M_1 \middle| a^T \cdot a \right\rangle \cdot \left\langle M_2 \middle| b^T \cdot b \right\rangle \end{aligned} \tag{5.5}$$

Analogously, we can easily extrapolate to the higher number of features (i.e textual and multiple visual features)

$$\begin{aligned} tr\left((\otimes_n M_n) \cdot \left(\otimes_n \left(a_n^T \cdot a_n\right)\right)\right) = \\ \prod_n \left\langle M_n \middle| a_n^T \cdot a_n \right\rangle \end{aligned} \tag{5.6}$$

Let $q_v$, $q_t$ denote the visual and textual representations of the query; $c^i$, $d^i$ denote visual and textual representations of the images in the feedback set; $D_q^v$, $D_f^v$ denote the density (co-occurrence) matrices of a visual query and its visual context (feedback images); $D_q^t$, $D_f^t$ denote the density matrices of a textual query and its textual context; $r_1$, $1 - r_1$ ($r_2$, $1 - r_2$) denote the weighting factors (constant, importance of a query and feedback density matrices, respectively); and n denote the number of images in the feedback set. Then, $M_1$ and $M_2$ can be defined as weighted combinations of co-occurrence matrices (a subspace generated by the query vector and vectors from the feedback set)

$$\begin{aligned} M_1 = r_1 \cdot D_q^v + \frac{1 - r_1}{n} \cdot D_f^v = \\ r_1 \cdot q_v^T \cdot q_v + \sum_i \left(\frac{1 - r_1}{n} \cdot \left(c^i\right)^T \cdot c^i\right) \end{aligned} \tag{5.7}$$

and

$$\begin{aligned} M_2 = r_2 \cdot D_q^t + \frac{1 - r_2}{n} \cdot D_f^t = \\ r_2 \cdot q_t^T \cdot q_t + \sum_i \left(\frac{1 - r_2}{n} \cdot \left(d^i\right)^T \cdot d^i\right) \end{aligned} \tag{5.8}$$

In our model, the utilized co-occurrence matrices represent document and image-level correlations (context is represented by the whole document). The matrices can be generated by multiplying the term-document matrix $M$ by its transpose (rows of the matrix represent the documents

$d_1, \dots d_m$), that is $D = M^T \cdot M$. This is equivalent to $D = \sum_{i=1}^{n} d_i^T \cdot d_i$.

Thus, the model can be simplified further

$$\begin{aligned}
\left\langle M_1 \otimes M_2 \middle| \left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right) \right\rangle = \\
\left\langle M_1 \middle| a^T \cdot a \right\rangle \cdot \left\langle M_2 \middle| b^T \cdot b \right\rangle = \\
\left\langle r_1 \cdot q_v^T \cdot q_v + \sum_i \left( \frac{1 - r_1}{n} \cdot \left(c^i\right)^T \cdot c^i \right) \middle| a^T \cdot a \right\rangle \cdot \\
\left\langle r_2 \cdot q_t^T \cdot q_t + \sum_i \left( \frac{1 - r_2}{n} \cdot \left(d^i\right)^T \cdot d^i \right) \middle| b^T \cdot b \right\rangle = \\
\left( \left\langle r_1 \cdot q_v^T \cdot q_v \middle| a^T \cdot a \right\rangle + \sum_i \frac{1 - r_1}{n} \left\langle \left(c^i\right)^T \cdot c^i \middle| a^T \cdot a \right\rangle \right) \cdot \\
\left( \left\langle r_2 \cdot q_t^T \cdot q_t \middle| b^T \cdot b \right\rangle + \sum_i \frac{1 - r_2}{n} \left\langle \left(d^i\right)^T \cdot d^i \middle| b^T \cdot b \right\rangle \right) = \\
\left( r_1 \cdot \langle q_v | a \rangle^2 + \frac{1 - r_1}{n} \cdot \sum_i \left\langle c^i \middle| a \right\rangle^2 \right) \cdot \\
\left( r_2 \cdot \langle q_t | b \rangle^2 + \frac{1 - r_2}{n} \cdot \sum_i \left\langle d^i \middle| b \right\rangle^2 \right)
\end{aligned} \tag{5.9}$$

The model decomposes into the weighted combinations of measurements performed on individual feature spaces. This makes the model fast and easy to implement. The squared inner products can be interpreted as the probabilities of the system transitions from one state to another (here, similarity measures).

Inspired by the original model, we can extrapolate to other similarity measures

$$\begin{aligned}
\left( r_1 \cdot s_{vis}^2 \left(q_v, a\right) + \frac{1 - r_1}{n} \cdot \sum_i s_{vis}^2 \left(c^i, a\right) \right) \cdot \\
\left( r_2 \cdot s_{text}^2 \left(q_t, b\right) + \frac{1 - r_2}{n} \cdot \sum_i s_{text}^2 \left(d^i, b\right) \right)
\end{aligned} \tag{5.10}$$

where $s_{text}\left(q_t, b\right)$ for example, represents textual similarity between the textual query (generated from image tags) representation and a textual representation of an image from the data collection.

We can also look at this model from different perspective, as an adaptive late fusion that can be utilized to capture user interactions with the retrieval system. Reverse engineering gives us then the theoretical justification for this particular combination of individual measurements.

### 5.1.1 Hybrid Relevance Feedback Model Based on the Orthogonal Projection

We can consider a variation of the aforementioned model, where just like in the original one $M_1 = r_1 \cdot D_q^v + \frac{r_2}{n} \cdot D_f^v$ and $M_2 = r_1 \cdot D_q^t + \frac{r_2}{n} \cdot D_f^t$. We can decompose (eigenvalue decomposition) the density matrices $M_1, M_2$ to estimate the bases[3] $(p_i^v, p_j^t)$ of the subspaces generated by the query and the images in the feedback set. Now, let us consider the measurement

$$\left\langle P_1 \otimes P_2 \right| \left(a^T a\right) \otimes \left(b^T b\right) \rangle \tag{5.11}$$

where $P_1$, $P_2$ are the projectors onto visual and textual subspaces generated by the query and the images in the feedback set ($\sum_i \left(p_i^v\right)^T p_i^v$, $\sum_j \left(p_j^t\right)^T p_j^t$), and $a$, $b$ are the visual and textual representations of an image from the data set. Because the tensor product of the projectors corresponding to visual and textual Hilbert spaces ($H_1$, $H_2$) is a projector onto the tensored Hilbert space ($H_1 \otimes H_2$), our similarity measurement can be interpreted as the probability of the relevance context, the probability that vector $a \otimes b$ was generated within the subspace (representing the relevance context) generated by $M_1 \otimes M_2$. Hence

$$\begin{aligned}
\left\langle P_1 \otimes P_2 \right| \left(a^T a\right) \otimes \left(b^T b\right) \rangle = \\
\left\langle P_1 | a^T a \right\rangle \cdot \left\langle P_2 | b^T b \right\rangle = \\
\left\langle \sum_i \left(p_i^v\right)^T p_i^v | a^T a \right\rangle \cdot \left\langle \sum_j \left(p_j^t\right)^T p_j^t | b^T b \right\rangle = \\
\sum_i \left\langle p_i^v | a \right\rangle^2 \cdot \sum_j \left\langle p_j^t | b \right\rangle^2 = \\
\sum_i Pr_i^v \cdot \sum_j Pr_j^t = \\
\left\| \left(\left\langle p_1^v | a \right\rangle, \ldots, \left\langle p_n^v | a \right\rangle\right) \otimes \left(\left\langle p_1^t | b \right\rangle, \ldots, \left\langle p_n^t | b \right\rangle\right) \right\|^2
\end{aligned} \tag{5.12}$$

We can see, that this measurement is equivalent to the weighted combinations of all the probabilities of projections for all the images involved. In quantum mechanics, the square of the absolute value of the inner product between the initial state and the eigenstate is the probability of the system collapsing to this eigenstate. In our case, the square of the absolute value of the inner product can be interpreted as a particular contextual factor influencing the measurement.

### 5.1.2 Hybrid Relevance Feedback Model for Image Re-ranking

Another version of the proposed hybrid relevance feedback model employs density matrices corresponding to feedback images only (no query density matrix). The quantum-like measurement

[3] It has been highlighted (Olshausen et al. 1996) that the orthogonal decomposition may not be the best option for visual spaces because the receptive fields that result from this process are not localized, and the vast majority do not at all resemble any known cortical receptive fields. Thus, in the case of visual spaces, we may want to utilize decomposition methods that produce non-orthogonal basis vectors.

is then utilized to re-rank the top images (from the first round retrieval). We have discovered that relevance feedback models which utilize both query and feedback information should employ measurement for re-scoring of the whole data collection. On the other hand, relevance feedback models which utilize the feedback information only should employ measurement for re-ranking of the top images.

Thus, the density matrices will now be generated by feedback images only

$$M_1 = \sum_i \left( \left(c^i\right)^T \cdot c^i \right) \tag{5.13}$$

and

$$M_2 = \sum_i \left( \left(d^i\right)^T \cdot d^i \right) \tag{5.14}$$

and the model will simplify to

$$\begin{aligned} \left\langle M_1 \otimes M_2 \middle| \left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right) \right\rangle = \\ \left\langle M_1 \middle| a^T \cdot a \right\rangle \cdot \left\langle M_2 \middle| b^T \cdot b \right\rangle = \\ \left\langle \sum_i \left( \left(c^i\right)^T \cdot c^i \right) \middle| a^T \cdot a \right\rangle \cdot \left\langle \sum_i \left( \left(d^i\right)^T \cdot d^i \right) \middle| b^T \cdot b \right\rangle = \\ \left( \sum_i \left\langle c^i \middle| a \right\rangle^2 \right) \cdot \left( \sum_i \left\langle d^i \middle| b \right\rangle^2 \right) \end{aligned} \tag{5.15}$$

This model was implemented in our prototype system with the interactive user interface.

### 5.1.3 Hybrid Relevance Feedback Model with Continuous and Discrete Levels of Relevance

Previously introduced hybrid relevance feedback models utilize positive relevance feedback only. Here, we also incorporate negative feedback into our hybrid relevance feedback model (we focus on the original model with query and feedback images, and data collection rescoring). Moreover, instead of restricting the relevance feedback to binary levels of relevance (relevant/not relevant), we exploit discrete and continuous levels of relevance. We can envision a relevance bar representing a spectrum of possible relevance levels (see Figure 5.1). A user can drag and drop feedback images on this spectrum according to their relevance degree.


Figure 5.1: Relevance bar

The first model incorporates discrete levels of relevance (we assume an arbitrary number of

relevance levels, i.e. 9)

$$w_i = -1 + 2 \cdot \frac{im_iPos}{8} \tag{5.16}$$

$$im_iPos = 0, 1, ..., 8 \tag{5.17}$$

$$w_i \in [-1, 1] \tag{5.18}$$

In the above, $i$ denotes a feedback image and $im_iPos$ denotes the position of this particular feedback image on the relevance spectrum. The discrete levels of relevance are shown by Figure 5.2.

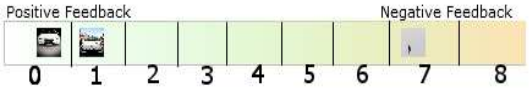


Figure 5.2: Relevance bar, discrete levels of relevance

Our second model incorporates continuous levels of relevance

$$w_i = -1 + 2 \cdot dist\,(0, im_iPos) \tag{5.19}$$

$$im_iPos = 0, 1, ..., 8 \tag{5.20}$$

$$w_i \in [-1, 1] \tag{5.21}$$

where $dist(0, im_iPos)$ represents distance from zero to the feedback image position on the relevance spectrum as shown in Figure 5.3.


Figure 5.3: Relevance bar, continuous levels of relevance

Now, thus calculated weights $w_i$ can be utilized to modify the importance of density matrices corresponding to individual feedback images. Let us remember our previously discovered property that a density matrix of the system (image-level correlation) can be decomposed into a sum of auto-density matrices (image-level auto-correlation) corresponding to all the individual images in this system. Thus

$$M_1 = r_1 \cdot D_q^v + \frac{r_2}{n} \cdot D_{f-pos}^v + \frac{r_3}{m} \cdot D_{f-neg}^v =$$

$$r_1 \cdot D_q^v + \frac{r_2}{n} \sum_i w_i D_i^v + \frac{r_3}{m} \sum_j w_j D_j^v =$$

$$r_1 \cdot q_v^T \cdot q_v + \frac{r_2}{n} \sum_i w_i \left( \left(c^i\right)^T \cdot c^i \right) + \frac{r_3}{m} \sum_j w_j \left( \left(c^j\right)^T \cdot c^j \right) \tag{5.22}$$

$$r_1 + r_2 + r_3 = 1 \tag{5.23}$$

and

$$M_2 = r'_1 \cdot D^t_q + \frac{r'_2}{n} \cdot D^t_{f-pos} + \frac{r'_3}{m} \cdot D^t_{f-neg} =$$

$$r'_1 \cdot D^t_q + \frac{r'_2}{n} \sum_i w_i D^t_i + \frac{r'_3}{m} \sum_j w_j D^t_j =$$

$$r'_1 \cdot q^T_t \cdot q_t + \frac{r'_2}{n} \sum_i w_i \left( \left(d^i\right)^T \cdot d^i \right) + \frac{r'_3}{m} \sum_j w_j \left( \left(d^j\right)^T \cdot d^j \right) \tag{5.24}$$

$$r'_1 + r'_2 + r'_3 = 1 \tag{5.25}$$

where $M_1$, $M_2$ denote the visual and textual density matrices corresponding to subspaces of query and feedback images, triples $r_1$, $r_2$, $r_3$ and $r'_1$, $r'_2$, $r'_3$ represent the importance of the query and the feedback images in both visual and textual feature spaces, $D_q$, $D_{f-pos}$, $D_{f-neg}$ represent density matrices of the query, positive feedback images and negative feedback images respectively, for visual and textual feature spaces. Entities $D_i$ and $D_j$ correspond to the individual feedback images auto-density matrices, $w_i$ and $w_j$ denote the weights representing the levels of relevance and irrelevance of particular feedback images, respectively, and $n$ and $m$ denote the number of positive and negative feedback images, respectively. As in the original model, $q_v$, $q_t$ denote the visual and textual representations of the query, $c^i$, $d^i$ and $c^j$, $d^j$ are the visual and textual representations of the positive and negative feedback images, respectively.

Now we can modify the original model to incorporate various levels of relevance. Hence

$$\left\langle M_1 \otimes M_2 \middle| \left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right) \right\rangle =$$

$$\left( r_1 \cdot \langle q_v | a \rangle^2 + \frac{r_2}{n} \cdot \sum_i w_i \left\langle c^i \middle| a \right\rangle^2 + \frac{r_3}{m} \cdot \sum_j w_j \left\langle c^j \middle| a \right\rangle^2 \right) \cdot$$

$$\left( r'_1 \cdot \langle q_t | b \rangle^2 + \frac{r'_2}{n} \cdot \sum_i w_i \left\langle d^i \middle| b \right\rangle^2 + \frac{r'_3}{m} \cdot \sum_j w_j \left\langle d^j \middle| b \right\rangle^2 \right) \tag{5.26}$$

$$\forall i : im_i Pos = 0, 1, ..., 3 \ \ (discrete) \tag{5.27}$$

$$\forall i : im_i Pos \in (0.5, 1] \ \ (continuous) \tag{5.28}$$

$$\forall j : im_j Pos = 5, 6, ..., 8 \ \ (discrete) \tag{5.29}$$

$$\forall j : im_j Pos \in [0, 0.5) \ \ (continuous) \tag{5.30}$$

This way we can naturally incorporate various levels of relevance into our hybrid relevance feedback model. More levels of relevance gives more freedom and accuracy in adapting and narrowing down the search.

## 5.2 On the Relationship Between Fusion Strategies

Here, we would like to refer again to the relationships between fusion strategies. As shown previously, factors $M_1 \otimes M_2$ and $\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)$ represent a relatively complex early fusion strategy (combination at the representation level). However, the interaction between the measurement (an inner product, our similarity measure) and the early fusion operators (here, the tensor product and the image-level notion of correlation) results in a combination of features at the decision level (late fusion). In other words, the measurement performed on the combined representations is equivalent to the relatively complex combination of measurements performed on individual feature spaces. We have discovered similar relationships for many widely used combinations of different measurements (similarity measures) and late fusion operators.

Researchers who utilize early fusion strategies do not take the interaction between the measurements and early fusion operators into account. This does not mean that their models are flawed. What this means is that in many cases the early fusion can be represented as a late fusion strategy and vice versa. Knowledge of this interchangeability could make the models easy to implement and significantly faster (computations performed on individual feature spaces).

Finally, it is commonly believed that late fusion cannot capture the correlation between different feature spaces. Considering the aforementioned interchangeability, does the interaction between certain measurements and certain early fusion operators decorrelate features, or is the belief unjustified and late fusion strategy can capture the correlations between different feature spaces. We are going to consider this interesting issue as an open question for further discussion.

## 5.3 Dynamic Weighting of Query and Its Context in the Relevance Hybrid Model

In the original model, the weights corresponding to textual query, textual context, visual query and visual context are fixed[4] (i.e. $r_1$, $1 - r_1$, $r_2$, $1 - r_2$, across all the queries). However, it has been highlighted (Teevan et al. 2005), that the query may be more or less related to its context.

We can further improve the feedback model by adjusting these weights with respect to the issued queries and feedback images based on the strength of the relationship between the query and the context.

Let us measure the strength of the aforementioned relationship by computing the similarity between co-occurrence matrices corresponding to the query and its context (feedback images). The higher the number of terms or visual terms (mid-level features) co-occurring between the current query and the context, the stronger the relationship and vice versa. Thus, the relationship

[4] In this paper, terms "textual query", "visual context" etc. refer to the textual (visual) representation of a query image (context images).

strength between the visual query and visual context can be measured as

$$\left\langle D_q^v \middle| D_f^v \right\rangle = \left\langle q_v^T \cdot q_v \middle| \sum_i \left(c^i\right)^T \cdot c^i \right\rangle = \sum_i \left\langle q_v \otimes q_v \middle| c^i \otimes c^i \right\rangle = \sum_i \left\langle q_v \middle| c^i \right\rangle^2 \tag{5.31}$$

Similarly, the relationship strength between the textual query and its textual context can be computed as

$$\left\langle D_q^t \middle| D_f^t \right\rangle = \sum_i \left\langle q_t \middle| d^i \right\rangle^2 \tag{5.32}$$

We can normalize these measurements. Thus, we can compute

$$\frac{\left\langle D_q^v \middle| D_f^v \right\rangle}{\left\| D_q^v \right\| \cdot \left\| D_f^v \right\|} \tag{5.33}$$

$$\frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\| D_q^t \right\| \cdot \left\| D_f^t \right\|} \tag{5.34}$$

where

$$\left\| D_q^v \right\| = \sqrt{\left\langle D_q^v \middle| D_q^v \right\rangle} = \sqrt{\langle q_v^T \cdot q_v | q_v^T \cdot q_v \rangle} = \sqrt{\langle q_v \otimes q_v | q_v \otimes q_v \rangle} = \sqrt{\langle q_v | q_v \rangle^2} = \langle q_v | q_v \rangle \tag{5.35}$$

and

$$\left\| D_f^v \right\| = \sqrt{\left\langle D_f^v \middle| D_f^v \right\rangle} = \sqrt{\left\langle \sum_i \left(c^i\right)^T \cdot c^i \middle| \sum_i \left(c^i\right)^T \cdot c^i \right\rangle} = \sqrt{\sum_i \sum_i \langle c^i \otimes c^i | c^i \otimes c^i \rangle} = \sqrt{n \cdot \sum_i \langle c^i | c^i \rangle^2} \tag{5.36}$$

Analogically, for the textual part

$$\left\|D_q^t\right\| = \langle q_t|q_t\rangle \tag{5.37}$$

and

$$\left\|D_f^t\right\| = \sqrt{n \cdot \sum_i \langle d^i|d^i\rangle^2} \tag{5.38}$$

Thus, the modified model becomes

$$\begin{gathered} tr\left(\left(M_1 \otimes M_2\right)\left(\left(a^T a\right) \otimes \left(b^T b\right)\right)\right) = \\ \left(str_v \langle q_v|a\rangle^2 + \left(1 - str_v\right)\frac{1}{n}\sum_i \left\langle c^i\middle|a\right\rangle^2\right) \cdot \\ \left(str_t \langle q_t|b\rangle^2 + \left(1 - str_t\right)\frac{1}{n}\sum_i \left\langle d^i\middle|b\right\rangle^2\right) \end{gathered} \tag{5.39}$$

where

$$str_v = \frac{\left\langle D_q^v\middle|D_f^v\right\rangle}{\left\|D_q^v\right\|\left\|D_f^v\right\|} = \frac{\sum_i \left\langle q_v\middle|c^i\right\rangle^2}{\langle q_v|q_v\rangle\sqrt{n\sum_i \langle c^i|c^i\rangle^2}} \tag{5.40}$$

and

$$str_t = \frac{\left\langle D_q^t\middle|D_f^t\right\rangle}{\left\|D_q^t\right\|\left\|D_f^t\right\|} = \frac{\sum_i \left\langle q_t\middle|d^i\right\rangle^2}{\langle q_t|q_t\rangle\sqrt{n\sum_i \langle d^i|d^i\rangle^2}} \tag{5.41}$$

Let us assume, that the relevance feedback is given after the first round retrieval to refine the query. The adaptive weighting can be interpreted in the following way:

1. small $\langle D_q|D_f\rangle$; weak relationship between query and its context, context becomes important. We adjust the probability of the original query terms, the adjustment will significantly modify the original query.

2. big $\langle D_q|D_f\rangle$; strong relationship (similarity) between query and its context, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

The experiments conducted on the ImageClef data collection have shown that the adaptive weighting can indeed outperform the fixed weighting scheme (in the context of hybrid models and user feedback).

### 5.3.1 Generalization of the Model

This enhanced model can be naturally expanded to accommodate other features, i.e. various visual features

$$tr\left(\left(\otimes_n M_n\right)\cdot\left(\otimes_n\left(a_n^T\cdot a_n\right)\right)\right)=\prod_n\left\langle M_n\middle|a_n^T\cdot a_n\right\rangle \tag{5.42}$$

Thus, for 3 features (i.e. two visual and a textual feature) our enhanced model becomes

$$\begin{aligned}tr\left(\left(M_1\otimes M_2\otimes M_3\right)\left(\left(a_1^Ta_1\right)\otimes\left(a_2^Ta_2\right)\otimes\left(b^Tb\right)\right)\right)=\\ \left(str_{v1}\left\langle q_{v1}\middle|a_1\right\rangle^2+\left(1-str_{v1}\right)\frac{1}{n}\sum_i\left\langle c_1^i\middle|a_1\right\rangle^2\right)\cdot\\ \left(str_{v2}\left\langle q_{v2}\middle|a_2\right\rangle^2+\left(1-str_{v2}\right)\frac{1}{n}\sum_i\left\langle c_2^i\middle|a_2\right\rangle^2\right)\cdot\\ \left(str_t\left\langle q_t\middle|b\right\rangle^2+\left(1-str_t\right)\frac{1}{n}\sum_i\left\langle d^i\middle|b\right\rangle^2\right)\end{aligned} \tag{5.43}$$

where

$$str_{v1}=\frac{\left\langle D_q^{v1}\middle|D_f^{v1}\right\rangle}{\left\|D_q^{v1}\right\|\left\|D_f^{v1}\right\|}=\frac{\sum_i\left\langle q_{v1}\middle|c_1^i\right\rangle^2}{\left\langle q_{v1}\middle|q_{v1}\right\rangle\sqrt{n\sum_i\left\langle c_1^i\middle|c_1^i\right\rangle^2}} \tag{5.44}$$

and

$$str_{v2}=\frac{\left\langle D_q^{v2}\middle|D_f^{v2}\right\rangle}{\left\|D_q^{v2}\right\|\left\|D_f^{v2}\right\|}=\frac{\sum_i\left\langle q_{v2}\middle|c_2^i\right\rangle^2}{\left\langle q_{v2}\middle|q_{v2}\right\rangle\sqrt{n\sum_i\left\langle c_2^i\middle|c_2^i\right\rangle^2}} \tag{5.45}$$

and

$$str_t=\frac{\left\langle D_q^t\middle|D_f^t\right\rangle}{\left\|D_q^t\right\|\left\|D_f^t\right\|}=\frac{\sum_i\left\langle q_t\middle|d^i\right\rangle^2}{\left\langle q_t\middle|q_t\right\rangle\sqrt{n\sum_i\left\langle d^i\middle|d^i\right\rangle^2}} \tag{5.46}$$

Here, for example, $M_1$, $a_1$ and $M_2$, $a_2$ may correspond to different visual features (density matrices and vector representations of images from the data collection), and $M_3$, $b$ corresponds to a textual feature.

## 5.4 Query Modification as a Combination of Relevance Scores

Query modification strategies utilize relevance feedback to modify the query in order to narrow down the search. Here, we show that the query modification can be represented as a late fusion strategy. This observation has a few important implications. First, some complex combinations of measurements performed on individual feature spaces may be regarded as a query modification technique. Second, query modification represented as a late fusion does not require an actual modification of query representation. The actual query modification would often lead to query renormalization. While normalization in text IR is often desirable, normalization of mid-level visual representations may even hamper the retrieval performance. Third, it is often easier to implement and work with relevance scores. Finally, knowing how to represent query modification as a late fusion can help us develop a family of hybrid relevance models (combinations of scores computed on lower dimensional feature spaces as opposed to high dimensional hybrid representations).

### 5.4.1 Similarity Measurements and Query Modification

Query reformulation techniques are often used in multimedia retrieval to narrow down the search based on the relevance feedback. We are going to show that the Rocchio and "Rocchio-like" query modification algorithms can be represented as a late fusion, a combination of a number of individual relevance scores. This observation will also help us generate hybrid Rocchio models, which can combine visual and textual sub-systems in the context of relevance feedback.

#### Aggregated Representations and Late Fusion

Query modification models often aggregate tensored or concatenated representations. Based on the previously discovered properties, we can show that these aggregations can be represented as aggregated relevance scores.

#### Aggregated Representations and Late Fusion. Inner Product

$$\left\langle \sum_i \left(d_i^v \oplus d_i^t\right) | d^v \oplus d^t \right\rangle = \sum_i \left( \langle d_i^v | d^v \rangle + \left\langle d_i^t | d^t \right\rangle \right) \tag{5.47}$$

$$\left\langle \sum_i \left(d_i^v \otimes d_i^t\right) | d^v \otimes d^t \right\rangle = \sum_i \left( \langle d_i^v | d^v \rangle \cdot \left\langle d_i^t | d^t \right\rangle \right) \tag{5.48}$$

*Proof.* (5.47)

$$\left\langle \sum_i \left(d_i^v \oplus d_i^t\right) | d^v \oplus d^t \right\rangle = \sum_i \left\langle d_i^v \oplus d_i^t | d^v \oplus d^t \right\rangle = \sum_i \left( \langle d_i^v | d^v \rangle + \left\langle d_i^t | d^t \right\rangle \right) \tag{5.49}$$

□

(5.48)

$$\left\langle \sum_i \left(d_i^v \otimes d_i^t\right) | d^v \otimes d^t \right\rangle = \sum_i \left\langle d_i^v \otimes d_i^t | d^v \otimes d^t \right\rangle = \sum_i \left( \langle d_i^v | d^v \rangle \cdot \left\langle d_i^t | d^t \right\rangle \right) \tag{5.50}$$

□

**Aggregated Representations and Late Fusion. Cosine Similarity**

$$s_c \left( \sum_i \left(d_i^v \oplus d_i^t\right), d^v \oplus d^t \right) = \frac{1}{2n} \sum_i \left( s_c \left(d_i^v, d^v\right) + s_c \left(d_i^t, d^t\right) \right) \tag{5.51}$$

$$s_c \left( \sum_i \left(d_i^v \otimes d_i^t\right), d^v \otimes d^t \right) = \frac{1}{n} \sum_i \left( s_c \left(d_i^v, d^v\right) \cdot s_c \left(d_i^t, d^t\right) \right) \tag{5.52}$$

*Proof.* (5.51) Since

$$\left\| \sum_i \left(d_i^v \oplus d_i^t\right) \right\| = \sqrt{\left\langle \sum_i \left(d_i^v \oplus d_i^t\right) | \sum_i \left(d_i^v \oplus d_i^t\right) \right\rangle} = \sqrt{\sum_i \sum_i \left( \|d_i^v\|^2 + \|d_i^t\|^2 \right)} = \sqrt{n^2 \cdot 2} = n\sqrt{2} \tag{5.53}$$

we get

$$s_c\left(\sum_i \left(d_i^v \oplus d_i^t\right), d^v \oplus d^t\right) = \frac{\left\langle \sum_i \left(d_i^v \oplus d_i^t\right), d^v \oplus d^t \right\rangle}{\left\| \sum_i d_i^v \oplus d_i^t \right\| \cdot \left\| d^v \oplus d^t \right\|} =$$
$$\frac{\sum_i \left(\langle d_i^v | d^v \rangle + \left\langle d_i^t | d^t \right\rangle\right)}{n\sqrt{2} \cdot \sqrt{2}} = \frac{1}{2n}\left(\sum_i \frac{\langle d_i^v | d^v \rangle}{\|d_i^v\| \cdot \|d^v\|} + \sum_i \frac{\left\langle d_i^t | d^t \right\rangle}{\|d_i^t\| \cdot \|d^t\|}\right) =$$
$$\frac{1}{2n}\left(\sum_i \left(sc\left(d_i^v, d^v\right) + s_c\left(d_i^t, d^t\right)\right)\right) \tag{5.54}$$

□

(5.52) Similarly, because

$$\left\| \sum_i \left(d_i^v \otimes d_i^t\right) \right\| = \sqrt{\sum_i \sum_i \left\|d_i^v\right\|^2 \cdot \left\|d_i^t\right\|^2} = n \tag{5.55}$$

we can show, that

$$s_c\left(\sum_i \left(d_i^v \otimes d_i^t\right), d^v \otimes d^t\right) = \frac{\sum_i \left(\langle d_i^v | d^v \rangle \cdot \left\langle d_i^t | d^t \right\rangle\right)}{n \cdot 1} =$$
$$\frac{1}{n}\left(\sum_i \left(\frac{\langle d_i^v | d^v \rangle}{\|d_i^v\| \cdot \|d^v\|} \cdot \frac{\left\langle d_i^t | d^t \right\rangle}{\|d_i^t\| \cdot \|d^t\|}\right)\right) = \frac{1}{n}\left(\sum_i \left(s_c\left(d_i^v, d^v\right) \cdot s_c\left(d_i^t, d^t\right)\right)\right) \tag{5.56}$$

□

**Aggregated Representations and Late Fusion. Euclidean Metric**

$$s_e\left(\sum_i \left(d_i^v \oplus d_i^t\right), d^v \oplus d^t\right) = \sqrt{2\left(n-1\right)^2 + \sum_i \left(s_e^2\left(d_i^v, d^v\right) + s_e^2\left(d_i^t, d^t\right)\right)} \tag{5.57}$$

$$s_e\left(\sum_i \left(d_i^v \otimes d_i^t\right), d^v \otimes d^t\right) =$$
$$\sqrt{\left(n-1\right)^2 + \sum_i \left(s_e^2\left(d_i^v, d^v\right) + s_e^2\left(d_i^t, d^t\right) - \frac{1}{2} s_e^2\left(d_i^v, d^v\right) \cdot s_e^2\left(d_i^t, d^t\right)\right)} \tag{5.58}$$

where $i = 1, \ldots, n$.

*Proof.* (5.57)

$$s_e\left(\sum_i \left(d_i^v \oplus d_i^t\right), d^v \oplus d^t\right) = \sqrt{\left\langle \sum_i \left(d_i^v \oplus d_i^t\right) | \sum_i \left(d_i^v \oplus d_i^t\right) \right\rangle + \langle d^v \oplus d^t | d^v \oplus d^t \rangle - 2\left\langle \sum_i \left(d_i^v \oplus d_i^t\right) | d^v \oplus d^t \right\rangle} = \sqrt{2n^2 + 2 - 2\sum_i \left(\langle d_i^v | d^v \rangle + \langle d_i^t | d^t \rangle\right)} = \sqrt{2\left(n-1\right)^2 + \sum_i \left(s_e^2\left(d_i^v, d^v\right) + s_e^2\left(d_i^t, d^t\right)\right)} \tag{5.59}$$

□

(5.58) Similarly, for the tensor product

$$s_e\left(\sum_i \left(d_i^v \oplus d_i^t\right), d^v \oplus d^t\right) = \sqrt{n^2 + 1 - 2\sum_i \left(\langle d_i^v | d^v \rangle \langle d_i^t | d^t \rangle\right)} = \sqrt{\left(n-1\right)^2 + \sum_i \left(s_e^2\left(d_i^v, d^v\right) + s_e^2\left(d_i^t, d^t\right) - \frac{1}{2} s_e^2\left(d_i^v, d^v\right) s_e^2\left(d_i^t, d^t\right)\right)} \tag{5.60}$$

□

**Rocchio Algorithm and Late Fusion**

The Rocchio algorithm modifies the query so that it moves closer to the centroid of relevant documents and further away from the centroid of irrelevant ones

$$Q_m = \left(a \cdot Q_o\right) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{D_j \in D_r} D_j\right) - \left(c \cdot \frac{1}{D_{nr}} \cdot \sum_{D_k \in D_{nr}} D_k\right) \tag{5.61}$$

where
$Q_m$ - modified query vector
$Q_o$ - original query vector
$D_j$ - related document vector
$D_k$ - non-related document vector
$a$ - original query weight
$b$ - related documents' weights
$c$ - non-related documents' weights
$D_r$ - set of related documents
$D_{nr}$ - set of non-related documents

We will show that the modification of the query can be interpreted as a weighted combination

of the measurements (scores, similarities) between a query and a document from the data collection and between a query and each document from the feedback set. After modifying the query, we need to re-compute the scores

$$\begin{aligned}
\langle Q_m|Q_d\rangle = \\
\left\langle aQ_o + b\frac{1}{|D_r|}\sum_{D_j\in D_r} D_j - c\frac{1}{|D_{nr}|}\sum_{D_k\in D_{nr}} D_k|Q_d \right\rangle = \\
\langle aQ_o|Q_d\rangle + \left\langle \frac{b}{|D_r|}\sum_{D_j\in D_r} D_j|Q_d \right\rangle - \left\langle \frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}} D_k|Q_d \right\rangle = \\
a\langle Q_o|Q_d\rangle + \frac{b}{|D_r|}\sum_{D_j\in D_r}\langle D_j|Q_d\rangle - \frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}}\langle D_k|Q_d\rangle
\end{aligned} \tag{5.62}$$

And similarly for other similarity measures. For the cosine similarity, we get

$$\begin{aligned}
s_c(Q_m, Q_d) = \\
\frac{1}{\|Q_m\|}\left( as_c(Q_o, Q_d) + \frac{b}{|D_r|}\sum_j s_c(D_j, Q_d) - \frac{c}{|D_{nr}|}\sum_k s_c(D_k, Q_d)\right) \\
\|Q_m\|^2 = a^2 + c^2 + \frac{2ab}{|D_r|}\sum_j s_c(Q_o, D_j) - \frac{2ac}{|D_{nr}|}\sum_k s_c(Q_o, D_k) - \\
\frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_c(D_j, D_k)
\end{aligned} \tag{5.63}$$

*Proof.*

$$\begin{aligned}
s_c(Q_m, Q_d) = \frac{\langle Q_m|Q_d\rangle}{\|Q_m\|\cdot\|Q_d\|} = \\
\frac{1}{\|Q_m\|}\left( as_c(Q_o, Q_d) + \frac{b}{|D_r|}\sum_j s_c(D_j, Q_d) - \frac{c}{|D_{nr}|}\sum_k s_c(D_k, Q_d)\right)
\end{aligned} \tag{5.64}$$

where

$$\begin{aligned}
\|Q_m\|^2 = \langle Q_m|Q_m\rangle = \\
\left\langle aQ_o + \frac{b}{|D_r|}\sum_j D_j - \frac{c}{|D_{nr}|}\sum_k D_k | aQ_o + \frac{b}{|D_r|}\sum_j D_j - \frac{c}{|D_{nr}|}\sum_k D_k \right\rangle = \\
a^2\langle Q_o|Q_o\rangle + \frac{2ab}{|D_r|}\sum_j \langle Q_o|D_j\rangle - \frac{2ac}{|D_{nr}|}\sum_k \langle Q_o|D_k\rangle - \\
\frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k \langle D_j|D_k\rangle + \frac{c^2}{|D_{nr}|^2}\sum_k\sum_k \langle D_k|D_k\rangle = \\
a^2 + c^2 + \frac{2ab}{|D_r|}\sum_j s_c(Q_o, D_j) - \frac{2ac}{|D_{nr}|}\sum_k s_c(Q_o, D_k) - \\
\frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_c(D_j, D_k)
\end{aligned} \tag{5.65}$$

□

For the Euclidean distance

$$\begin{aligned}
s_e^2(Q_m, Q_d) = a^2 + c^2 + 2ab - 2ac - 2bc - 2a - 2b - 2c + 1 - \\
\frac{ab}{|D_r|}\sum_j s_e(Q_o, D_j) + \frac{ac}{|D_{nr}|}\sum_k s_e(Q_o, D_k) + \\
\frac{bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_e(D_j, D_k) + \\
as_e(Q_o, Q_d) + \frac{b}{|D_r|}\sum_j s_e(D_j, Q_d) + \frac{c}{|D_{nr}|}\sum_k s_e(D_k, Q_d)
\end{aligned} \tag{5.66}$$

*Proof.* Based on (5.63), we get

$$
\begin{aligned}
s_e^2\left(Q_m, Q_d\right) = \\
a^2 + c^2 + \frac{2ab}{|D_r|}\sum_j \langle Q_o|D_j\rangle - \frac{2ac}{|D_{nr}|}\sum_k \langle Q_o|D_k\rangle - \frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k \langle D_j|D_k\rangle + \\
1 - 2a\langle Q_o|Q_d\rangle - \frac{2b}{|D_r|}\sum_j \langle D_j|Q_d\rangle - \frac{2c}{|D_{nr}|}\sum_k \langle D_k|Q_d\rangle = \\
a^2 + c^2 + 1 + \frac{2ab}{|D_r|}|D_r| - \frac{2ab}{|D_r|}|D_r| + \frac{2ab}{|D_r|}\sum_j \langle Q_o|D_j\rangle + \\
\frac{2ac}{|D_{nr}|}|D_{nr}| - \frac{2ac}{|D_{nr}|}|D_{nr}| - \frac{2ac}{|D_{nr}|}\sum_k \langle Q_o|D_k\rangle + \\
\frac{2bc}{|D_r|\cdot|D_{nr}|}|D_r|\cdot|D_{nr}| - \frac{2bc}{|D_r|\cdot|D_{nr}|}|D_r|\cdot|D_{nr}| - \frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k \langle D_j|D_k\rangle + \\
2a - 2a - 2a\langle Q_o|Q_d\rangle + \frac{2b}{|D_r|}|D_r| - \frac{2b}{|D_r|}|D_r| - \frac{2b}{|D_r|}\sum_j \langle D_j|Q_d\rangle + \\
\frac{2c}{|D_{nr}|}|D_{nr}| - \frac{2c}{|D_{nr}|}|D_{nr}| - \frac{2c}{|D_{nr}|}\sum_k \langle D_k|Q_d\rangle = \\
a^2 + c^2 + 2ab - 2ac - 2bc - 2a - 2b - 2c + 1 - \\
\frac{ab}{|D_r|}\sum_j s_e\left(Q_o, D_j\right) + \frac{ac}{|D_{nr}|}\sum_k s_e\left(Q_o, D_k\right) + \frac{bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_e\left(D_j, D_k\right) + \\
as_e\left(Q_o, Q_d\right) + \frac{b}{|D_r|}\sum_j s_e\left(D_j, Q_d\right) + \frac{c}{|D_{nr}|}\sum_k s_e\left(D_k, Q_d\right)
\end{aligned}
\tag{5.67}
$$

□

### 5.4.2 Generation of Hybrid Relevance Feedback Models

We can tensor or concatenate the modified query vectors in order to generate hybrid models. Then ($v$, $t$ indexes denote visual and textual representations respectively)

$$
\begin{aligned}
\left\langle Q_m^v \otimes Q_m^t | Q_d^v \otimes Q_d^t \right\rangle = \left\langle Q_m^v | Q_d^v \right\rangle \left\langle Q_m^t | Q_d^t \right\rangle = \\
(a\left\langle Q_o^v | Q_d^v \right\rangle + \frac{b}{|D_r|}\sum_{D_j \in D_r}\left\langle D_j^v | Q_d^v \right\rangle - \frac{c}{|D_{nr}|}\sum_{D_k \in D_{nr}}\langle D_k^v | Q_d^v\rangle)\cdot \\
(a\left\langle Q_o^t | Q_d^t \right\rangle + \frac{b}{|D_r|}\sum_{D_j \in D_r}\left\langle D_j^t | Q_d^t \right\rangle - \frac{c}{|D_{nr}|}\sum_{D_k \in D_{nr}}\left\langle D_k^t | Q_d^t \right\rangle)
\end{aligned}
\tag{5.68}
$$

and for concatenation

$$\left\langle Q_m^v \oplus Q_m^t | Q_d^v \oplus Q_d^t \right\rangle = \langle Q_m^v | Q_d^v \rangle + \left\langle Q_m^t | Q_d^t \right\rangle = a\left(\langle Q_o^v | Q_d^v \rangle + \left\langle Q_o^t | Q_d^t \right\rangle\right) + \frac{b}{|D_r|} \sum_{D_j \in D_r} \left(\left\langle D_j^v | Q_d^v \right\rangle + \left\langle D_j^t | Q_d^t \right\rangle\right) - \frac{c}{|D_{nr}|} \sum_{D_k \in D_{nr}} \left(\langle D_k^v | Q_d^v \rangle + \left\langle D_k^t | Q_d^t \right\rangle\right) \tag{5.69}$$

We can use other similarity measures (and substitute individual measurements with (5.63) and (5.66))

$$s_c\left(Q_m^v \otimes Q_m^t, Q_d^v \otimes Q_d^t\right) = s_c\left(Q_m^v, Q_d^v\right) \cdot s_c\left(Q_m^t, Q_d^t\right) \tag{5.70}$$

$$s_c\left(Q_m^v \oplus Q_m^t, Q_d^v \oplus Q_d^t\right) = \frac{1}{2}\left(s_c\left(Q_m^v, Q_d^v\right) + s_c\left(Q_m^t, Q_d^t\right)\right) \tag{5.71}$$

$$s_e\left(Q_m^v \otimes Q_m^t, Q_d^v \otimes Q_d^t\right) = \sqrt{s_e^2\left(Q_m^v, Q_d^v\right) + s_e^2\left(Q_m^t, Q_d^t\right) - \frac{1}{2} s_e^2\left(Q_m^v, Q_d^v\right) s_e^2\left(Q_m^t, Q_d^t\right)} \tag{5.72}$$

$$s_e\left(Q_m^v \oplus Q_m^t, Q_d^v \oplus Q_d^t\right) = \sqrt{s_e^2\left(Q_m^v, Q_d^v\right) + s_e^2\left(Q_m^t, Q_d^t\right)} \tag{5.73}$$

$$s_e\left(Q_m^v \oplus Q_m^t, Q_d^v \oplus Q_d^t\right) = \sqrt{s_e^2\left(Q_m^v, Q_d^v\right) - 2 s_c\left(Q_m^t, Q_d^t\right) + 2} \tag{5.74}$$

$$s_e\left(Q_m^v \otimes Q_m^t, Q_d^v \otimes Q_d^t\right) = \sqrt{s_e^2\left(Q_m^v, Q_d^v\right) s_c\left(Q_m^t, Q_d^t\right) - 2 s_c\left(Q_m^t, Q_d^t\right) + 2} \tag{5.75}$$

where the last formula would be a suggested combination choice (Euclidean distance for measuring the similarity between visual representations and cosine similarity for textual representations).

Hence, the modifications of the query vector and presented combinations of query vectors corresponding to different feature spaces would all decompose into products or sums of the weighted combinations of individual measurements. These models can be used to combine visual and textual systems within the context of user feedback, for example.

## 5.5 Chapter Summary

In this chapter, we propose a novel hybrid relevance feedback model for the combination of visual and textual feature spaces in the context of relevance feedback.

Existing hybrid relevance feedback models, similarly to the hybrid models, combine the features in a rather heuristic manner. They do not exploit the inter and intra feature relationships intrinsic to both feature spaces. The intra relationships correspond to individual features subspaces, while the inter relationships correspond to correlation and complementarity of both visual and textual feature spaces.

We introduce a novel model for the combination of visual and textual sub-systems within the relevance feedback context. The model was inspired by the measurement utilized in quantum

mechanics (QM) and the tensor product of co-occurrence (density) matrices, which represents a density matrix of the composite system in QM. It provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context. The proposed approach takes into account both intra (via co-occurrence matrices) and inter (via tensor operator) relationships between features' dimensions. It is also computationally cheap and scalable to large data collections.

In this chapter, we also propose three variations of the original hybrid relevance feedback model. One is an orthogonal projection based approach. Another variation builds the density matrices from the feedback images only and utilizes the inter and intra feature relationships between feature spaces to re-rank images. Finally, the last variation integrates different levels of relevance into the original model in a natural way.

Because a query can be more or less related to its context, and because it is desirable not to use arbitrary weights in the original hybrid model, we propose a dynamic weighting scheme.

Thus, the respective weights corresponding to a query and its context (feedback images) are automatically modified, depending on the relationship strength between visual query and its visual context and textual query and its textual context; the number of terms or visual terms (mid-level visual features) co-occurring between current query and the context.

In addition, this chapter presents a generalization of the hybrid relevance feedback model to multiple visual features and a textual feature. This generalization was utilized in our prototype system.

Finally, in this chapter we continue our theoretical investigation of fusion strategies. Here, we show that the query modification approaches can be represented as combinations of relevance scores (complex late fusion). This discovery has a few important implications. First, some complex combinations of measurements performed on individual feature spaces may be regarded as a query modification technique. Second, query modification represented as a late fusion does not require an actual modification of query representation. The actual query modification would often lead to query renormalization. While normalization in text IR is often desirable, normalization of mid-level visual representations may even hamper the retrieval performance. Third, it is often easier to implement and work with relevance scores. Finally, knowing how to represent query modification as a late fusion can help us develop a family of hybrid relevance feedback models (combinations of scores computed on lower dimensional feature spaces as opposed to high dimensional hybrid representations). Here, we present a family of such hybrid relevance feedback models.

# Chapter 6

# Experimental Evaluation

In addition to the theoretical validation of our observations in the previous chapters, an experimental evaluation of the implemented models will be presented here.

We will start by describing the data collections utilized in the experiments. We test the models on three different data collections, apart from the user simulation framework, where the realistic ground truth data is needed (not just a list of general categories).

We evaluate our novel global and local features on ImageCLEF2007, MIRFlickr, and BGS data collections. The local features were also evaluated in the ImageCLEF2010 Wikipedia Retrieval Task. In addition, we test the performance of local versus global visual features as well as the performance of various sampling techniques for the Bag of Visual Words framework.

The enhancement of the local features based on the correlations between the visual words was tested within the pseudo relevance feedback framework. This is the fully automated method that helped us improve the standard local features model.

The combination of textual and visual features, as well as the auto-annotation model, were evaluated in a small-scale experiment first. Next, the proper large-scale evaluation of the hybrid model was performed as part of our participation in the ImageCLEF2010 competition.

Our hybrid relevance model (and the model with the adaptive weighting scheme) for the combination of textual and visual features in the context of relevance feedback was tested within the user simulation framework.

## 6.1 Datasets

We utilize three image collections in our experiments, namely MIRFlickr, ImageCLEF and BGS. The collections differ significantly in terms of image diversity and some visual properties. We also participated in the ImageCLEF2010 Wikipedia Retrieval Task, which allowed us to test some of our models on a large scale image collection. Here, we describe the collections in more detail.

### 6.1.1 ImageCLEFphoto2007

ImageCLEFphoto2007 consists of 20000 still natural images taken from locations around the world. It is a standard collection used by the Information Retrieval (IR) community for evaluation purposes. This allows comparison with published results. The images comprise different sports and actions, people, animals, cities, landscapes and many other aspects of contemporary life. The history, design, and implementation of this particular data collection is described in (Grubinger, Clough, Hanbury & Müller 2008). This data collection is freely available to the participants of the ImageCLEF competition. It can be downloaded from (Author n.d.a).

There are 60 query topics provided, along with the ground truth data. Example topics are shown in Table 6.1.

Table 6.1: Example topics in the ImageCLEFphoto2007 data collection

| |
|---|
| accommodation with swimming pool |
| church with more than two towers |
| religious statue in the foreground |
| people with a flag |
| straight road in the USA |
| group standing in salt pan |
| host family posing for a photo |
| tourist accommodation near Lake Titicaca |
| destinations in Venezuela |
| people observing football match |

The ImageCLEFphoto2007 data collection is considered to be very difficult for retrieval systems because of the abstract semantic content of many queries. For example, the topic “straight road in the USA” could be difficult for visual features whereas “church with more than two towers” could be hard for textual features. This is indeed the motivation for the incorporation of hybrid models in CBIR.

While the MIRFlickr and BGS data collections possess only a number of categories to which the images belong to, the ImageCLEF dataset comes with realistic ground-truth data. This relevance assessment data was generated from image pools, which were judged for relevance by assessors. Of course, relevance is generally a subjective concept and this is why the relevance models, especially the hybrid relevance models are needed. Hybrid relevance models exploit not only the relevance based on the visual relevance (e.g. a plane on the ground) but also captions required by some topics (i.e. pictures of beaches in northern Peru).

The relevance of the images to the query topics was judged in the following scale: relevant, partially relevant, and not relevant.

### 6.1.2 MIRFlickr25000

The MIRFlickr collection comprises 25000 images from the Flickr website which are freely redistributable for research purposes. For a detailed description of the collection the reader is referred to (Huiskes & Lew 2008). This data collection is freely available and can be downloaded from (Author n.d.b). The topics (categories) that were used for evaluation purposes are listed in Table 6.2.

Table 6.2: Topics in MIRFlickr collection

| **General topics** |
|---|
| sky |
| water |
| people |
| night |
| plant life |
| animals |
| man-build structures |
| sunset |
| indoor |
| transport |

The relevance assessment on this data collection can be lenient - an image is relevant to the query if both share at least one category, or strict - an image from data collection must belong to all the query categories.

### 6.1.3 British Geological Survey Data, BGS

The last data collection consists of 7432 images from the British Geological Survey. All the categories are listed in Table 6.3. Some of the categories are obviously very difficult for the visual features to capture. This is due to the highly abstract categories with semantic meanings like "economic geology" or "geological hazards". Thus, the overall performance of the visual features on this collection is rather low.

The relevance assessment on this data collection can be lenient - an image is relevant to the query if both share at least one category, or strict - an image from data collection must belong to all the query categories.

### 6.1.4 ImageCLEF2010 Wikipedia

The ImageCLEF2010 Wikipedia collection comprises 237434 images with unstructured and noisy textual annotations.

An objective of the ImageCLEF2010 Wikipedia Task was to find as many as possible images relevant to the given query describing a user's information need. This data collection is freely available to the participants of the ImageCLEF competition.

Table 6.3: Topics in the BGS data collection

| **Topics** |
|---|
| archeology and early history |
| economic geology |
| fossils |
| general views |
| geological hazards |
| igneous features |
| landforms, erosion |
| landforms, glaciation |
| landforms, karst |
| landforms, lakes |
| landforms, marine |
| landforms, mountains and hills |
| landforms, river |
| landforms, weathering |
| landforms, wind |
| metamorphic features |
| minerals |
| named locality |
| rocks, igneous |
| rocks, metamorphic |
| rocks, sedimentary |
| sedimentary features |
| stratigraphical |
| structural features |
| structural geology |

## 6.2 Novel Visual Features for Generic Image Retrieval

Here, we evaluate our novel visual features designed for generic image retrieval (Chapter 3.2). We also compare the performance of global and local features and the performance of different sampling techniques.

### 6.2.1 Experimental Setup

The global features implemented for comparison purposes are: colour histogram in HSV (Hue, Saturation and Value) colour space, colour histogram in RGB space, texture features based on co-occurrence matrix, colour moments, colour correlogram and methods based on edge detectors.

Our colour histogram was computed as three independent colour distributions. The image colours were discretized and counted according to how many colours belonged to each bin. The number of bins used was 64. Because histograms do not take the spatial relations into account, two images perceived by humans as different may have similar representations.

Colour space may play an important role in visual information retrieval. Hue and saturation components for instance can help to retain light independent colour properties.

The texture features are based on the co-occurrence matrix. Another version of this method incorporates colour by computing the matrices for individual colour channels in HSV colour space.

A colour correlogram is a simple colour feature that takes the spatial information into account. An image is first divided into six sub-images and then the three colour moments are extracted from each sub-image.

The methods based on edge detectors were developed during the course of our research. They incorporate a Canny edge detector and our novel detector based on bilateral filtering (image smoothing that preserves the edges), four directional derivatives and thresholding. The distribution of edges is captured by co-occurrence matrices.

As for the local features, we experimented with the following sampling techniques: pure corner-based, random, dense and hybrid. Most existing approaches set the number of keypoints to be between 300 and 1400. The larger number of sample points is expected to give more accurate results, but the trade-off is the higher computational cost. In our experiments, we tested two different settings: 250 and 900 respectively. For each point of an image, a $10 \times 10$ square patch around it was characterized as multidimensional vector by applying a local descriptor. In the case of dense sampling an image is divided into the $15 \times 15$ (250 sample points) and $30 \times 30$ (900 sample points) sub-images, and the local patches' dimensions are $(x/15, y/15)$ and $(x/30, y/30)$ accordingly, where $x$ denotes width and y height of an image. We also tested another technique based on fixed extracted points, which is similar to dense sampling but the size of a single patch is $10 \times 10$. The influence of the spatial information is taken into account by dividing an image into 9 sub-images of equal size. Each image patch has 9 and 12 dimensions accordingly, and the codebook size is 40.

For the ImageCLEF data collection, for each query topic (60 topics provided) we retrieve 1000 images and calculate Mean Average Precision (MAP).

For the MIRFlickr, we retrieve 1000 images for each query, which were selected randomly from the collection. We select 100 query images and compute MAP based on the ground truth data and general topics. Some images belong to a few categories. The evaluation approach was the "lenient" one. We assumed that an image is relevant if it shares at least one category with the query image.

For BGS, 100 query images were randomly selected, with 1000 images retrieved for each individual query. The Mean Average Precision computed for the local features was compared with MAPs obtained by applying global methods. We also evaluate the different sampling techniques on this data collection. Again, some images in the BGS collection belong to a few categories. The evaluation approach was the "lenient" one. We assumed that an image is relevant if it shares at least one category with the query image.

### 6.2.2 Results and Discussion

Tables 6.4 and 6.5 illustrate the performance of our local features on the ImageCLEF2007 data collection, for 250 and 900 sample points accordingly.

Table 6.4: ImageCLEF2007 results, 250 sample points

| | **Hybrid** | **Random** | **Dense** | **Corner** |
|---|---|---|---|---|
| **MAP** | 0.042 | 0.041 | 0.037 | 0.031 |

Table 6.5: ImageCLEF2007 results, 900 sample points

| | **Hybrid** | **Random** | **Dense** | **Corner** |
|---|---|---|---|---|
| **MAP** | 0.050 | 0.051 | 0.048 | 0.040 |

In terms of the MAP, for the relatively small number of sample points, the best performing method is the one with hybrid sampling. When the number of sample points increases, however, the discriminative power of randomly sampled patches grows and the set of keypoints detected by corner detector gets saturated. In other words, the corner detector is unable to find more discriminative patches. That is why the best performing method, for 900 sample points, is the one incorporating random sampling. We can also observe that it is desirable to have a large number of sample points which improves the overall performance of local features. Nevertheless, the larger the number of local patches is, the higher the computational and the storage costs will be.

We have also tested the influence of the spatial information on the retrieval performance. For a small number of sample points, the inclusion of spatial information improves the retrieval performance. Thus, the MAP for the local features with random sampling and 250 patches increases to MAP$\approx$0.0427. However, when the number of sample points is high, e.g. 900, the spatial information can even hamper the retrieval. For instance, the retrieval performance of local features with random sampling, spatial information included and 900 patches, drops to MAP$\approx$0.045.

The influence of the number of sample points and the inclusion of spatial information on retrieval performance validates the findings reported in (Yang et al. 2007). It turns out that when it comes to content based retrieval of generic real-life images, the sophisticated sampling methods are effective only if the number of sample points is relatively low. Otherwise, simple random sampling is the best one due to its ability to find a higher number of discriminative image patches.

The local features with a different descriptor based on the co-occurrence matrix (with hybrid sampling and 250 sample points) performed worse, obtaining MAP$\approx$0.011.

According to (Overell, Llorente, Liu, Hu, Rae, Zhu, Song & Rüger 2008), the best performing global method obtained MAP$\approx$0.026. They conducted their experiments on ImageCLEF 2008 collection which consists of the same 20000 images and same 60 topics as ImageCLEF 2007. Thus, in this case, the MAP of the local features is almost twice as high as MAP of the global methods.

The performance of local features with different sampling techniques on MIRFlickr dataset

is quite similar (Table 6.6). This is rather surprising, considering the differences in performance on individual queries. This suggests that the combination of the features with different sampling should improve the results. This will be investigated further in the future.

Table 6.6: MIRFlickr results, 250 sample points

| | **Hybrid** | **Random** | **Dense** | **Corner** |
|---|---|---|---|---|
| **MAP** | 0.610 | 0.607 | 0.609 | 0.592 |

Figure 6.1 presents the retrieval results for the global features. The MAP of local features with hybrid sampling is also depicted for comparison purposes. We can observe that the local features performed better on this image collection with our hybrid sampling approach (local hybrid) being the best in terms of MAP.

The local features with a different descriptor based on the co-occurrence matrix (with hybrid sampling) performed worse, obtaining MAP≈0.599.

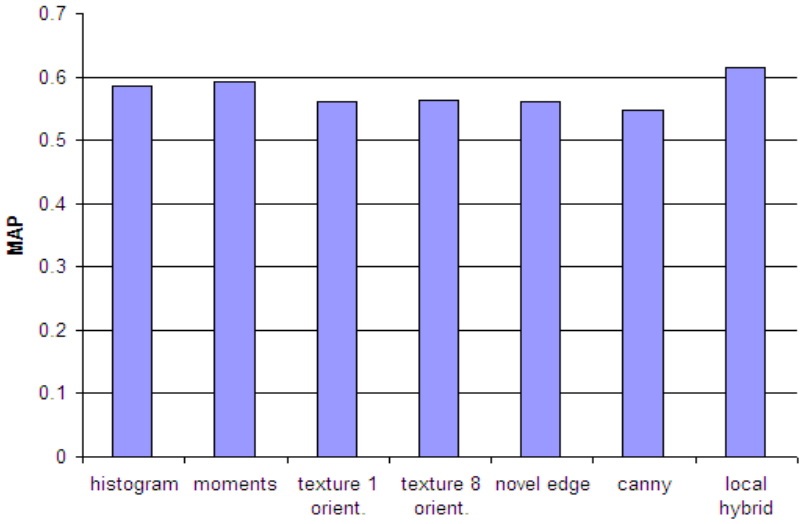


Figure 6.1: Global versus local features, MIRFlickr.

The best results for the BGS data collection are obtained by the local features with hybrid sampling as shown in Table 6.7. The difference between this best performing method and others (approximately 17% of improvement over the most commonly used corner/blob detector-based sampling) is more significant than the difference on the ImageCLEF and MIRFlickr collections.

Table 6.7: BGS - results, 250 sample points

| | **Hybrid** | **Random** | **Dense** | **Corner** |
|---|---|---|---|---|
| **MAP** | 0.183 | 0.171 | 0.178 | 0.156 |

Figure 6.2 depicts the MAP for global features and the best performing local method (hybrid sampling). Again, the local features outperformed the global ones.

The local features with a different descriptor based on the co-occurrence matrix (with hybrid sampling) performed worse, obtaining MAP≈0.159.

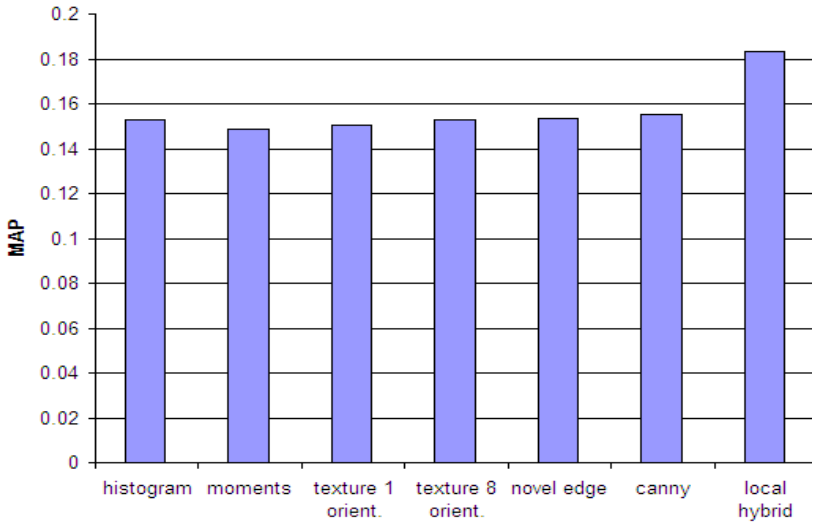


Figure 6.2: Global versus local features, BGS.

The novel global methods based on the edge detectors performed comparably with other standard, widely used global methods. Moreover, both Canny and our novel edge detector performed similarly in this context.

The average scores across all system runs of purely visual retrieval at ImageCLEF2007 were 0.068 in terms of MAP. The best performing visual systems combined various visual features. Moreover, most of the groups participating in the retrieval task exploited high dimensional, computationally expensive visual features. Hence, our local features, despite their low dimensionality prove to be as effective as more sophisticated methods in the generic image retrieval task.

**RGU at ImageCLEF2010 Wikipedia Retrieval Task**

In order to test our local features on a large-scale dataset, we have participated in the ImageCLEF2010 Wikipedia Retrieval Task.

The participants of the Wikipedia Task were required to submit ranked lists of the top 1000 images ranked in descending order of similarity. Then, the MAP (major criteria) and a few other evaluation metrics were calculated for each run.

The performance of our local features was 0.0069 in terms of MAP. The visual features provided by the organizer, despite being high dimensional and more sophisticated, performed worse obtaining MAP of only 0.0003. Out of the seven purely visual runs, our features were ranked in the middle.

### 6.2.3 Conclusions and Future Work

Real-life images often consist of different patches of uniform pattern, making it hard for the global description of images to capture all the properties. That is why the the local features approach based on the "bag of visual words" has recently gathered much attention. We propose a novel method based on the local features, incorporating an easy to implement descriptor and a hybrid sampling technique. We also compare different sampling methods: hybrid (a combination of random and detector-based sampling), purely random, purely detector-based and dense on three large data collections. The hybrid sampling produces more discriminative image patches than commonly used detector-based method. The proposed descriptor is easy to implement and produces low dimensional feature vectors which, in turn, reduces the computational and data storage cost. Empirical evaluation has been done on three large image collections, namely ImageCLEF 2007, MIRFlickr 25000 and BGS datasets.

The evaluation results show that for a relatively small number of sample points the hybrid sampling is superior than other methods. This is because the points detected by corner/blob detectors tend to concentrate on objects, which is good for object instance recognition but not necessarily good for generic image retrieval. The background also plays an important role in the retrieval of real-life generic images. When it comes to the large number of sample points the random sampling technique performs best which is due to its high discriminative power. The random sampling is always able to find discriminative image patches while detector-based set of keypoints gets at some point saturated. Moreover, the higher number of sample points usually results in a higher Mean Average Precision (MAP). However, the increase in the number of sample points produces higher storage and computational requirements so there is always some kind of a trade-off. In order to further improve the performance of local features, we can also take the spatial information into account. This works well for the relatively small number of sample points but can hamper the overall performance if the number of points is too high.

The comparison of local features against global methods shows that the approach based on "bag of visual words" performs better in terms of Mean Average Precision. This is hardly surprising, considering the discriminative properties of local features for generic real-life image retrieval. However, in our opinion this is domain specific and global features might be a better choice when it comes to the different retrieval problems. The evaluation performed on all three collections shows that the results obtained by our local features are comparable with the current state-of-the-art in CBIR.

Our approach is easy to implement, not sophisticated, with low computational and data storage cost (mostly because our vectors are low dimensional), and the hybrid sampling technique can be used in other methods based on the "bag of visual words" to improve the retrieval performance. Moreover, the evaluation of the proposed method has been conducted on three different large data collections without changing the initial setup. In this way we avoided "fine-tuning" of the parameters to the specific data collection which makes the results more general and reliable.

Purely content-based image information retrieval systems are still unable to properly imitate

human visual perception. This is because of the semantic gap, the difference between the human perception and computer representation of the multimedia. Therefore, the focus should be on hybrid approaches incorporating user relevance feedback, social information like tags, ratings and comments, user profiling (e.g. to simulate individual user's perception by changing the similarity measure), semantics and textual information. However, it is important to try to improve individual components like visual features.

We are planning to extend our evaluation of different sampling techniques to various other detectors and descriptors like Scale Invariant Feature Transform for example. We will extend the experiments to a different number of sample points. The influence of the size of the codebook on the retrieval performance will also be measured. We believe that the optimal number of "visual words" may depend on the sort of detector and descriptor used and their representative ability. To the best of our knowledge the existing evaluation assumes that the size of the codebook is independent of the descriptors and detectors or test the influence on one specific method. This is an open issue that remains to be tested experimentally.

## 6.3 Enhancing Local Features by Exploiting the Correlation Between Visual Words

Recall that the standard Bag of Visual Words framework treats the visual words as independent. In fact, the bins in the histograms of visual words counts are often correlated. We utilize this correlation to adapt the similarity measure and improve the standard local features model (Chapter 3.3). We also introduce and test various notions of correlation within this framework.

### 6.3.1 Experimental Setup

For each of the 100 query topics (60 for ImageCLEF) we retrieve 16 images and calculate Mean Average Precision (MAP). To test the influence of the correlations on the retrieval performance, we generate the correlation matrix from these 16 images, weight the similarity measure, and re-rank the top images. Next, we compute the Mean Average Precision and compare it with the baseline (which does not take the correlations into account). The assumption that the top retrieved images are relevant to the query is characteristic of the so-called Pseudo Relevance Feedback (PRF).

Some images belong to a few categories. The evaluation on MIRFlickr and BGS collections was the "lenient" one. We assumed that an image is relevant if it shares at least one category with the query image (based on the ground truth data provided).

The implemented local features utilize the random sampling technique. We set the number of sample points to 900. A large number of sample points (in random sampling) is expected to give better results than other sampling methods (see (Nowak et al. 2006a)). For each sample point of an image, a $10 \times 10$ square patch around it was characterized as multidimensional vector by applying a local descriptor. Each image patch has 9 dimensions (3 for each colour channel), and the codebook

size is 40. The visual features, despite using low dimensional vectors and small vocabulary, are comparable with more sophisticated approaches (ImageCLEF2010 Wikipedia Retrieval Task).

When exploiting the correlations between visual words, we identify 5 most and 1 least correlated pair. The $p$ and $c$ parameters' values in the similarity measure are set to 0.5 and 1.31 for all three data collections and were determined experimentally. In the case of proximity-based correlation, we will consider two instances of visual words to be correlated if they both appear within a circle of radius 14.15. This is approximately the sum of two diagonals of square sub-images (image patches may overlap).

### 6.3.2 Results and Discussion

Tables 6.8, 6.9, and 6.3.2 show the experimental results. They present the results for the case when no Pseudo Relevance Feedback was incorporated (NP), when only the dominant correlations were taken into account (D), and the MAPs for both dominant and least correlated pairs (DL). The performance of five notions of correlation is also depicted in the tables. Labels C1, C2, C3 and C4 correspond to correlations 1, 2, 3 and 4 accordingly, whereas C0 denotes proximity-based correlation (subsection 3.3).

The computation of correlation 1 and then addition of matrices is equivalent to the commonly used multiplication of the transpose of an image representation matrix by itself. It is one of the standard ways for capturing correlation. Therefore, correlation 1 can also be considered as another baseline. Results presented in bold font are significantly different (two-tailed t-test, 0.05) from the baseline.

Table 6.8: ImageCLEF2007 results (MAP)

| | **C4** | **C3** | **C2** | **C1** | **C0** |
|---|---|---|---|---|---|
| **NP** | 0.0204 | 0.0204 | 0.0204 | 0.0204 | 0.0204 |
| **D** | 0.0211 | 0.0211 | 0.0210 | 0.0208 | 0.0206 |
| **DL** | 0.0213 | 0.0213 | 0.0211 | 0.0209 | 0.0207 |

Table 6.9: MIRFlickr results (MAP)

| | **C4** | **C3** | **C2** | **C1** | **C0** |
|---|---|---|---|---|---|
| **NP** | 0.6794 | 0.6794 | 0.6794 | 0.6794 | 0.6794 |
| **D** | **0.6938** | **0.6936** | 0.6859 | 0.6869 | 0.6802 |
| **DL** | **0.6951** | **0.6936** | 0.6854 | 0.6871 | 0.6807 |

Table 6.10: BGS results (MAP)

| | **C4** | **C3** | **C2** | **C1** | **C0** |
|---|---|---|---|---|---|
| **NP** | 0.3158 | 0.3158 | 0.3158 | 0.3158 | 0.3158 |
| **D** | **0.3286** | **0.3286** | 0.3187 | 0.3199 | 0.3172 |
| **DL** | 0.3268 | 0.3265 | 0.3194 | 0.3193 | 0.3176 |

It can be seen that the C4 and C3 correlations obtained the best results on all three data collections. We argue that this is due to some contradictions with our intuition of correlation.

Let us focus on Correlation 1. If the frequencies of two pairs of visual words are $\{5, 10\}$ and $\{5, 100\}$ then the latter will be assigned a higher correlation value. We would, however, expect the former pair to be at least equally correlated.

Normalization (Correlation 2) helps to overcome the above issue. However, if the frequencies are proportional, for example $\{10, 20\}$ and $\{40, 80\}$ then the former will score higher. But, intuitively, the latter is more correlated.

Correlation 3 seems to be intuitively right, but will ignore the additional information from the frequencies (see example for Correlation 1). Normalization of correlation 3 will produce similar side effects to Correlation 2. Therefore, we introduced the Correlation 4, which does not seem to contradict our intuition.

The addition of information about the least correlated visual words often further improves the performance. Moreover, image level correlations outperformed proximity based one. This may be due to the notion that an image may contain correlated visual words not because of their proximity but because they refer to the same topic.

We should be aware, however, that the assumption in the PRF framework that all the top documents are relevant to the query may produce a number of false correlations. The process will therefore depend on the adequacy (the ability to capture relevant properties) of the image representation and the retrieval performance of the implemented methods.

### 6.3.3 Conclusions and Future Work

We propose a new approach for identifying and utilizing the information about correlations between visual words. We implement and test various notions of correlation at different contextual levels (we refer to them as image-level and proximity based). To the best of our knowledge, this is the first time these two were compared within this type of framework in image retrieval. Our local features consist of low dimensional histograms, where bins representing visual words are highly correlated. We identify the most and the least correlated coefficients and use the thus obtained information, along with the visual terms' frequencies from the current query, to weight the similarity measure. Certain coefficients in the similarity measure corresponding to the most correlated terms are then increased, while the coefficients related to the least correlated pairs are deemphasized. The evaluation was performed on three large data collections, namely ImageCLEF 2007, MIRFlickr 25000 and BGS. The evaluation was performed within the Pseudo Relevance Feedback framework.

Experimental results show the superiority of two notions of correlation, C4 and C3, which are image level correlations. For these two correlations, we report significant improvement in terms of Mean Average Precision on two data collections within PRF evaluation framework. Moreover, the addition of information about the least correlated visual words often further improves the performance. The proximity based notion of correlation does not show a significant improvement

in the context of this model.

The proposed method is computationally and data storage cheap, utilizes correlation at different contextual levels, and avoids the normalization of histograms.

We are planning to extend our evaluation to other various weighting schemes and similarity measures.

## 6.4 Combining Systems. Tensor Product of Correlated Text and Visual Features

Here, we present the evaluation results of our hybrid model for the combination of visual and textual features (Chapter 4.1). Instead of heuristically combining the ranking scores independently of different media types, we develop a tensor product based model to represent textual and visual content features of an image as a non-separable composite system.

### 6.4.1 Experimental Setup

The initial image retrieval experiments based on a tensor space were performed on ImageClef2007, a widely used benchmarking collection for image retrieval.

When creating a density matrix for each document, we set the probability of each term to the normalized TF-IDF. The visual feature that we are using in the initial experiment is the colour histogram in HSV colour space, which is the cylindrical representation of the RGB colour space. Hue and saturation components help to retain light independent colour properties. The HSV colour histogram of an image is computed as three independent distributions. First, we split an image into individual colour channels (grayscale representations of primary colours). Next, we discretize the colours and count how many pixels belong to each colour bin.

The large-scale evaluation of the hybrid model (RGU at ImageCLEF2010 Wikipedia Retrieval Task) incorporates our novel local features.

Notice, that the feature space can be replaced by any other visual feature. The tensor space can also be expanded to incorporate more visual features.

Clearly, all the dimensions of the HSV colour space are pair-wise orthogonal, as the colours do not overlap. However, this is not the case in textual feature space, where the dimensions correspond to textual words. Some words may have the same or similar meaning. This can be captured by replacing the synonyms with one unique term, or utilizing LSI to find its latent semantic space. As a first step, we simply assume that the words are orthogonal, and focus on the tensor model itself.

In our experiments, we compare the proposed tensor-based model with methods based on the pure visual and pure textural features only. We also compare them with simple concatenation of textual and visual feature vectors. Because of the slight adjustment of each model during the experiments, we list all the runs

- cbF: pure visual feature based method using the city block distance measure (following the recommendation from a systematic study on distance measurements in (Liu et al. 2008)).
- cosT: pure text-based method using cosine similarity
- cos T+F: cosine similarity based on the concatenation of textual and visual features
- cosT(e)+F: cosine similarity based on the concatenation of textual and visual features (first, each image will be annotated.)
- tensor(T+F): quantum-like measurement in tensor space
- tensorT(e)+F: quantum-like measurement in tensor space (a correlation density matrix will be included into each image's density representation.)

Suppose we have a document and a query, each of them represented by feature vectors: $d^t = (t_1, t_2, ...t_n)$ and $d^v = (f_1, f_2, ...f_m)$. $m$ and $n$ are visual and textual feature dimensions, respectively. Then, the retrieval functions utilized in our experiment are given by

1. City block (for cbF)
$$s(d,q) = \sum_{i=1}^{m} |f_i^d - f_i^q| \tag{6.1}$$
2. Cosine similarity

   For cosT:
$$s(d,q) = \sum_{i=1}^{n} t_i^d \cdot t_i^q \tag{6.2}$$

   For cosT+F and cosT(e)+F:
$$s(d,q) = \sum_{i=1}^{n} t_i^d \cdot t_i^q + \sum_{i=1}^{m} f_i^d \cdot f_i^q \tag{6.3}$$

   Our cosine similarity measurement is an approximate cosine similarity, as it can be observed that the similarity score in Equation 6.3 is not divided by vector length. We report the result of this model rather than the standard cosine similarity for two reasons: the feature values have been normalized within their own feature space; and our experimental results show that the approximate cosine similarity has better performance than the standard one.
3. Measurement in the tensor space.

   Based on the quantum measurement, we score a document according to the observable's expectation on the document. With orthogonal assumption of textual basis $|t_i\rangle$ and visual feature basis $|f_i\rangle$, we have:

$$\begin{aligned} s(d,q) &= tr(\sum_i (t_i^d \cdot f_j^d)^2 |t_i f_j\rangle\langle t_i f_j| \\ &\cdot \quad (t_i^q \cdot f_j^q)^2 |t_i f_j\rangle\langle t_i f_j|) \\ &= trace(\rho_d \cdot \rho_q) \\ &= \sum_{ij} (t_i^d \cdot f_j^d)^2 (t_i^q \cdot f_j^q)^2 \end{aligned} \tag{6.4}$$

This shows the same result of transition probability, which is explained as the probability that a system in state $d$ will be found in state $q$ (Aharonov et al. 1981), and it is computed as $P(q|d) = |\langle q|d\rangle|^2$. When this classical quantum view is applied to the retrieval model, $|\langle q|d\rangle|^2$ can be explained as the probability that a document can be observed containing the information described by the query.

Let us take the superposed document and a query as an example:

$$|d\rangle = \sum_{ij} \gamma_{ij}^d |t_i f_j\rangle,\ |q\rangle = \sum_{ij} \gamma_{ij}^q |t_i f_j\rangle \tag{6.5}$$

Then, the transition probability between them is:

$$\begin{aligned} s(d,q) &= P(d \to q) \\ &= |\langle d|q\rangle|^2 \\ &= \sum_{i,j} (\gamma_{ij}^d)^2 (\gamma_{ij}^q)^2 \end{aligned} \tag{6.6}$$

In such a case, the measurement on the document density matrix is the same as the inner product of two states, which equals to the cosine similarity of two flattened tensors, where the document and query are represented in a tensor form. This is also our current experimental setting.

We use two widely adopted IR performance measures: Average Precision (AP) and Precision at top 10 retrieved documents (P@10). Precision measures the percentage of relevant documents in the whole returned document list. However, the ability to return the relevant documents with higher ranks is also a desirable performance for a retrieval system. Average precision measures both, it is the average of the precisions computed at the point of each of the relevant documents in the ranked list:

$$\text{AvgPrecision} = \frac{\sum_{r=1}^{N} (Precision(r) \times \text{relevant}(r))}{\text{number of relevant documents}} \tag{6.7}$$

where r is the rank, N is the number of retrieved documents, relevant(r) denotes a binary function on the relevance of a document with rank $r$, and Precision(r) denotes precision at a given

cut-off rank $r$:

$$\text{Precision(r)} = \frac{|\{\text{relevant retrieved document of rank r or less}\}|}{r} \tag{6.8}$$

Note that the denominator in Equation 6.7 is the number of relevant documents in the entire collection, hence the average precision reflects the performance over all relevant documents, regardless of the retrieval cut-off.

In this small-scale evaluation, we focus on testing the effectiveness of the model, and therefore do not include additional efficiency measures. Nonetheless, the tensor model is obviously more computationally expensive than standard approaches. [1]

### 6.4.2 Results and Discussion

The list of evaluation results for each run (Table 6.11, Table 6.12) shows that the pure visual content based retrieval has the lowest performance. Pure text retrieval on images is significantly better than content based visual retrieval. However, even the simple concatenation of text and visual features can improve the retrieval.

The tensor of visual and textual features can capture certain relationships between them. Even the pure tensor product without taking into account the correlation between text and visual features, can improve mean AP by 17% compared to cosT+F, and 34 individual queries have better retrieval results. Still, for some queries, their APs drop compared with the cosine similarity on the features concatenation.

If text or content features alone cannot retrieve any relevant images, the pure tensor product cannot retrieve any relevant images either. This can be observed on queries 06, 24, 30, 41, 49, and 56. This is not a surprise, since if a document does not project to the space spanned by $|t_i\rangle$, it will not project to the space spanned by $|t_i f_j\rangle$ either. Therefore, the pure tensor product will not solve the problem that the images without proper annotation will be ranked low. For example, even if an image has very distinctive visual features related to the concept "house" but the word "house" does not appear in its text description, the image's ranking score will be low with respect to a query whose text contains the term "house". However, the same image can be ranked high through the correlation of its visual features with text "house". This is the reason why identifying and exploiting the correlation between visual and textual features is an important aspect of image retrieval.

Unfortunately, our two simple methods for correlating visual and textual features did not bring much improvement to the retrieval results. We observed that some words associated with the feature dimensions do not match any query terms. As a result, the ranking scores of the images would not change at all. This is due to several reasons. For example, the annotation of some images do not account for the content of the image. e.g., "the destination of the tourist". The

[1] This problem can be solved by our observations from the previous chapters, namely, that the cosine similarity of the tensored vectors can be represented as a late fusion.

| | cos T+F | | cos T | | cb F | | cos T(e)+F | | tensor T+F | | tensor T(e)+F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Qid | AP | P@10 | AP | P@10 | AP | P@10 | AP | P@10 | AP | P@10 | AP | P@10 |
| 01 | 0.1395 | 0.4000 | 0.0811 | 0.4000 | 0.0070 | 0.0000 | 0.1161 | 0.4000 | 0.0906 | 0.6000 | 0.0890 | 0.6000 |
| 02 | 0.0038 | 0.0000 | 0.0050 | 0.0000 | 0.0229 | 0.2000 | 0.0094 | 0.0000 | 0.0134 | 0.0000 | 0.0065 | 0.0000 |
| 03 | 0.0576 | 0.0000 | 0.1646 | 0.0000 | 0.0003 | 0.0000 | 0.0798 | 0.2000 | 0.1993 | 0.0000 | 0.1656 | 0.0000 |
| 04 | 0.0187 | 0.2000 | 0.0108 | 0.0000 | 0.0022 | 0.0000 | 0.0168 | 0.0000 | 0.0100 | 0.0000 | 0.0144 | 0.2000 |
| 05 | 0.0065 | 0.2000 | 0.0040 | 0.0000 | 0.0024 | 0.0000 | 0.0160 | 0.2000 | 0.0208 | 0.2000 | 0.0180 | 0.2000 |
| 06 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0005 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 07 | 0.4916 | 1.0000 | 0.4141 | 0.8000 | 0.0025 | 0.0000 | 0.4998 | 0.8000 | 0.4369 | 0.8000 | 0.4338 | 0.8000 |
| 08 | 0.5173 | 0.8000 | 0.5651 | 1.0000 | 0.0059 | 0.2000 | 0.4934 | 0.8000 | 0.3929 | 0.8000 | 0.3913 | 0.6000 |
| 09 | 0.0002 | 0.0000 | 0.0002 | 0.0000 | 0.0006 | 0.0000 | 0.0005 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 10 | 0.2374 | 0.8000 | 0.5474 | 0.6000 | 0.0012 | 0.0000 | 0.2826 | 0.8000 | 0.5829 | 0.0000 | 0.5816 | 0.0000 |
| 11 | 0.8284 | 1.0000 | 0.0685 | 0.0000 | 0.7115 | 1.0000 | 0.9135 | 1.0000 | 0.7548 | 0.4000 | 0.7523 | 0.4000 |
| 12 | 0.0110 | 0.0000 | 0.0185 | 0.0000 | 0.0000 | 0.0000 | 0.0254 | 0.2000 | 0.0449 | 0.2000 | 0.0459 | 0.2000 |
| 13 | 0.1137 | 1.0000 | 0.0998 | 1.0000 | 0.0001 | 0.0000 | 0.0964 | 0.8000 | 0.0753 | 0.8000 | 0.0758 | 0.8000 |
| 14 | 0.0032 | 0.0000 | 0.0055 | 0.0000 | 0.0588 | 0.2000 | 0.0236 | 0.0000 | 0.0042 | 0.0000 | 0.0043 | 0.0000 |
| 15 | 0.3720 | 0.6000 | 0.5572 | 1.0000 | 0.0061 | 0.0000 | 0.2680 | 0.0000 | 0.2313 | 0.0000 | 0.2294 | 0.0000 |
| 16 | 0.1385 | 0.0000 | 0.1367 | 0.0000 | 0.0004 | 0.0000 | 0.1409 | 0.0000 | 0.2280 | 0.6000 | 0.1926 | 0.2000 |
| 17 | 0.1611 | 0.6000 | 0.1563 | 0.2000 | 0.0031 | 0.0000 | 0.1571 | 0.2000 | 0.2674 | 0.2000 | 0.2693 | 0.2000 |
| 18 | 0.2423 | 0.6000 | 0.2378 | 0.4000 | 0.0048 | 0.0000 | 0.2844 | 1.0000 | 0.2962 | 1.0000 | 0.2950 | 1.0000 |
| 19 | 0.0198 | 0.2000 | 0.0096 | 0.2000 | 0.0139 | 0.2000 | 0.0318 | 0.2000 | 0.0463 | 0.2000 | 0.0459 | 0.2000 |
| 20 | 0.0055 | 0.0000 | 0.0098 | 0.0000 | 0.0000 | 0.0000 | 0.0096 | 0.0000 | 0.0341 | 0.4000 | 0.0241 | 0.2000 |
| 21 | 0.2822 | 0.2000 | 0.2669 | 0.0000 | 0.0153 | 0.2000 | 0.2892 | 0.0000 | 0.4504 | 0.6000 | 0.3319 | 0.0000 |
| 22 | 0.0004 | 0.0000 | 0.0010 | 0.0000 | 0.2804 | 1.0000 | 0.0319 | 0.0000 | 0.0143 | 0.2000 | 0.0114 | 0.2000 |
| 23 | 0.0212 | 0.2000 | 0.0267 | 0.0000 | 0.0015 | 0.0000 | 0.0891 | 0.6000 | 0.1272 | 0.0000 | 0.1258 | 0.0000 |
| 24 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0015 | 0.0000 | 0.0080 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25 | 0.0012 | 0.0000 | 0.0016 | 0.0000 | 0.0019 | 0.0000 | 0.0015 | 0.0000 | 0.0025 | 0.0000 | 0.0025 | 0.0000 |
| 26 | 0.0000 | 0.0000 | 0.0530 | 0.0000 | 0.0019 | 0.0000 | 0.0000 | 0.0000 | 0.0530 | 0.0000 | 0.0530 | 0.0000 |
| 27 | 0.5682 | 0.8000 | 0.3932 | 0.0000 | 0.1154 | 0.6000 | 0.6863 | 1.0000 | 0.6657 | 0.8000 | 0.6618 | 0.8000 |
| 28 | 0.1008 | 0.4000 | 0.1119 | 0.6000 | 0.0087 | 0.0000 | 0.0916 | 0.4000 | 0.1796 | 0.8000 | 0.2029 | 1.0000 |
| 29 | 0.0975 | 0.0000 | 0.1557 | 0.0000 | 0.0074 | 0.0000 | 0.1139 | 0.0000 | 0.1320 | 0.0000 | 0.1283 | 0.0000 |
| 30 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| mean | 0.1440 | 0.3167 | 0.1313 | 0.2300 | 0.0354 | 0.1267 | 0.1596 | 0.3133 | **0.1685** | 0.3067 | 0.1638 | 0.2867 |

Table 6.11: AP and P@10 for first 30 queries

| | cos T+F | | cos T | | cb F | | cos T(e)+F | | tensor T+F | | tensor T(e)+F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Qid | AP | P@10 | AP | P@10 | AP | P@10 | AP | P@10 | AP | P@10 | AP | P@10 |
| 31 | 0.0605 | 0.8000 | 0.0427 | 0.2000 | 0.0018 | 0.0000 | 0.0696 | 0.4000 | 0.1076 | 0.8000 | 0.1071 | 0.8000 |
| 32 | 0.2017 | 0.4000 | 0.2779 | 0.6000 | 0.0003 | 0.0000 | 0.2204 | 0.6000 | 0.2922 | 0.4000 | 0.2926 | 0.4000 |
| 33 | 0.0290 | 0.0000 | 0.0001 | 0.0000 | 0.0475 | 0.0000 | 0.0909 | 0.0000 | 0.0001 | 0.0000 | 0.0001 | 0.0000 |
| 34 | 0.0810 | 0.2000 | 0.0865 | 0.2000 | 0.0019 | 0.0000 | 0.0785 | 0.2000 | 0.0542 | 0.2000 | 0.0537 | 0.2000 |
| 35 | 0.1765 | 1.0000 | 0.2202 | 1.0000 | 0.0329 | 0.2000 | 0.2133 | 1.0000 | 0.2500 | 1.0000 | 0.2495 | 1.0000 |
| 36 | 0.5584 | 0.8000 | 0.5392 | 0.8000 | 0.0171 | 0.2000 | 0.5790 | 0.8000 | 0.5766 | 0.8000 | 0.5769 | 0.8000 |
| 37 | 0.1213 | 0.4000 | 0.0957 | 0.2000 | 0.0490 | 0.6000 | 0.1415 | 0.4000 | 0.1256 | 0.4000 | 0.1090 | 0.4000 |
| 38 | 0.1698 | 0.4000 | 0.1058 | 0.0000 | 0.0382 | 0.0000 | 0.1976 | 0.2000 | 0.2769 | 0.6000 | 0.2780 | 0.6000 |
| 39 | 0.0015 | 0.0000 | 0.0007 | 0.0000 | 0.0012 | 0.0000 | 0.0008 | 0.0000 | 0.0008 | 0.0000 | 0.0008 | 0.0000 |
| 40 | 0.0049 | 0.2000 | 0.0012 | 0.0000 | 0.0027 | 0.0000 | 0.0031 | 0.0000 | 0.0010 | 0.0000 | 0.0010 | 0.0000 |
| 41 | 0.0003 | 0.0000 | 0.0002 | 0.0000 | 0.0003 | 0.0000 | 0.0006 | 0.0000 | 0.0004 | 0.0000 | 0.0004 | 0.0000 |
| 42 | 0.2484 | 0.0000 | 0.2798 | 0.0000 | 0.0016 | 0.0000 | 0.2931 | 0.0000 | 0.3256 | 0.0000 | 0.3231 | 0.0000 |
| 43 | 0.2097 | 0.4000 | 0.1767 | 0.4000 | 0.0231 | 0.2000 | 0.2163 | 0.4000 | 0.1583 | 0.4000 | 0.1479 | 0.4000 |
| 44 | 0.0429 | 0.4000 | 0.0606 | 0.4000 | 0.0044 | 0.2000 | 0.0643 | 0.8000 | 0.0634 | 0.2000 | 0.0636 | 0.2000 |
| 45 | 0.0491 | 0.4000 | 0.0243 | 0.4000 | 0.0377 | 0.2000 | 0.0511 | 0.4000 | 0.0698 | 0.6000 | 0.0698 | 0.6000 |
| 46 | 0.0021 | 0.0000 | 0.0019 | 0.0000 | 0.0075 | 0.0000 | 0.0031 | 0.0000 | 0.0109 | 0.0000 | 0.0066 | 0.0000 |
| 47 | 0.0143 | 0.2000 | 0.0048 | 0.0000 | 0.0028 | 0.0000 | 0.0286 | 0.2000 | 0.0143 | 0.2000 | 0.0143 | 0.2000 |
| 48 | 0.1106 | 0.0000 | 0.0931 | 0.0000 | 0.0385 | 0.2000 | 0.1252 | 0.0000 | 0.1805 | 0.4000 | 0.1913 | 0.4000 |
| 49 | 0.0004 | 0.0000 | 0.0000 | 0.0000 | 0.0172 | 0.2000 | 0.0012 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 50 | 0.0919 | 0.4000 | 0.0270 | 0.0000 | 0.0004 | 0.0000 | 0.1550 | 0.0000 | 0.1956 | 0.0000 | 0.1917 | 0.0000 |
| 51 | 0.0921 | 0.6000 | 0.0867 | 0.6000 | 0.1321 | 0.6000 | 0.0990 | 0.6000 | 0.0896 | 0.6000 | 0.0896 | 0.6000 |
| 52 | 0.0012 | 0.0000 | 0.0001 | 0.0000 | 0.0003 | 0.0000 | 0.0024 | 0.0000 | 0.0001 | 0.0000 | 0.0001 | 0.0000 |
| 53 | 0.2254 | 0.8000 | 0.1463 | 0.6000 | 0.0049 | 0.0000 | 0.2656 | 1.0000 | 0.2477 | 0.6000 | 0.2452 | 0.6000 |
| 54 | 0.0584 | 0.2000 | 0.0575 | 0.0000 | 0.0243 | 0.2000 | 0.1008 | 0.2000 | 0.1033 | 0.0000 | 0.0925 | 0.0000 |
| 55 | 0.1323 | 0.2000 | 0.0013 | 0.0000 | 0.2829 | 0.8000 | 0.2468 | 0.8000 | 0.0211 | 0.4000 | 0.0212 | 0.4000 |
| 56 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 57 | 0.9751 | 1.0000 | 0.9536 | 1.0000 | 0.0004 | 0.0000 | 0.9255 | 0.8000 | 0.9255 | 0.8000 | 0.9073 | 0.8000 |
| 58 | 0.2491 | 0.8000 | 0.2050 | 0.6000 | 0.0044 | 0.0000 | 0.3272 | 1.0000 | 0.3057 | 0.8000 | 0.3148 | 0.8000 |
| 59 | 0.0513 | 0.0000 | 0.0727 | 0.2000 | 0.0021 | 0.0000 | 0.0858 | 0.0000 | 0.0918 | 0.2000 | 0.0908 | 0.2000 |
| 60 | 0.2394 | 0.4000 | 0.2132 | 0.4000 | 0.0691 | 0.4000 | 0.2136 | 0.4000 | 0.2640 | 0.4000 | 0.2389 | 0.2000 |
| mean | 0.1440 | 0.3167 | 0.1313 | 0.2300 | 0.0354 | 0.1267 | 0.1596 | 0.3133 | **0.1685** | 0.3067 | 0.1638 | 0.2867 |

Table 6.12: AP and P@10 for remaining 30 queries

query text sometimes does not associate with any specific content feature, e.g., "Asian traffic". Moreover, a simple visual feature such as a colour histogram may not be suitable to be associated with semantic meaning (Wang, Hoiem & Forsyth 2009). Extracting correlation information for each query sample can be a solution.

**RGU at ImageCLEF2010 Wikipedia Retrieval Task**

In order to evaluate our models on the large-scale dataset, we have participated in the ImageCLEF2010 Wikipedia Retrieval Task. We have submitted the following runs

- TFL: search by text, then re-rank with our local features
- TFC: search by text, then re-rank with visual features provided by the organizer
- TXFL: quantum-like measurement on tensor product space of annotation vector and our local features vector
- TXFC: quantum-like measurement on tensor product space of annotation vector and visual features vector (provided by organizer)
- combine: TFC based retrieval. When the length of the result from TFC is less than 1000, then the images from content retrieval are appended into the result list
- WFC: retrieve on Wikipedia file first, then re-rank with visual features provided by the organizer
- cleszek: city block distance with our local features (pure visual)
- cadd: city block distance with all visual features provided by ImageCLEF organizer

We did not run the mixed retrieval process on the whole collection due to the computational cost. We ran the text retrieval first, then applied mixed retrieval to the re-ranking.

From Table 6.13, we can see that the retrieval results based on visual features only have very low MAP. The retrieval results of the mixed runs are also considerably lower than our results from ImageCLEF2007. After investigating the tensor-based experiments, we found a bug in the code. The $+$ operator was missing, hence only the last textual and content feature dimension had been used to re-rank images. This accounts for the poor performance of the submission runs. We have corrected the code and re-run the experiments. The text based result was MAP=0.0939 and P@10=0.3485; The tensor product based result with our local features was MAP=0.0665 and P@10=0.2000. There was a small improvement after removing the bug from the experiment, but it was still worse than the text based retrieval, which needs a further investigation.

We did not include the correlations between dimensions across different feature spaces in the 2010 year submission. In the ImageCLEF2010 experiments, we assume that all dimensions are pair-wise orthogonal.

| Run | Modality | field(s) | MAP | P@10 |
|---|---|---|---|---|
| combine | Mixed | TITLEIMG | 0.0617 | 0.2271 |
| TFL | Mixed | TITLEIMG | 0.0617 | 0.2257 |
| TXFC | Mixed | TITLEIMG | 0.0486 | 0.1443 |
| TXFL | Mixed | TITLEIMG | 0.0484 | 0.1500 |
| TFC | Mixed | TITLEIMG | 0.0325 | 0.1143 |
| WFC | Mixed | TITLEIMG | 0.0031 | 0.0086 |
| cleszek | Visual | IMG | 0.0069 | 0.0614 |
| cadd | Visual | IMG | 0.0003 | 0.0100 |

Table 6.13: Our ImageCLEF2010 runs

### 6.4.3 Conclusion and Future Work

Here, we evaluated a quantum theory inspired multimedia retrieval framework based on the tensor product of feature spaces, where the similarity measurement between a query and a document corresponds to the quantum measurement. At the same time, the correlations between dimensions across different feature spaces can also be naturally incorporated in the framework. The tensor based model provides a formal and flexible way to expand the feature spaces, and seamlessly integrate different features, potentially enabling multi-modal and cross media search in a principled and unified framework.

Experimental results on a standard multimedia benchmarking collection show that the quantum-like measurement on a tensored space leads to remarkable performance improvements in terms of average precision over the use of individual feature spaces separately or concatenation of them. However, the incorporation of dimension-wise correlation across feature spaces in the tensor model did not lead to performance improvement. Further investigation is needed in this direction.

In our experiments, we assumed that all dimensions are orthogonal to each other. This assumption, however, can be relaxed. We can either replace the synonyms with one representative word, or apply dimensionality reduction techniques in our tensor model. This is also sensible from a practical point of view, as a high dimensional textual feature space can make the computation of the ranking scores difficult on a large scale collection. Moreover, we would like to test the tensor product model on a wide selection of visual content features.

## 6.5 Tensor Product of Correlated Text and Visual Features. Associating Textual and Visual Features Dimensions

In the original hybrid model two relatively straightforward statistical approaches for making associations between dimensions of both feature spaces were employed, but with unsatisfactory results. The tensor model requires that all images have textual annotations. Here, we propose to alleviate the problem regarding unannotated images by projecting them onto subspaces representing visual

context and by incorporating a quantum-like measurement (Chapter 4.2).

### 6.5.1 Experimental Setup

Here, we solely test the subspace-based auto-annotation methods. We manually choose a few terms and construct a subspace for each term. The projection of the unannotated image to the semantic subspace can be utilized to decide whether the image is about the term. We experiment on ImageCLEF 2007 data collection.

The terms are selected from the query text, some of which have explicit visual characteristics (e.g. sea), while others do not have general visual characteristics (e.g. California).

The measurement operator is constructed from all the images with relevant content. To simplify the experiment, we look at the available ground-truth data and choose 5 images that are specifically about the term, and construct the correlation matrix from the term-document matrix M.

We manually select 10 relevant images belonging to each topic and 60 irrelevant images to investigate how the sub-space can distinguish the relevant from irrelevant images. The visual features we choose are: global colour histogram in HSV colour space, and local feature based on the bag of visual words approach. The latter consists of image sampling (random, dense sampling), description of local patches (three colour moments), quantization of descriptors (k-means) and generation of histograms of visual words counts.

The auto-annotation methods utilized in the experiments for the comparison are: orthogonal projection, quantum measurement, and distance based. The latter clusters the training images containing given tags and the distance between cluster centroids and the unannotated image is used as the score for image - text association.

- Projection

$$P = \sum_i |u_i\rangle\langle u_i| \tag{6.9}$$

  The probability of document out of a space can be compute as:

$$P_i = tr(D_i \cdot P) \tag{6.10}$$

- Quantum measurement

$$P_i = tr(D_i \cdot M) \tag{6.11}$$

  where $D_i$ is the density matrix for a document $d_i$, $D = |d_i\rangle\langle d_i|$. In order to represent the density matrix of the document $d_i$ with basis U, we need to do the decomposition $D' = UDU^T$.

- Distance-based. To compare the effectiveness of subspace based method, we look at a distance based approach. Suppose that all the relevant images are clustered together, then a

new image should belong to this cluster if it is close to the centroid of the cluster. Thus, we cluster the images containing a given tag.

$$w_i = |d_i - c| \tag{6.12}$$

### 6.5.2 Results and Discussion

If a measurement operator can filter the relevant images with success, then this operator can be used to associate the text with visual features. The test results are shown in the Table 6.14. We can observe that cluster distance based measurement outperforms the other two. Here, *localdense* denotes local feature with dense sampling, *localrand* denotes local feature with random sampling, and *histHSV* is a colour histogram in HSV colour space.

The observation is out of our expectations, as the subspace based measurement is supposed to capture the relevance of context as well as the latent information. This may be due to the small-scale experiment that was performed.

These results may be also related to our assumption that the correlations at the image-level may be stronger that the correlations based on the proximity between visual terms (pixels, local patches). An image may contain correlated terms (pixels, visual words) not because of their proximity, but because they refer to the same topic (context represented by image). We were inspired by (Biancalana et al. 2009), where the page-based correlations (text) performed better than proximity based ones. We were aware, however, that this does not have to be transferable to image retrieval. Future experiments will verify this hypothesis.

### 6.5.3 Conclusion and Future Work

Here, we describe and test two novel approaches for making associations between tags and images. We also experiment with mid-level semantic image representations based on the "bag of visual words" model. This is a follow-up work on our tensor-based unified image retrieval framework. In order to prepare the data for the quantum measurement in the tensor space, we need to alleviate the problem regarding the unannotated images. The first proposed approach projects the unannotated images onto the subspaces generated by subsets of training images (containing given textual terms). We calculate the probability of an image being generated by the contextual factors related to the same topic. In this way, we should be able to capture the visual contextual properties of images, taking advantage of this extended vector space model framework. The other method performs quantum like measurement on the density matrix of an unannotated image, with respect to the density matrix representing the probability distribution obtained from the subset of training images. These approaches can be seamlessly integrated into our unified framework for image retrieval.

The experimental results show that the standard approach based on clustering works better than other methods. This may be due to the small-scale experiments conducted. Another reason

| | localdense | | | | | | localrand | | | | | | histHSV | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *q* | | *p* | | *d* | | *q* | | *p* | | *d* | | *q* | | *p* | | *d* | |
| | p5 | p10 | p5 | p10 | p5 | p10 | p5 | p10 | p5 | p10 | p5 | p10 | p5 | p10 | p5 | p10 | p5 | p10 |
| mountain | 1 | 2 | 1 | 2 | 1 | 2 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 0 | 1 | 1 | 2 |
| sea | 2 | 6 | 4 | 6 | 5 | 6 | 4 | 6 | 4 | 6 | 2 | 4 | 4 | 8 | 4 | 8 | 3 | 4 |
| straight | 0 | 1 | 0 | 1 | 2 | 4 | 1 | 2 | 1 | 3 | 4 | 5 | 0 | 2 | 0 | 1 | 3 | 4 |
| black_white | 4 | 9 | 4 | 8 | 5 | 9 | 4 | 8 | 3 | 8 | 5 | 9 | 5 | 9 | 5 | 9 | 5 | 8 |
| girl | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 3 | 4 | 1 | 1 | 0 | 1 | 1 | 2 |
| California | 1 | 1 | 0 | 3 | 2 | 4 | 2 | 2 | 0 | 1 | 2 | 4 | 0 | 1 | 0 | 1 | 1 | 1 |

Table 6.14: Accuracy of different measurements on various visual features. Here, $q$ denotes quantum-like measurement, and $p$ and $d$ correspond to projection and distance based measurements respectively; the values in the table correspond to the number of positive associations at different precision levels.

for these surprising results may be related to the assumption we made, that the correlations at the image-level may be stronger that the correlations based on the proximity between visual terms

(pixels, local patches). Recent works build the correlations based on the proximity between image patches to capture the spatial information, as researchers believe that the relative distance between them is important. Thus, we need to test this alternative method for correlation matrix generation and perform large scale experiments.

## 6.6 Combining Systems in the Context of Relevance Feedback

Relevance feedback can narrow down the search and put the query in the right context. In the hybrid relevance model, apart from the visual and textual query representations, we would have additional information in the form of sets of feedback images. The sets of feedback images would themselves consist of subsets (in Hilbert space - subspaces) of visual and textual feedback representations. Moreover, there are some inherent inter and intra relationships between visual and textual feature spaces. These relationships have not yet been exploited in the context of user feedback.

Here, we evaluate a model for the combination of visual and textual sub-systems within the aforementioned user feedback context (Chapter 5.1). The model was inspired by the measurement utilized in quantum mechanics (QM) and the tensor product of co-occurrence (density) matrices, which represents a density matrix of the composite system in QM. It provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context. The proposed approach takes into account both intra (via co-occurrence matrices) and inter (via tensor operator) relationships between features' dimensions. It is also computationally cheap and scalable to large data collections. We test our approach on the ImageCLEF2007photo data collection and present interesting findings.

### 6.6.1 Experimental Setup

We evaluate the proposed model on the ImageCLEFphoto2007 data collection, within the user simulation framework. We only test the hybrid relevance model on one collection because of the nature of our evaluation framework (user simulation).

We test our model (expectation value with a tensor product of density matrices) within a simulated user feedback framework. First, we perform the first round retrieval for a topic from the query set based on the visual features only (we retrieve 1000 images). We use the visual features only because in the real life scenario many images would not have textual descriptions. We also do not combine the features in the first round retrieval as this would represent a different task. In this work we want to focus on testing the feature combination models within the user feedback framework.

Next, we identify 1, 2 and 3 relevant images respectively from the highest ranked images based on the ground truth data. The thus obtained images simulate the user feedback and are utilized in the proposed model to re-score the data collection. For each query topic (60 in total) we calculate

mean average precision (MAP) for the top 20 retrieved images, as most users would only look at this number of documents. We set the weights $r_1$, $r_2$ to 1 and 0.8 respectively (standard weight values for the query and its context as in the classic Rocchio algorithm, for example).

The visual features used in the experiment are based on the Bag of Visual Words framework, and were introduced in the previous chapters. They are regarded as a mid-level representation.

The textual features were obtained by applying the standard Bag of Words technique, with Porter stemming, stop words removal, and term frequency - inverse document frequency weighting scheme.

As aforementioned, we modify existing models in order to incorporate the user feedback. We use several baselines for comparison purposes.

Thus, early fusion is represented by a modified Rocchio algorithm ($earlyFusion$). The only difference between this variation and the classic model is that we apply it to concatenated visual and textual vectors, as opposed to visual or textual representations only. Let $\oplus$ denote the concatenation operation (other notation as in the previous sections). Then, in this model we modify the query in the following way

$$newQuery = q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \tag{6.13}$$

After the query modification the scores are recomputed.

Another baseline, which we will refer to as $lateFusion$ will be represented as a combination of all the scores

$$s(q_v, a) + \frac{0.8}{n} \sum_i s(c_i, a) + s(q_t, b) + \frac{0.8}{n} \sum_i s(d_i, b) \tag{6.14}$$

where $s$ denotes the similarity between given vectors. In this work $s$ is an inner product between two vectors.

Our third baseline $rerankText$ denotes the re-ranking of the results obtained from the first round retrieval based on the aggregated textual representations of the feedback images. Similarly, $rerankVis$ represents re-ranking of the top retrieved images based on the aggregated visual representations of the images from the feedback set.

The next model $trMedia$ represents, as the label implies, inter-media feedback query modification. Here, textual annotations from the feedback images (identified by visual features) are used to expand a textual query.

The system performance without simulated feedback will be denoted as $noFeedback$ and the proposed model for combination of visual and textual features within the context of simulated relevance feedback will be denoted as $prMeanMeasure$.

### 6.6.2 Results and Discussion

Table 6.15 presents the obtained results.

Table 6.15: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP)

| | **1 Feedback Image** | **2 Feedback Images** | **3 Feedback Images** |
|---|---|---|---|
| $noFeedback$ | 0.013 | 0.013 | 0.013 |
| **prMeanMeasure** | **0.076** | **0.094** | **0.11** |
| $earlyFusion$ | 0.066 | 0.082 | 0.085 |
| $lateFusion$ | 0.066 | 0.082 | 0.085 |
| $rerankText$ | 0.055 | 0.069 | 0.075 |
| $rerankVis$ | 0.034 | 0.036 | 0.031 |
| $trMedia$ | 0.061 | 0.078 | 0.081 |

From the experimental results we can see that the best performing model is based on the proposed predicted mean value of the measurement ($prMeanMeasure$) with the density matrix of the composite system (tensor product of the subspaces). The difference (in terms of means) between $prMeanMeasure$ and the rest of the baselines is statistically significant (paired t-test, $p < 0.05$). The inter-media feedback query expansion ($trMedia$) also performed well, albeit worse than early and late fusion ($earlyFusion$, $lateFusion$). In general, all the models' performances suggest that they are quite effective in utilizing users' feedback.

An interesting observation is that both early ($earlyFusion$, modified Rocchio) and late fusion strategies ($lateFusion$, combination of scores) show exactly the same performance. This is because

$$newQuery = q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \tag{6.15}$$

$$imagesInDataset = a \oplus b \quad forAll \quad a, b \in Dataset \tag{6.16}$$

$$\begin{aligned} \langle newQuery | imagesInDataset \rangle = \\ \left\langle q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \, | a \oplus b \right\rangle = \\ \langle q_v \oplus q_t | a \oplus b \rangle + \frac{0.8}{n} \sum_i \langle c_i \oplus d_i | a \oplus b \rangle = \\ \langle q_v | a \rangle + \frac{0.8}{n} \sum_i \langle c_i | a \rangle + \langle q_t | b \rangle + \frac{0.8}{n} \sum_i \langle d_i | b \rangle \end{aligned} \tag{6.17}$$

Thus, in our case the early and late fusion strategies (modified Rocchio algorithm operating on concatenated representations and weighted linear combination of scores) are interchangeable. We have addressed this interesting discovery in the previous chapters.

We observe that even one feedback image can help to narrow down the search, thus increasing the match between the user's preferences (in this case, a human expert who assessed the relevance of images in ground truth data). Let us assume, that the visual query pictures a person wearing sunglasses. In the first round retrieval, the system may recognize (return more images of) a concept representing sunglasses without a person present on the picture. However, the human assessor

might have deemed an image relevant only if both concepts were present in the image. A user feedback can then reinforce the subjective (perceived) relevance of the query to the retrieved images. In case of using the visual representations only in the user feedback ($rerankVis$), more images in the feedback set can sometimes confuse the visual features (especially if they significantly differ in terms of colour, texture, viewpoint or illumination). Thus, approaches like $rerankVis$ may strongly depend on the type of visual features used (while visual features A may be suitable for the particular feedback set C, visual features B may not work so well on C and vice versa).

In this work, the MAP is calculated for the 20 top images only as this is a more realistic scenario (especially for user simulation/user feedback context). However, for 1000 top and 3 feedback images, the system performance is approximately $MAP \approx 0.206$. If we consider the ImageCLEF2007photo results of other systems (the best models utilize both visual and textual information) which can be found on the ImageCLEF website (Author n.d.a), the proposed model places itself among the best performing approaches. However, it must be noted that our model combines visual and textual features within the context of user feedback framework ( a different task).

We also need to take into consideration the disadvantages of automatic evaluation methods. The ultimate test for every retrieval system (especially for user simulation/user feedback context) should be the real user evaluation (although it is a time consuming task). The relevance of an image is a highly subjective concept and automatic evaluation seems to fail to address this problem. Moreover, there is a glitch in the trec-eval evaluation software, that can bring the reported results into question. To be more specific, if some images obtain the same similarity score, they will be re-ordered by the software. The result is that two identical submissions may get different performance scores.

### 6.6.3 Conclusions and Future Work

We have evaluated the model for visual and textual feature combination within the context of user feedback. The approach is based on mathematical tools also used in quantum mechanics - the predicted mean value of the measurement and the tensor product of the density matrices, which represents a density matrix of the combined systems. It was designed to capture both intra-relationships between features' dimensions (visual and textual correlation matrices) and inter-relationships between visual and textual representations (tensor product). The model provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context by applying quantum-like measurement. It opens a door for a series of theoretically well-founded further exploration routes, e.g. by considering the interference among different features. It is easily scalable to large data collections as it is general and computationally cheap. The results of the experiment conducted on the ImageCLEF data collection show significant improvement over other baselines.

Future work will involve testing different notions of correlation within the proposed framework

(we can construct correlation matrices in such a way that they can be regarded as density matrices). In this experiment, we incorporate document/image level correlations only. However, in the case of textual representations, we can also experiment with Hyperspace Analogue to Language (HAL). In the aforementioned approach, the context is represented by a sliding window of a fixed size (while in document level correlation the context is represented by the whole document). We can also consider a visual counterpart to HAL, where a window of a fixed size (e.g. square, circular) is shifted from one instance of a visual word to another. Then, the number of instances of visual words that appear in the proximity of the visual word on which the window is centered can be calculated. In the case of a dense sampling, the window would be shifted analogously to HAL in text IR. If the sparse sampling was utilized, however, the window would shift from one instance of a visual word to another.

## 6.7 Dynamic Weighting of the Query and Its Context

Here, we enhance the original hybrid model by incorporating an adaptive weighting scheme (Chapter 5.3). Thus, the respective weights are automatically modified, depending on the relationship strength between the visual query and its visual context and textual query and its textual context; the number of terms or visual terms (mid-level visual features) co-occurring between the current query and the context. The user simulation experiment shows that this kind of adaptation can indeed further improve the effectiveness of CBIR.

### 6.7.1 Experimental Setup

We evaluate the proposed approach on the ImageCLEFphoto2007 data collection.

For the consistency and comparison with the fixed weight model, we test the adaptive weight approach in a simulated user feedback framework. First, we perform the first round retrieval for a topic from the query set based on the visual features only (we retrieve 1000 images). We use the visual features only because in a real life scenario many images would not have textual descriptions. We also do not combine the features in the first round retrieval as this would represent a different task. In this work we want to focus on testing the feature combination models within the user feedback framework.

We identify 1, 2 and 3 relevant images respectively from the highest ranked images based on the ground truth data. Thus obtained images simulate the user feedback and are utilized in the proposed model to re-score the data collection. For each query topic (60 in total) we calculate mean average precision (MAP) for the top 20 retrieved images, as it is unlikely that users would look at more than this number of documents. For the fixed weight model, we test different combinations of parameters and choose the ones performing best for a fair comparison with our adaptive weighting scheme.

The visual features used in the experiment are based on the Bag of Features (BOF) framework. Here, we utilize random sampling (best in generic image retrieval when the number of

sample points is high (Nowak et al. 2006a)) to get 900 sample points. We use colour moments as descriptors and generate 40 dimensional vectors of visual words counts.

In addition to the local features, we also experiment with a global method - colour histogram in RGB colour space. First, each image is split into individual colour channels (a greyscale representation of an individual colour). Next, pixel intensities corresponding to each colour channel are quantized into 8 bins. Thus obtained three histograms are concatenated to form a 24 dimensional colour histogram.

The textual features were obtained by applying the standard Bag of Words technique, with Porter stemming, stop words removal, and term frequency - inverse document frequency (TF.IDF) weighting scheme.

Let *prMMFixed* and *prMMAdapt* denote the hybrid CBIR relevance feedback model with fixed weights and the enhanced model with the adaptive weighting scheme, respectively. We also test the adaptive capabilities of the visual and textual elements of the model. Thus, by *vOnlyFix*, *tOnlyFix* we will denote the visual and textual parts of the model with fixed weights and *vOnlyAd*, *tOnlyAd* will represent the visual and textual parts of the model with the adaptive weighting scheme.

Early fusion is represented by a modified Rocchio algorithm (*eFus*). The only difference between this variation and the classic model is that it is applied to concatenated visual and textual vectors, as opposed to visual or textual representations only. Let $\oplus$ denote the concatenation operation. Then, this model modifies the query in the following way

$$newQuery = q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \tag{6.18}$$

After the query modification the scores are recomputed.

Another baseline, which we will refer to as *lFus* will be represented as a combination of all the scores

$$\begin{aligned} s(q_v, a) + s(q_t, b) + \frac{0.8}{n} \sum_i s(c_i, a) + \\ \frac{0.8}{n} \sum_i s(d_i, b) \end{aligned} \tag{6.19}$$

where $s$ denotes the similarity between given vectors. In this work $s$ is an inner product between two vectors.

We can observe, that the performance of the two aforementioned baselines must be exactly the

same. This stems from the fact, that

$$newQuery = q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \tag{6.20}$$

$$imagesInDataset = a \oplus b \;\; forAll \;\; a, b \in Dataset \tag{6.21}$$

$$\langle newQuery | imagesInDataset \rangle = \left\langle q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \middle| a \oplus b \right\rangle = \langle q_v \oplus q_t | a \oplus b \rangle + \frac{0.8}{n} \sum_i \langle c_i \oplus d_i | a \oplus b \rangle = \langle q_v | a \rangle + \frac{0.8}{n} \sum_i \langle c_i | a \rangle + \langle q_t | b \rangle + \frac{0.8}{n} \sum_i \langle d_i | b \rangle \tag{6.22}$$

Thus, in our case the early and late fusion strategies (modified Rocchio algorithm operating on concatenated representations and weighted linear combination of scores) are interchangeable.

Our third baseline *rrText* denotes the re-ranking of the results obtained from the first round retrieval based on the aggregated textual representations of the feedback images. Similarly, *rrVis* represents re-ranking of the top retrieved images based on the aggregated visual representations of the images from the feedback set.

Our next model *trMed* represents inter-media feedback query modification. Here, textual annotations from the feedback images (identified by visual features) are used to modify a textual query.

The system performance without simulated feedback will be denoted as *noFeed*.

### 6.7.2 Results and Discussion

In this work, the Mean Average Precision (MAP) is calculated for the 20 top images only as this is a more realistic scenario (especially for a user simulation/user feedback context).[2]

First, let us check the performance of individual components of the original model for different values of parameters $r_1$, $r_2$ (fixed weights). We will select the best combination of parameters for a fair comparison with the adaptive weighting scheme. Tables 6.16 and 6.17 show the performance of visual and textual components of the hybrid CBIR relevance feedback model for different parameters' values (fixed weights). The last row displays the adaptive capabilities of weights $r_1$, $r_2$. Significantly different results (adaptive part against the fixed weights models) are displayed in bold font in Tables 6.16, 6.17, 6.18 ($p = 0.05$; paired t-test).

We can observe, that different values of parameter $r$ have different impacts on the component's performance. Moreover, the visual part of the model with adaptive weights performed significantly

[2]For 1000 top and 3 feedback images, the original system's performance is approximately $MAP \approx 0.206$. If we consider the ImageCLEF2007photo results of other systems (the best models utilize both visual and textual information) which can be found on the ImageCLEF website (Author n.d.a), the hybrid CBIR relevance feedback model places itself among the best performing approaches.

Table 6.16: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP), Visual part only

| | **1 Feed Im** | **2 Feed Im** | **3 Feed Im** |
|---|---|---|---|
| $noFeed$ | 0.013 | 0.013 | 0.013 |
| $r_1 = 0$ | 0.041 | 0.053 | 0.061 |
| $r_1 = 0.2$ | 0.036 | 0.060 | 0.070 |
| $r_1 = 0.4$ | 0.036 | 0.046 | 0.047 |
| $r_1 = 0.5$ | 0.033 | 0.038 | 0.044 |
| $r_1 = 0.6$ | 0.027 | 0.034 | 0.041 |
| $r_1 = 0.8$ | 0.020 | 0.031 | 0.039 |
| $r_1 = 1$ | 0.018 | 0.029 | 0.038 |
| $\mathbf{r_1 adapt}$ | 0.036 | 0.063 | **0.081** |

Table 6.17: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP), Textual part only

| | **1 Feed Im** | **2 Feed Im** | **3 Feed Im** |
|---|---|---|---|
| $noFeed$ | 0.013 | 0.013 | 0.013 |
| $r_2 = 0$ | 0.058 | 0.072 | 0.075 |
| $r_2 = 0.2$ | 0.063 | 0.076 | 0.079 |
| $r_2 = 0.4$ | 0.063 | 0.076 | 0.081 |
| $r_2 = 0.5$ | 0.062 | 0.080 | 0.080 |
| $r_2 = 0.6$ | 0.062 | 0.079 | 0.080 |
| $r_2 = 0.8$ | 0.062 | 0.079 | 0.080 |
| $r_2 = 1$ | 0.052 | 0.071 | 0.071 |
| $\mathbf{r_2 adapt}$ | **0.085** | **0.095** | **0.112** |

better than the fixed weights part for the higher number of feedback images. The textual part with adaptive weights shows even better adaptive capabilities of its weights.

Table 6.18 shows the performance of the hybrid CBIR relevance feedback model for different combinations of parameters values (fixed weights), and the results of the enhanced model with the adaptive weighting scheme.

Table 6.18: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP)

| | **1 Feed Im** | **2 Feed Im** | **3 Feed Im** |
|---|---|---|---|
| $noFeed$ | 0.013 | 0.013 | 0.013 |
| $r_1 = 0.2; r_2 = 0.4$ | 0.080 | 0.098 | 0.115 |
| $r_1 = 0.4; r_2 = 0.2$ | 0.082 | 0.101 | 0.114 |
| $r_1 = 0.5; r_2 = 0.5$ | 0.082 | 0.097 | 0.115 |
| $r_1 = 0.2; r_2 = 0.8$ | 0.081 | 0.098 | 0.116 |
| $r_1 = 0.8; r_2 = 0.2$ | 0.084 | 0.096 | 0.113 |
| $r_1 = 0.2; r_2 = 0.2$ | 0.081 | 0.096 | 0.113 |
| $r_1 = 0.8; r_2 = 0.8$ | 0.084 | 0.097 | 0.115 |
| $\mathbf{r_1, r_2 adapt}$ | 0.091 | **0.12** | **0.142** |

Although individual components of the hybrid CBIR relevance feedback model exhibited some sensitivity to the changing values of weights $r_1$, $r_2$, the combined model's performance is relatively stable, regardless of the values of the fixed weights. However, if the weights are automatically adjusted for each individual query, the enhanced model performs significantly better (for more than two images in the feedback set).

Finally, the overall comparison of different models is shown in Table 6.19 (In Tables 6.19 and 6.20, the significantly different results are displayed in bold font in the tables ($p = 0.05$; paired t-test). We will denote the statistical significance over the fixed weights model by *. The bold font and the absence of the * symbol will then represent the statistical significance over the baselines).

Table 6.19: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP)

| | **1 Feed Im** | **2 Feed Im** | **3 Feed Im** |
|---|---|---|---|
| $noFeed$ | 0.013 | 0.013 | 0.013 |
| $eFus$ | 0.066 | 0.082 | 0.085 |
| $lFus$ | 0.066 | 0.082 | 0.085 |
| $rrText$ | 0.055 | 0.069 | 0.075 |
| $rrVis$ | 0.034 | 0.036 | 0.031 |
| $trMed$ | 0.061 | 0.078 | 0.081 |
| $tOnlyFix$ | 0.063 | 0.080 | 0.081 |
| $vOnlyFix$ | 0.041 | 0.060 | 0.070 |
| $tOnlyAd$ | 0.085 | 0.095 | 0.112 |
| $vOnlyAd$ | 0.036 | 0.063 | 0.081 |
| $r_1, r_2 fixed$ | **0.084** | **0.101** | **0.116** |
| $\mathbf{r_1, r_2 adapt}$ | 0.091 | **0.12*** | **0.142*** |

Our main focus here should be on the difference in performance of the original fixed weight model and our enhanced model with the adaptive weighting scheme. It has been shown that the hybrid CBIR relevance feedback model outperformed other state-of-the-art hybrid systems that can be modified to incorporate user feedback. Our experiments confirm the previous findings. In general, the enhanced model significantly outperformed the original one (for more than one image in the feedback set).

Let us now add another visual feature, a global colour histogram computed in RGB colour space. In the case of an early fusion model, this extra visual feature will be concatenated with the combined (concatenated) vector of local and text features. Late fusion will naturally incorporate colour histogram as additional aggregation factor. Pre-filtering is going to involve an additional, last step, re-ranking by colour histogram. Similarly for the transmedia fusion approach, we add an extra step, aggregation of the colour histogram representation scores corresponding to the top retrieved images.

The results are presented in Table 6.20. The hybrid CBIR relevance feedback model with fixed parameters values performed best for $r_1 = 0.2$, $r_2 = 0.2$, $r_3 = 0.8$, where $r_1$, $r_2$ correspond to

global and local visual features respectively, and $r_3$ corresponds to the textual feature.

Table 6.20: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP) with additional visual feature

| | **1 Feed Im** | **2 Feed Im** | **3 Feed Im** |
|---|---|---|---|
| $noFeed$ | 0.013 | 0.013 | 0.013 |
| $eFus$ | 0.066 | 0.082 | 0.085 |
| $lFus$ | 0.066 | 0.082 | 0.085 |
| $rrText$ | 0.053 | 0.068 | 0.075 |
| $rrVis$ | 0.034 | 0.035 | 0.032 |
| $trMed$ | 0.064 | 0.080 | 0.083 |
| $r_1, r_2, r_3 fixed$ | **0.088** | **0.107** | **0.120** |
| $\mathbf{r_1, r_2, r_3 adapt}$ | 0.093 | **0.129*** | **0.153*** |

From the results table we can see that the baselines did not benefit much from the addition of global visual features. However, both the hybrid CBIR relevance feedback model and our enhanced one recorded an improvement in terms of MAP.

It is evident that the query and its context can be less or more related. We can utilize the information about this relationship strength to automatically adjust the weights corresponding to query and its context in relevance feedback. Thus, each query (visual example) can be associated with a particular combination of weights, unique for this query.

### 6.7.3 Conclusion and Future Work

A combination of visual and textual systems in the context of user feedback can help to reduce the semantic gap, the difference between human perception and machine representation of images. Because the feature spaces are complementary and correlated (describing the same information object), it is possible to model and exploit the inter and intra relationships between them. A hybrid model based on the tensor product of co-occurrence matrices proved to be an effective tool in combining the features in the context of user feedback. It outperformed other state-of-the-art hybrid approaches that could be modified to incorporate user feedback.

The aforementioned model for the visual and textual feature combination utilizes fixed weights corresponding to the importance of query and its context. However, the query can be more or less related to its context. Inspired by this observation, we incorporate an adaptive weighting scheme into the hybrid CBIR relevance feedback model. Thus, each query is associated with unique set of weights corresponding to the relationship strength between the visual query and its visual context as well as the textual query and its textual context. The higher the number of terms or visual terms (mid-level features) co-occurring between the current query and the context, the stronger the relationship and vice versa. If the relationship between query and its context is weak, context becomes important. If we adjust the probability of the original query terms, the adjustment will significantly modify the original query. If the aforementioned relationship (similarity) between the query and its context is strong, however, context will not help much. The original query terms

will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

We tested the enhanced model within the user simulation framework. For fair comparison purposes, the best performing sets of fixed weights were selected. We have shown, that our enhanced model with the adaptive weighting scheme can outperform the original one with fixed weights. Moreover, an addition of another visual feature (colour histogram, global feature) further improved both the hybrid CBIR relevance feedback model and the enhanced model's performance, whereas the performance of the baselines did not change much.

Our contribution here is related to showing how to measure the relationship strength between the query and its context, and how to incorporate the adaptive weighting scheme into the state-of-the-art existing model (hybrid, user feedback context) to further improve the retrieval. The proposed adaptive weighting approach is relatively easy to implement and does not require any training of features.

Our future work may involve an adaptation of the weights associated with the importance of visual and textual representations in the context of user feedback. Both the original and our enhanced models do not allow for adaptive weighting of visual and textual features. However, it could be advantageous to be able to infer from the user feedback, the subjective importance of the visual and textual parts of the system.

The original model (represented in the following form, not the end product) can be interpreted as a fusion-like strategy similar to combPROD late fusion

$$\left\langle M_1 \middle| a^T \cdot a \right\rangle \left\langle M_2 \middle| b^T \cdot b \right\rangle \tag{6.23}$$

However, the weighted linear combination with adaptive weights corresponding to the importance of visual and textual features can have even better capability for performance improvement (model personalization). Thus, we can treat our model as a kind of a weighted linear combination

$$\begin{aligned} r_v \left\langle M_1 \middle| a^T \cdot a \right\rangle + r_t \left\langle M_2 \middle| b^T \cdot b \right\rangle = \\ r_v \left( str_v \left\langle q_v | a \right\rangle^2 + (1 - str_v) \frac{1}{n} \sum_i \left\langle c^i \middle| a \right\rangle^2 \right) + \\ r_t \left( str_t \left\langle q_t | b \right\rangle^2 + (1 - str_t) \frac{1}{n} \sum_i \left\langle d^i \middle| b \right\rangle^2 \right) \end{aligned} \tag{6.24}$$

$$str_v = \frac{\left\langle D_q^v \middle| D_f^v \right\rangle}{\left\| D_q^v \right\| \left\| D_f^v \right\|} \tag{6.25}$$

$$str_t = \frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\| D_q^t \right\| \left\| D_f^t \right\|} \tag{6.26}$$

where the weights corresponding to the importance of visual and textual features, $r_v$ and $r_t$, could

be somehow automatically modified depending on the inferred user preferences and interests.

The future work may also involve testing different notions of correlation within the proposed framework (we can construct correlation matrices in such a way that they can be regarded as density matrices). In this paper, we incorporate document/image level correlations only. However, in the case of textual representations, we can also experiment with Hyperspace Analogue to Language (HAL). In the aforementioned approach, the context is represented by a sliding window of a fixed size (while in document level correlation the context is represented by the whole document). We can also consider a visual counterpart to HAL, where a window of a fixed size (e.g. square, circular) is shifted from one instance of a visual word to another. Then, the number of instances of visual words that appear in the proximity of the visual word on which the window is centered can be calculated. In case of a dense sampling, the window would be shifted analogously to HAL in text IR. If the sparse sampling was utilized, however, the window would shift from one instance of a visual word to another.

## 6.8 Chapter Summary

In addition to the theoretical validation of our discoveries, we empirically evaluate our models. Thus, we start by describing the data collections utilized in the experiments.

We evaluate our novel global and local features on ImageCLEF2007, MIRFlickr, and BGS data collections. The local features were also evaluated in the ImageCLEF2010 Wikipedia Retrieval Task. In addition, we test the performance of local versus global visual features as well as the performance of various sampling techniques for the Bag of Visual Words framework.

The enhancement of the local features based on the correlations between the visual words was tested within the pseudo relevance feedback framework. This is the fully automated method that helped us improve the standard local features model.

The combination of textual and visual features, as well as the auto-annotation model, were evaluated in a small-scale experiment first. Next, the proper large-scale evaluation of the hybrid model was performed as part of our participation in the imageCLEF2010 competition.

Our hybrid relevance model (and the model with adaptive weighting scheme) for the combination of textual and visual features in the context of relevance feedback was tested within the user simulation framework.

In general, the experimental evaluation shows the effectiveness of our models.

# Chapter 7

# Interactive User Interface - Prototype System

In this chapter, we present an interactive user interface which has been synchronized with our prototype system. The user interface is the communication platform between the system and the user. In order to show a working demo of the prototype system, we have integrated it with the interactive user interface. Thus, we present a unified working framework (prototype system) for Content-based Image Retrieval comprising: novel visual features for generic image retrieval and their combinations with existing visual features, combination of visual and textual features in the first round retrieval, combination of visual and textual features in the context of relevance feedback (search refinement), interactive user interface. To the best of our knowledge, this is the first prototype system that utilizes the hybrid model for combination of visual and textual features as well as a hybrid relevance feedback model, and allows for a full interaction with the system. Figure 7.1 shows our interactive user interface. Let us describe the panels and functionalities of our interface.

**Visual Example Panel**

Here, the user can browse the data collection by clicking the browse button or exploiting the quick browse panel (*visual example*). Then, an image can be selected which will represent a visual example, to query the system. Thus the selected current query will be displayed in the panel *current query*. A double mouse click on the query image will result in a full screen image display. Having selected the visual example, the user can click on the search button which will perform the data collection search based on the hybrid model (more details later in the chapter).

**Positive Results Panel**

The positive results panel displays the positive results of a current search (first round retrieval

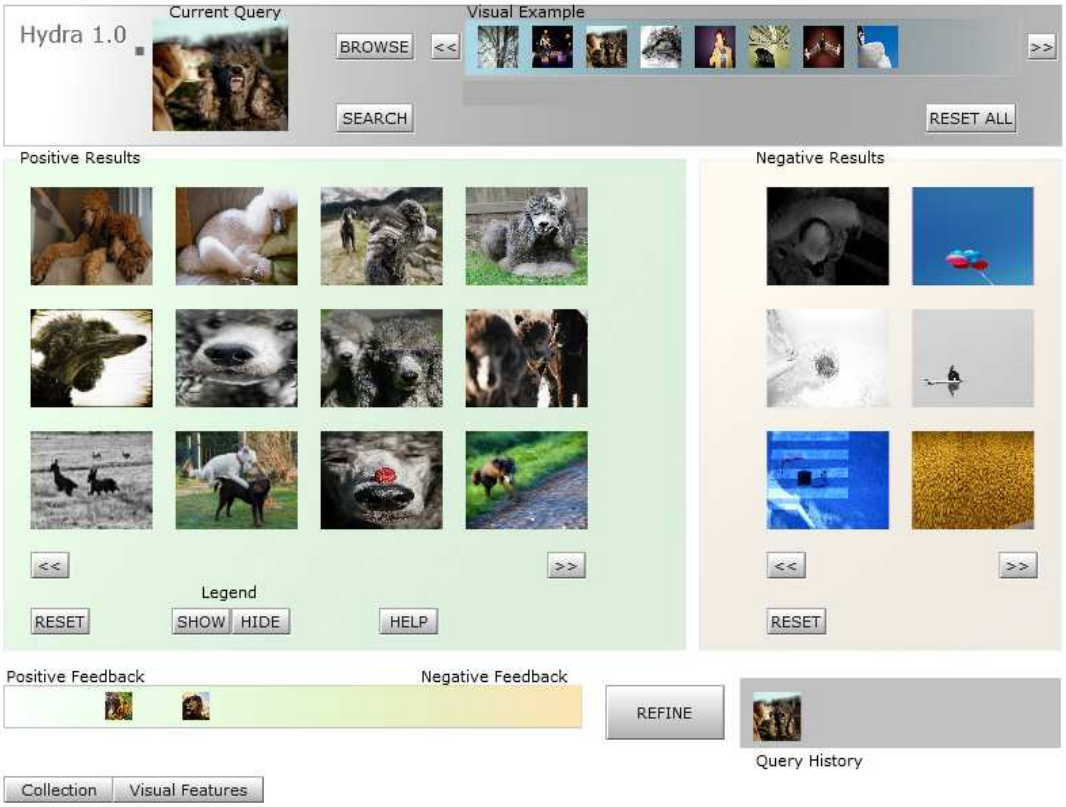


Figure 7.1: Interactive interface for the prototype hybrid system

or refined results based on the relevance feedback) starting from the most relevant (top left) image. Each result image can be displayed in a full screen mode by double clicking on it. Additionally, each result image can be dragged and dropped in the *current query* panel and thus become a current query itself. Buttons marked as "<<" and ">>" allow the user to browse the search results (exploratory search). Images from the positive panel results can be also dragged and dropped in the relevance bar, thus indicating the level of relevance of specific images.

**Negative Results Panel**

The negative results panel displays the negative results of a current search (first round retrieval or refined results based on the relevance feedback) starting from the least relevant (top left) image. Each result image can be displayed in a full screen mode by double clicking on it. Additionally, each result image can be dragged and dropped in the *current query* panel and thus become a current query itself. Buttons marked as "<<" and ">>" allow the user to browse the search results (exploratory search). Images from the negative panel results can be also dragged and dropped in the relevance bar, thus indicating the level of relevance of specific images. Thus, the user can always correct the performance of the system because even an image from the negative results panel

can become the current query or can be suggested as a relevant example (relevance feedback).

**Relevance Feedback Panel (Relevance Bar)**

The relevance bar indicates the levels of relevance of feedback images. The user can always change the levels of relevance of the feedback images already placed on the relevance bar. The degree of relevance is naturally represented in a form of a spectrum, with one end corresponding to positive feedback, and the other end corresponding to negative feedback. If the current query is selected, the user needs to continue the search and gave the feedback to the system, then clicking the refine button will narrow down and also correct the search.

**Query History Panel**

The expandable query history panel, as the name suggests, shows the previously issued queries. These queries are also utilized during the search refinement process.

**Other Functionalities**

Other functionalities provided by our interface are: reset buttons for resetting either the relevance feedback or query history, or both, show/hide buttons display or hide the panels' descriptions, the help button displays descriptions of provided functionalities, a left mouse click on the collection button displays a list of data collections (three image datasets), a left click on the visual features button displays a list of visual features, and various combinations of visual features.

## 7.1 Models Implemented in the Prototype System

We have implemented two hybrid models in our prototype system demo.

First is the hybrid model for the combination of visual and textual features

$$\sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2 s_c\left(d_1^t, d_2^t\right) + 2} \tag{7.1}$$

where $s_e$ denotes Euclidean distance, $s_c$ represents cosine similarity measure, $d_1^v$ and $d_1^t$ denote visual and textual representations of the query respectively, $d_2^v$ and $d_2^t$ denote visual and textual representations of an image in the data collection respectively, and $\otimes$ is the tensor operator. Thus, we utilize Euclidean distance to measure the similarity between visual representations, and the cosine similarity to measure the similarity between textual representations. Euclidean distance, in case of our mid-level visual features, performs better than cosine similarity. It is due to the fact that normalization of our local features hampers the retrieval performance. On the other hand, cosine similarity in textual space performs better than other similarity measurements.

Recall, that the aforementioned combination of measurements performed on individual feature spaces is equivalent to

$$\sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2s_c\left(d_1^t, d_2^t\right) + 2} = s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) \tag{7.2}$$

Thus, the implemented model is equivalent to computing the Euclidean distance on a tensored representation. Knowing about this equivalence helps us avoid the curse of dimensionality (no need to perform the tensor operation).

The second implemented hybrid model is the hybrid relevance feedback model for image re-ranking.

$$tr\left(\left(\otimes_n M_n\right) \cdot \left(\otimes_n \left(a_n^T \cdot a_n\right)\right)\right) = \prod_n \left\langle M_n \middle| a_n^T \cdot a_n \right\rangle \tag{7.3}$$

For 3 features (i.e. two visual and a textual feature) our enhanced model becomes

$$tr\left(\left(M_1 \otimes M_2 \otimes M_3\right)\left(\left(a_1^T a_1\right) \otimes \left(a_2^T a_2\right) \otimes \left(b^T b\right)\right)\right) = \left(\sum_i \left\langle c_1^i \middle| a_1 \right\rangle^2\right) \cdot \left(\sum_i \left\langle c_2^i \middle| a_2 \right\rangle^2\right) \cdot \left(\sum_i \left\langle d^i \middle| b \right\rangle^2\right) \tag{7.4}$$

Here, for example, $M_1$, $a_1$ and $M_2$, $a_2$ may correspond to different visual features (density matrices and vector representations of images from the data collection), and $M_3$, $b$ corresponds to a textual feature. Analogously, $c_1^i$ and $c_2^i$ denote visual feedback images corresponding to different visual features, $d^i$ denotes textual feedback images, $tr$ denotes the trace operator, and $\langle \cdot | \cdot \rangle$ is a standard inner product.

The visual features implemented in our prototype model comprise: edge histogram, homogeneous texture, bag of visual words (our low dimensional model), colour histogram, co-occurrence matrix, and the feature combinations.

## 7.2 Prototype System's Design

The image content modelling and visual feature extraction is performed in the C programming language. Implementation of image processing techniques is facilitated by INTEL's open source library OpenCV 1.0. Both visual and textual representations are stored in a MySQL database in the form of multidimensional vectors. The interactive user interface was developed in OpenLaszlo, an open source platform for the development and delivery of rich internet applications. The OpenLaszlo platform consists of the LZX programming language and the OpenLaszlo Server. The LZX programming language is an Extensible Markup Language (XML) and JavaScript descrip-

tion language similar to XUL, MXML, and Extensible Application Markup Language (XAML). The communication between the interactive user interface and the core search engine is facilitated by the GuestBlox technology, a simple yet complete "end-to-end" Web application powered by an OpenLaszlo frontend and a PHP backend. Because our models are fast and easy to implement, the core search technology is implemented as SQL scripts. The implementation comprises: the similarity measurement, the hybrid model for visual and textual features combination, and the hybrid relevance feedback model.

## 7.3 Prototype System at Work

Figures 7.2 and 7.3 present our prototype system at work.

## 7.4 Chapter Summary

This chapter presents our prototype system consisting of an interactive user interface, a hybrid model for the combination of visual and textual features, and a hybrid model for the combination of features in the context of relevance feedback.

The novel interactive interface offers the functionalities of monomodal prototype systems and more (various degrees of relevance represented as a relevance bar to which a user drags and drops feedback images in more natural way, integrated with the hybrid relevance feedback model).

Other functionalities provided are: exploratory search (browsing through positive and negative results, positive and negative results may become the current query, continuous control over the query history and images in the feedback set), query history, zoom in and out on the images of interest, positive and negative results which can both be utilized as feedback, selection of visual features and their combinations through the application of the generalization of our hybrid relevance feedback model and the hybrid model.

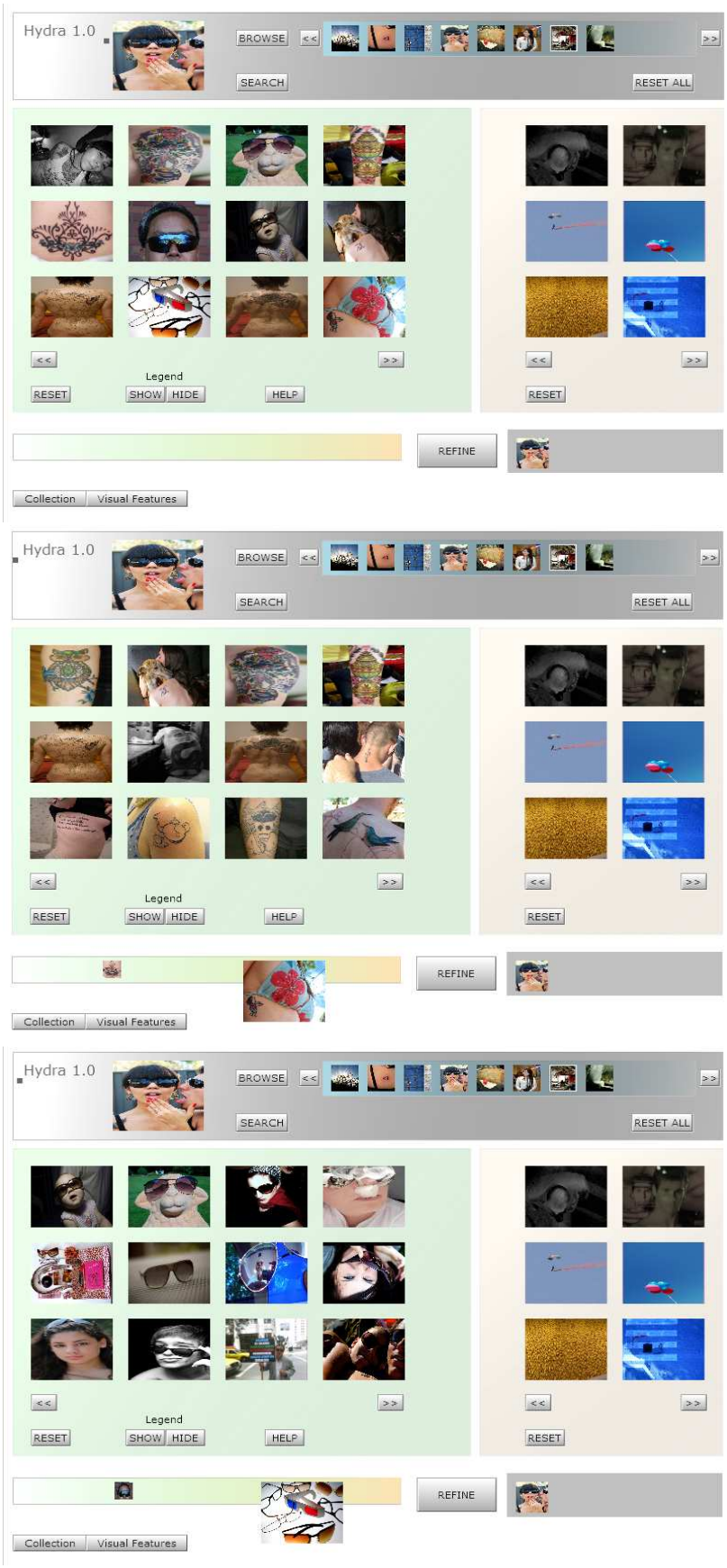


Figure 7.2: Narrowing down the search (combination of textual and visual features). Top: Users A and B query the system by visual example. The system identifies a few concepts and displays the results. Middle: User A specifies his interests by giving the feedback (tattoos) and refines the search. Bottom: User B specifies his interests by giving the feedback (sunglasses) and refines the search.

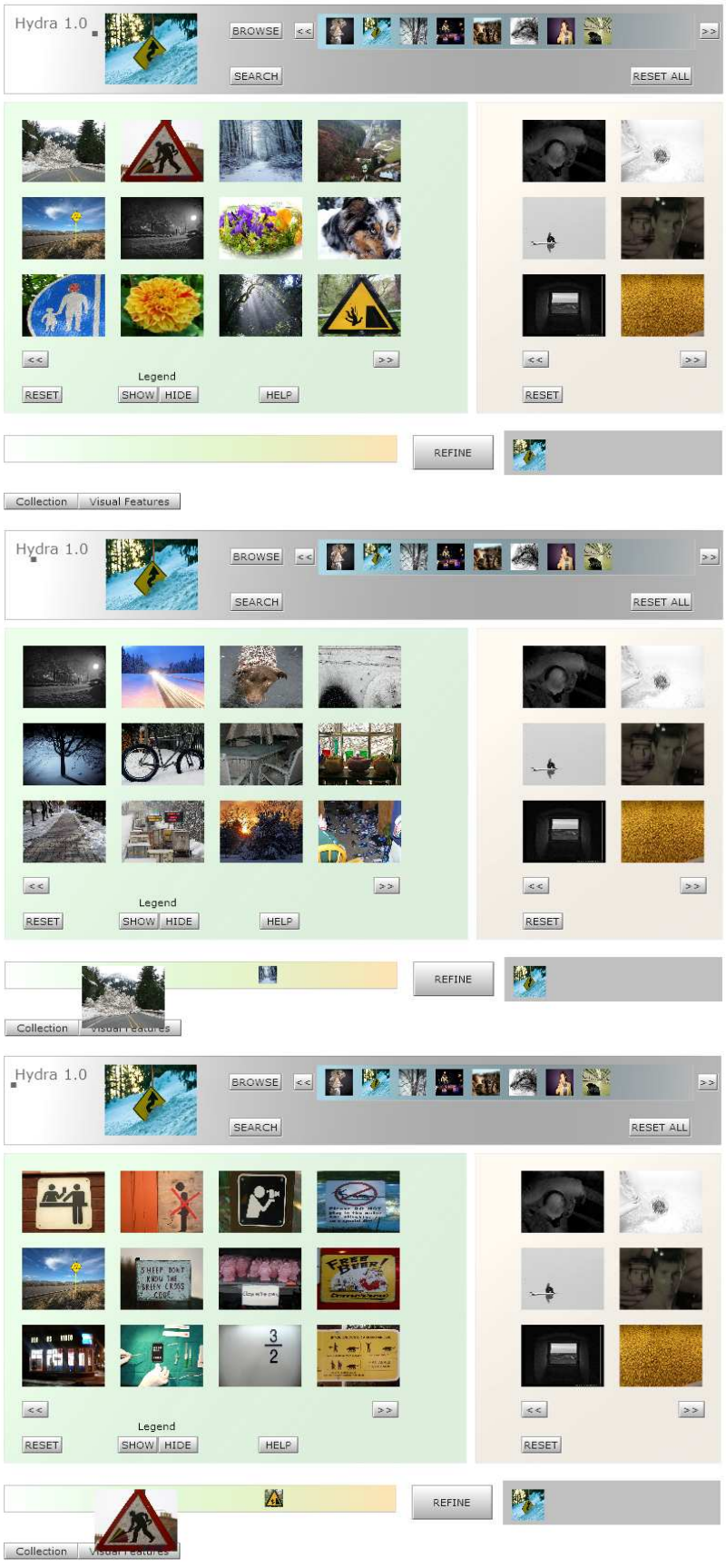


Figure 7.3: Narrowing down the search (combination of textual and visual features). Top: Users A and B query the system by visual example. The system identifies a few concepts and displays the results. Middle: User A specifies his interests by giving the feedback (snow) and refines the search. Bottom: User B specifies his interests by giving the feedback (signs) and refines the search.

# Chapter 8

# Other Applications of Our Novel Local Features

In this chapter, we would like to present two interesting applications of our novel local visual features, outside of the generic image retrieval task. First is the music genre classification based on the visual features. Second application is the food recognition task, with a prototype system developed for our industrial partner.

Music genre classification is based on the audio content features extracted from the audio signal. These features are often hard to extract and define, and represent such musical aspects as melody, timbre and rhythm, to name a few. We present an alternative approach to music genre classification, and classify music based on the visual similarities of music spectrograms generated in a colour space. This approach, and the possibility of treating real life images as spectrograms, points at an interchangeability of visual and auditory perception.

The application of visual features to food recognition domain is not a novel idea. However, the visual comparison of retrieval results shows, that the proposed low dimensional visual features can be quite effective in this particular domain.

## 8.1 Enhancing Music Genre Classification by Incorporating Image-Based Local Features

Existing approaches to music genre classification try to model such musical concepts as melody, timbre and rhythm, which is a difficult task.

The research questions are: Can we abstract from these musical concepts and thus avoid the modelling problems? Will two music tracks, represented by visually similar spectrograms, be classified to the same music genre?

Here, we present a novel approach to music genre classification. Having represented music tracks in the form of two dimensional images, we apply the “Bag of Visual Words” method from visual IR in order to classify the songs into 19 genres. By switching to the visual domain, we can

abstract from musical concepts such as melody, timbre and rhythm. We obtained classification accuracy of 46% (with 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.

The motivation behind this work was the hypothesis that 2D images of music tracks (spectrograms) perceived as similar would correspond to the same music genres. Conversely, it is possible to treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This points to an interesting interchangeability between visual and music information retrieval.

We apply our novel local features to the music genre classification task. Hence, we mention a few works that represent current state-of-the-art in this research area.

Almost every representation of music utilized in the field of Music Information Retrieval (MIR) involves extracting features from music transformed into the frequency domain. These features include chromatic, melodic, harmonic, rhythmic, and timbral measures.

Thus, (Suyoto, Uitdenbogerd & Scholer 2008) represents the distribution of chroma within a song as a histogram. The songs with similar chroma histogram distributions are considered similar. The temporal aspects of pitch are taken into account by (Hu, Dannenberg & Tzanetakis 2003) and (Collins 2005). In (Foote & Cooper 2001) researchers try to capture the rhythm by constructing a self-similarity matrix based upon the similarity of each short time frequency spectra extracted from the audio. (Bello, Daudet, Abdallah, Duxbury, Davies & Sandler 2005) presents a method to describe a novelty function (a common method for identifying onsets) of a waveform inspired by (Foote & Cooper 2001) similarity measure.

One of the challenges in MIR is how to interpret the classification confusions. Some incorrectly classified instances cast doubt on whether the ground truth is correct (for example a pop song that could be labelled as funk). The solution to this problem might be the incorporation of fuzzy logic.

Let us give a detailed description of the proposed model. The framework establishes the link between MIR and VIR research areas.

Our algorithm consists of the following stages:

1. **Transformation to the frequency domain:** Transform the music data from the time to frequency domain using Fast Fourier Transform (FFT). Since audio signals are periodic over time, it is convenient to represent them as a sum of an infinite number of sinusoidal waves. It makes it easier to analyse sinusoidal functions than general shaped functions.

2. **Music representation, visual data generation:** Generate spectrograms in two dimensional space, where the geometric dimensions represent frequency and time, and the colour of each point in the image indicates the amplitude of a particular frequency at a particular time. These spectrograms are generated from the signal transformed by FFT. Our method of

spectrograms generation was designed in such a way as to produce images containing easy to capture visual properties.

3. **Image sampling:**

   **Keypoints detection:** Apply Shi and Tomasi method (see (Shi & Tomasi 1994)) to find the points of interest.

   **Random sampling:** Apply random points generator to produce another half of the sample points.

   **Random sampling:** Alternatively, dense sample images (divide images into a number of uniform, non-overlapping rectangular sub-images).

4. **Description of local patches:** Characterize local patches in the form of a co-occurrence matrix or colour moments. In the case of a co-occurrence matrix extract the meaningful statistics - energy, entropy, contrast, and homogeneity. Compute the features for individual colour channels.

5. **Feature vector construction:** Represent local patterns as colour moments or statistics extracted from a co-occurrence matrix.

6. **Visual dictionary generation:** Apply K-means clustering to the training set in order to obtain the codebook of visual words.

7. **Histogram computation:** Create a histogram of visual words counts by calculating the distance between image patches and cluster centroids.

8. **Music genre classification:** The classification of music data into music genres is performed by the k-nearest neighbour algorithm, based on Minkowski's fractional similarity measure.

Steps 3 to 7 are related to the generation of visual representations of the spectrograms. In the course of this research, global methods like colour moments, co-occurrence matrix (texture), colour correlograms were also tested. We utilize local features because of their superior performance over various global methods. The local features may also have another advantage over other models. Interesting future work would be to investigate if image patches identified by corner detectors (roughly speaking - locations of a sudden change of pixel intensities) and "visual words" which correspond to some important characteristics of audio signal.

**The Fast Fourier Transform**

Let $x_0, ..., x_{N-1}$ be complex numbers. The Discrete Fourier Transform (DFT) is defined by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, ..., N-1. \tag{8.1}$$

Computing DFT requires $O(N^2)$ operations, while FFT reduces the number to $O(N \log N)$. The implemented FFT method incorporates the Cooley-Tukey algorithm, which breaks down a DFT into smaller DFTs. The audio sampled size is 65536 bytes (1.486 seconds), with sampling rate of 44100Hz.

**Spectrogram Generation**

The resulting spectrum is split into 512 bins (64Hz / bin). The power of each bin is converted into a pixel as follows:

```
Let colour = power / meanPower

IF {colour > 1}
   r = ((1 - (1 / colour)) / 2) + 0.5f
ELSE
   IF {colour > 0.5}
      g = (((colour-0.5)/0.5) / 2) + 0.5f
ELSE
   b = ((colour / 0.5) / 2) + 0.5f

   ENDIF
ENDIF
```

The horizontal dimension of each image represents the time (1 pixel = 1.486sec), the vertical dimension represents frequency (1 pixel = 64Hz), and pixel intensities represent power.

This method of spectrogram generation produces images that vary in colour and texture (see Figure 8.1). These properties make the images suitable for the application of visual features. An interesting observation is that some genres are easily recognizable directly from our spectrograms. Classical music, for instance, is characterized by the presence of the blue colour joining the top and the bottom part of an image. Figure 8.2 shows the query by visual example retrieval based on the local features and our music representation.

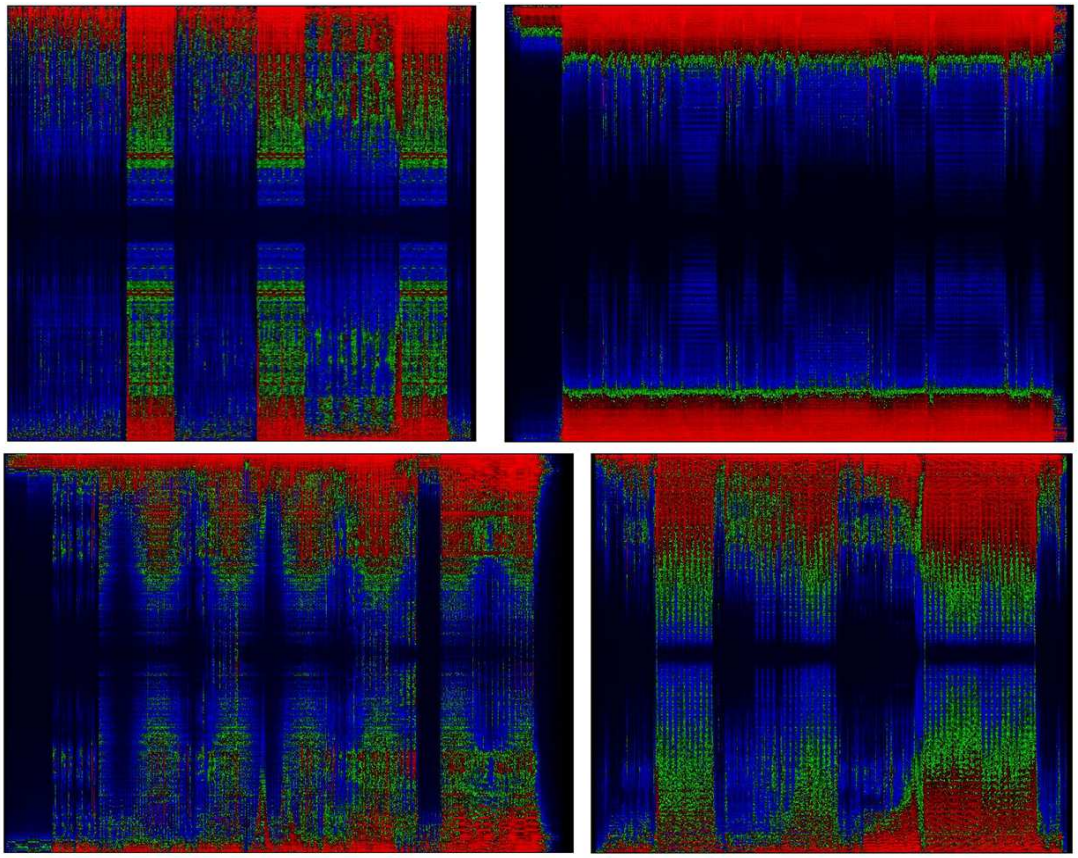

Figure 8.1: Music representation in the form of 2D images

**The Sampling Technique**

The implemented hybrid sampling method combines Shi and Tomasi corner detection (Shi & Tomasi 1994) with a random number generator. The Shi and Tomasi method is based on the Harris corner detector.

Another sampling technique implemented for comparison purposes was dense sampling. In this case, each image was divided into the same number of 900 identical rectangular sub-images.

**Region Descriptors**

Each local patch in an image was represented as

- The 8 orientation co-occurrence matrix.
- Colour moments.

A simple co-occurrence matrix is defined as follows:

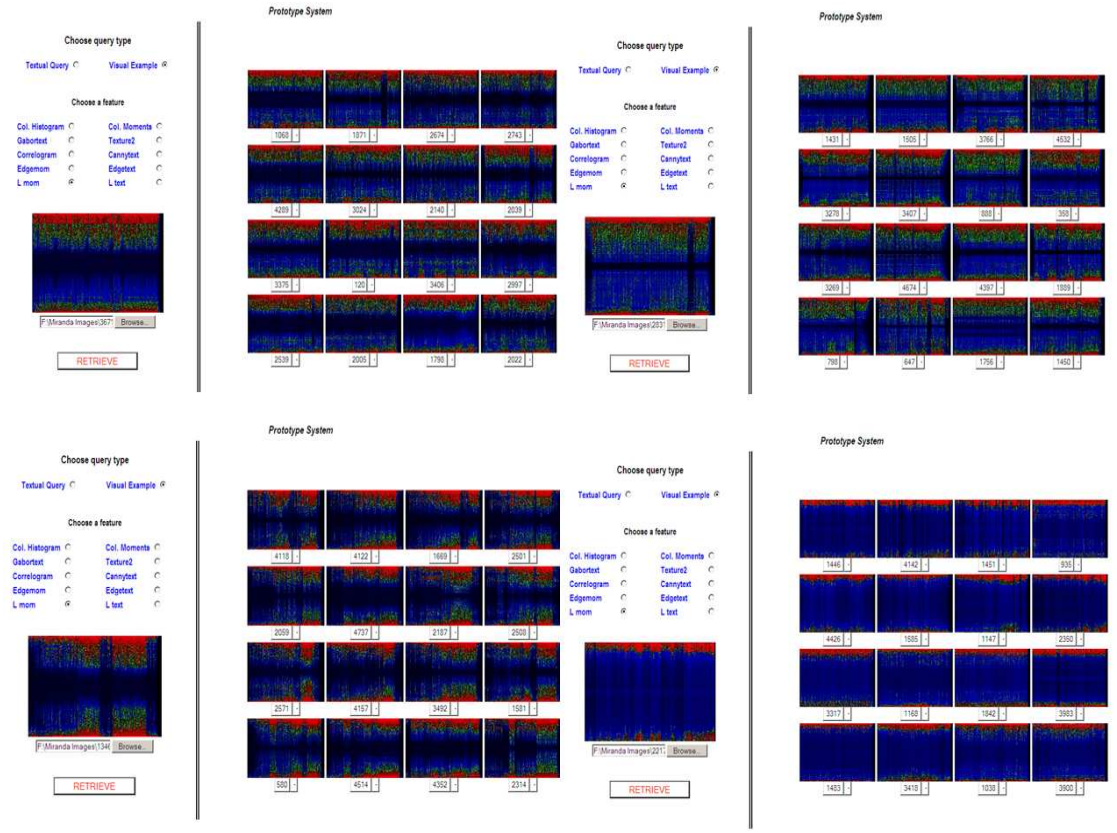


Figure 8.2: Local features at work. Four visual examples and spectrograms retrieved. It would be interesting to investigate if the retrieved music tracks are similar to the visual example (not only belong to the same music genre)

$$C_{\Delta x,\Delta y}(i,j) =$$

$$\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases} 1, & if \quad I(p,q)=i \quad and \quad I(p+\Delta x, q+\Delta y)=j \\ 0, & otherwise \end{cases} \tag{8.2}$$

The matrix describes the way certain grayscale pixel intensities occur in relation to other grayscale pixel intensities. It counts the number of such patterns. The most discriminating statistics extracted from the co-occurrence matrix are: contrast, inverse difference moment, entropy, energy, homogeneity, and variance.

The method based on three colour moments assumes that the distribution of colour can be treated as a probability distribution. Three statistics extracted from individual colour channels are

- Mean $E_i = \sum_{j=1}^{n} \frac{1}{N} p_{ij}$
- Standard Deviation $\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij}-E_i)^2\right)}$

- Skewness $s_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^3\right)}$

The first moment can be interpreted as an average colour value, the second as a square root of the variance of the distribution, and the third as the measure of asymmetry in the distribution. One can construct the weighted similarity measure as an analogy to the Manhattan metric, for example:

$$s(H, I) = \sum_{i=1}^{r} w_{i1}\left|E_i^1 - E_i^2\right| + w_{i2}\left|\sigma_i^1 - \sigma_i^2\right| + w_{i3}\left|s_i^1 - s_i^2\right|. \tag{8.3}$$

Colour moments can also capture the textural properties of an image and are fairly insensitive to viewpoint changes. By computing them in HSV colour space we can make the statistics insensitive to illumination changes.

**Feature Vector Construction, Visual Dictionary Generation, and Histogram Computation**

The local patches are represented as multidimensional vectors constructed from different statistics, extracted from individual colour channels. By taking a sample training set consisting of the collection's representative images, we can generate the so-called visual vocabulary. The K-means clustering algorithm has been used for that purpose. Each cluster characterizes a local pattern, representing a specific "visual word". The histogram of visual word counts is created by computing the Manhattan distance between individual patches and cluster centroids, and calculating how many patches belong to specific clusters.

**Music Genre Classification**

The classification of music data into music genres is performed by the k-nearest neighbour algorithm, using Minkowski's fractional similarity measure

$$s(x, y) = \left(\sum_{i=1}^{n}\sqrt{|x_i - y_i|}\right)^2 \tag{8.4}$$

where $x = (x_i)$ and $y = (y_i)$ are the $n$ dimensional feature vectors.

The main advantages of our method are: more intuitive, easy way to automatic music classification, classification accuracy comparable with state-of-the-art and new promising research direction.

### 8.1.1 Enhancing Music Information Retrieval, Experimental Results

Next, we evaluate our novel approach to music genre classification.

### Experimental Setup

For experimental purposes we used a data collection consisting of 4759 music tracks. The genre distribution is presented in Table 8.1.

Table 8.1: Genre classes

| Genre | Tr. | Genre | Tr. |
|---|---|---|---|
| Pop | 1024 | Country | 82 |
| Alternative and Punk | 919 | Hip-Hop | 81 |
| Rock | 862 | Reggae | 80 |
| R&B | 516 | Easy Listening | 80 |
| Classical | 293 | Musicals | 75 |
| Dance | 265 | Latin | 62 |
| Alternative | 139 | Christmas | 42 |
| Folk | 115 | Rap | 15 |
| Metal | 89 | Soundtrack | 11 |
| Blues | 9 | | |

Genre labels were extracted from iTunes. The local feature algorithm uses 900 sample points per image, for each sample point we open a square window 10 by 10 pixels wide. The dimensionality of the histograms of visual words counts is 40. The applied k-nearest neighbour algorithm uses 9-fold cross-validation, 12 nearest neighbours, distance weighting and Manhattan metric.

### Results and Discussion

The classification accuracy we obtained with dense sampling was approximately **46 per cent** (2176 tracks) of correctly classified instances. The hybrid sampling scored lower, resulting in **43 per cent** (2051 tracks) of retrieval accuracy. The reason for this lies in the worse performance of corner detector in this domain. The local features with hybrid sampling performed better than the one with dense sampling on the ImageCLEF2007 and MIRFlickr25000 collections which consist of real-life images.

From the confusion analysis we observed that most incorrectly classified instances were confused with similar genres, and the song-genre correspondence was arguable and subjective. That is why it is so hard to improve the retrieval performance. Good, natural solution to this problem could be the incorporation of fuzzy logic, and associate each song with certain probability of it being in one of the classes. The problem with comparisons with other methods arises because of the lack of standardized data collections in MIR. Many data collections have unequally distributed data sets, different number of genres, more specialized or generalized classes.

All of that affects the behaviour of a classifier. (Meng, Ahrendt, Larsen & Hansen 2007) used a multivariate autoregressive feature model, considered as current state-of-the-art in MIR, to capture the temporal information in the window. The data set used consisted of 1210 music tracks

with 11 genres. The best mean classification accuracy they obtained were 44% and 40% for the LM and GLM classifiers. It should be noted though that the accuracies obtained by the automatic classification need to be relative to the theoretical baseline for random classification which is 9% for (Meng et al. 2007), and 5% for our collection. It means that the performance of our method is actually much better. There are also other aspects, mentioned previously, that make the evaluation difficult.

In his PhD thesis on music genre classification, Serra presents a "non exhaustive list for the most relevant papers presented in journals and conferences for the last years" (Serra 2009). He concludes that "although accuracies are not completely comparable due to the different datasets the authors use, similar approaches have similar results. This suggest that music genre classification, as it is known today, seems to reach a "glass ceiling"". The reported accuracies were then plotted with respect to the number of genres (Figure 8.3; adapted from (Serra 2009)).

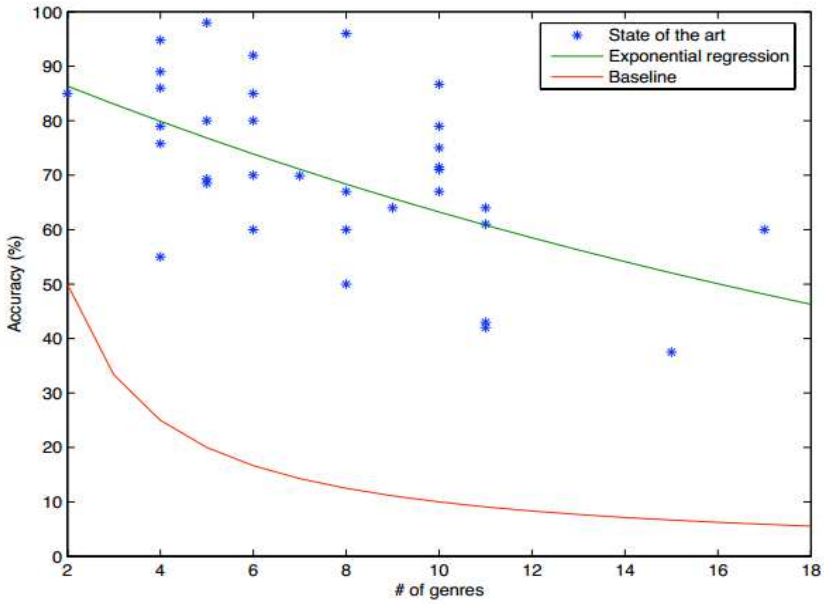


Figure 8.3: State of the art in music genre classification

The human performance in classifying music genres (10 genres) is around 53% correctly classified for 250ms samples and around 70% for samples longer than 3s (Perrot & Gjerdigen 1999). Thus, the performance of current state-of-the-art models for genre classification is comparable with human performance.

**Conclusion and Future Work**

In this chapter we evaluated a novel approach to MIR. Having represented the music tracks in the form of two dimensional images, we apply the "Bag of Visual Words" method from visual IR

in order to classify the songs into 19 genres. The motivation behind this work was the hypothesis, that 2D images of music tracks (spectrograms) perceived as similar would correspond to the same music genres (perhaps even similar music tracks). Conversely, we can treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This would point to an interesting interchangeability between visual and music information retrieval.

We obtained classification accuracy of 46% (with a 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.

The main advantages of our method are: more intuitive, an easy way to automate music classification, classification accuracy comparable with state-of-the-art and a new promising research direction.

The future work may include incorporation of the spatial information for local image patches, experimentation with different sampling techniques and incorporation of temporal information (short time Fourier transform, wavelets), which should further improve the classification accuracy. Additionally, interesting future work would be to investigate whether image patches identified by corner detectors and "visual words" correspond to some important characteristics of the audio signal. In other words, new specialized visual features can be developed for this particular task.

## 8.2 Food Recognition

The food recognition task is an example of the domain specific application of visual features. It can be utilized to build a mobile phone application software for restaurants and users concerned about their nutrition (sports diet, etc.). For example, a user (restaurant's customer) would take a picture of a concrete dish served in a restaurant with his/her mobile phone. This picture would be sent to the server where the retrieval system would search for a similar image in the database based on the visual content. The images in the database can have descriptions related to the nutritional value of this particular dish. Thus, the user can quickly obtain the desired information about the food that he/she is going to eat (e.g. number of calories). We have prepared a food recognition prototype system based on our local visual features. Our industrial partner, a chain of restaurants interested in this task, provided a data collection representing a range of their products (food selection). Figures 8.4, 8.5, 8.6, and 8.7 show examples of the prototype system at work.

## 8.3 Chapter Summary

This chapter presents two applications of our novel local features to music genre classification and the food recognition task.

The food recognition task is an example of the domain specific application of visual features. It can be utilized to build a mobile phone application software for restaurants and users (customers)

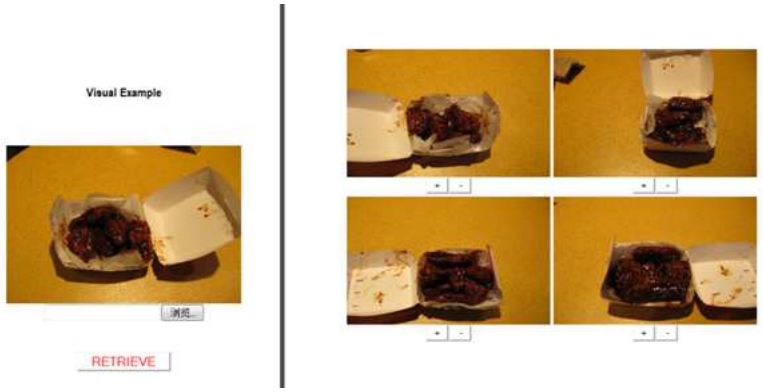


Figure 8.4: Food Recognition System. Left - example image (i.e. taken by a customer), right - images retrieved from the database

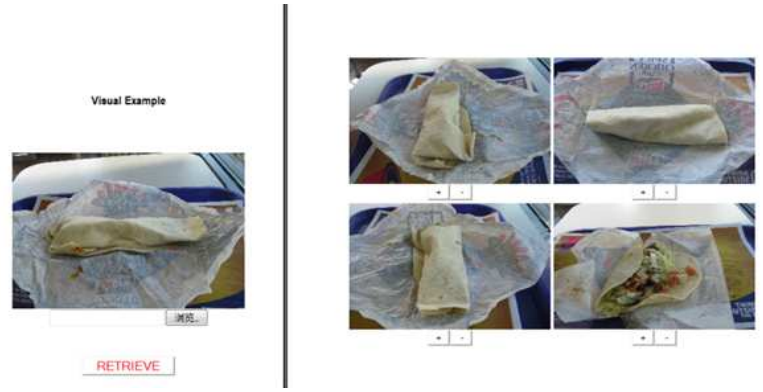


Figure 8.5: Food Recognition System. Left - example image (i.e. taken by a customer), right - images retrieved from the database


Figure 8.6: Food Recognition System. Left - example image (i.e. taken by a customer), right - images retrieved from the database

concerned about their nutrition (sports diet, etc.). For example, a user would take a picture of a concrete dish served in a restaurant with his/her mobile phone. This picture would be sent to the server where the retrieval system would search for a similar image in the database based on the


Figure 8.7: Food Recognition System. Left - example image (i.e. taken by a customer), right - images retrieved from the database

visual content. The images in the database can have descriptions related to the nutritional value of this particular dish. Thus, the user can quickly obtain the desired information about the food that he/she is going to eat.

For the food recognition task we developed a working demo that shows the promising performance of our visual features.

Music genre classification is based on the audio content features extracted from the audio signal. These features are often hard to extract and define, and represent such musical aspects as melody, timbre and rhythm, to name a few.

We present an alternative approach to music genre classification, and classify music based on the visual similarities of music spectrograms generated in a colour space. Thus, having represented music tracks in the form of two dimensional images, we apply the "Bag of Visual Words" method from visual IR in order to classify the songs into 19 genres. By switching to the visual domain, we can abstract from musical concepts such as melody, timbre and rhythm. We obtained classification accuracy of 46% (with a 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.

The motivation behind this work was the hypothesis that 2D images of music tracks (spectrograms) perceived as similar would correspond to the same music genres. Conversely, it is possible to treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This points to an interesting interchangeability between visual and music information retrieval.

# Chapter 9

# Conclusions and Future Work

In this thesis, we have proposed a principled framework for the semantic gap reduction in large scale generic image retrieval. The proposed framework comprises development and enhancement of novel visual features, a hybrid model for visual and textual feature combination, and a hybrid model for the combination of features in the context of relevance feedback, with both fixed and adaptive weighting schemes (importance of a query and its context). Apart from the experimental evaluation of our models, theoretical validations of some interesting discoveries on feature fusion strategies were also performed. The proposed models were incorporated into our prototype demo system with an interactive user interface.

We shed more light on the existing hybrid and relevance feedback models. We prove that the interchangeability of specific fusion strategies (also suggested by other researchers through reported experimental results) result from the interaction between similarity measurements and early fusion operators. Both complex early fusion and relevance feedback models (query modification) can be represented as a late fusion strategy.

## 9.1 Contributions

Here, we present a list of our contributions, in relation to addressing the research questions and the gaps that we have identified in the literature survey.

All the identified gaps and posed research questions are important to the hybrid modelling problem, which we choose as our tool for semantic gap reduction. Thus, the key objectives (hypotheses) of this thesis, that were evaluated and theoretically verified, are:

1. experimental evaluation of novel visual features for generic large-scale CBIR and enhancement of local visual features;

2. experimental evaluation of hybrid model for the combination of visual and textual features;

3. experimental evaluation of hybrid model for the combination of visual and textual features in the context of relevance feedback;

4. experimental evaluation of adaptation of weights associated with the importance of a query and its context (feedback images) in our hybrid relevance feedback model;

5. theoretical verification of the hypothesis on the interchangeability of specific fusion strategies;

6. theoretical verification of the hypothesis on the representation of query modification in a late fusion form;

Particularly:

### 9.1.1 Novel Visual Features (Objective 1)

We introduce novel global methods based on our novel edge detector, bilateral filtering, directional derivatives, and pixel intensities thresholding.

We propose a novel method based on the local features, incorporating an easy to implement descriptor and a hybrid sampling technique. We also compare different sampling methods: hybrid (a combination of random and detector-based sampling), purely random, purely detector-based and dense on three large data collections. The hybrid sampling produces more discriminative of image patches than the commonly used detector-based method. The proposed descriptor is easy to implement and produces low dimensional feature vectors which, in turn, reduces the computational and data storage cost. Empirical evaluation has been performed on three large image collections, namely ImageCLEF 2007, MIRFlickr 25000 and BGS datasets.

Our approach is easy to implement, not sophisticated, with low computational and data storage cost (mostly because our vectors are low dimensional), and the hybrid sampling technique can be used in other methods based on the "bag of visual words" to improve the retrieval performance. Moreover, the evaluation of the proposed method has been conducted on three different large data collections without changing the initial setup. In this way we avoided "fine-tuning" of the parameters to the specific data collection which makes the results more general and reliable.

### 9.1.2 Enhancement of Local Features (Objective 1)

We propose a new approach for identifying and utilizing the information about correlations between visual words. We implement and test various notions of correlation at different contextual levels (we refer to them as image-level and proximity based). To the best of our knowledge, this is the first time these two were compared within this type of framework in image retrieval. Our local features consist of low dimensional histograms, where bins representing visual words are highly correlated. We identify the most and the least correlated coefficients and use thus obtained information, along with the visual terms' frequencies from the current query, to weight the similarity measure. Certain coefficients in the similarity measure corresponding to the most correlated terms are then increased, while the coefficients related to the least correlated pairs are deemphasized. The evaluation was performed on three large data collections, namely ImageCLEF 2007,

MIRFlickr 25000 and BGS. The evaluation was performed within the Pseudo Relevance Feedback framework.

Experimental results show the superiority of two notions of correlation, which are image level correlations. For these two correlations, we report significant improvement in terms of Mean Average Precision on two data collections within PRF evaluation framework. Moreover, the addition of information about the least correlated visual words often further improves the performance. The proximity based notion of correlation does not show a significant improvement in the context of this model.

The proposed method is computationally and data storage cheap, utilizes correlation at different contextual levels, and avoids the normalization of histograms.

### 9.1.3 Novel Hybrid Model (Objective 2)

We propose a quantum theory inspired multimedia retrieval framework based on the tensor product of feature spaces, where similarity measurement between a query and a document corresponds to the quantum measurement. At the same time, the correlations between dimensions across different feature spaces can also be naturally incorporated in the framework. The tensor based model provides a formal and flexible way to expand the feature spaces, and seamlessly integrate different features, potentially enabling multi-modal and cross media search in a principled and unified framework.

Experimental results on a standard multimedia benchmarking collection show that the quantum-like measurement on a tensored space leads to remarkable performance improvements in terms of average precision over the use of individual feature spaces separately or concatenation of them.

### 9.1.4 Novel Hybrid Model and Image Auto-Annotation (Objective 2)

We introduce and test two novel approaches for making associations between tags and images. We also experiment with mid-level semantic image representations based on the "bag of visual words" model. This was a follow-up work on our tensor-based unified image retrieval framework. In order to prepare the data for the quantum-like measurement in the tensor space, we need to alleviate the problem regarding the unannotated images. The first proposed approach projects the unannotated images onto the subspaces generated by subsets of training images (containing given textual terms). We calculate the probability of an image being generated by the contextual factors related to the same topic. In this way, we should be able to capture the visual contextual properties of images, taking advantage of this extended vector space model framework. The other method performs quantum-like measurement on the density matrix of unannotated image, with respect to the density matrix representing the probability distribution obtained from the subset of training images. These approaches can be seamlessly integrated into our unified framework for image retrieval.

### 9.1.5 Novel Hybrid Relevance Feedback Model (Objective 3)

We also extend our hybrid model for visual and textual feature combination within the context of relevance feedback. The approach is based on mathematical tools also used in quantum mechanics - the predicted mean value of the measurement and the tensor product of the density matrices, which represents a density matrix of the combined systems. It was designed to capture both intra-relationships between features' dimensions (visual and textual correlation matrices) and inter-relationships between visual and textual representations (tensor product). The model provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context by applying quantum-like measurement. It opens a door for a series of theoretically well-founded further exploration routes, e.g. by considering the interference among different features. It is easily scalable to large data collections as it is general and computationally cheap. The results of the experiment conducted on ImageCLEF data collection show the significant improvement over other baselines.

### 9.1.6 Dynamic Weighting of Query and its Context (Objective 4)

The aforementioned model for the visual and textual features combination utilizes fixed weights corresponding to the importance of the query and its context. However, the query can be more or less related to its context. Inspired by this observation, we incorporate adaptive weighting scheme into the hybrid CBIR relevance feedback model. Thus, each query is associated with a unique set of weights corresponding to the relationship strength between a visual query and its visual context as well as the textual query and its textual context. The higher the number of terms or visual terms (mid-level features) co-occurring between current query and the context, the stronger the relationship and vice versa. If the relationship between query and its context is weak, context becomes important. We adjust the probability of the original query terms, the adjustment will significantly modify the original query. If the aforementioned relationship (similarity) between query and its context is strong, however, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

We tested the enhanced model within the user simulation framework. For fair comparison purposes, the best performing sets of fixed weights were selected. We have shown that our enhanced model with adaptive weighting scheme can outperform the original one with fixed weights. Moreover, an addition of another visual feature (colour histogram, global feature) further improved both the hybrid CBIR relevance feedback model and the enhanced model's performance, whereas the performance of the baselines did not change much.

Our contribution here is related to showing how to measure the relationship strength between query and its context, and how to incorporate the adaptive weighting scheme into the state-of-the-art existing model (hybrid, user feedback context) to further improve the retrieval. The proposed

adaptive weighting approach is relatively easy to implement and does not require any training of features.

### 9.1.7 Duality of Fusion Strategies (Objective 5)

In this work, we theoretically investigate some interesting interactions between common similarity measurements and common operators related to early fusion strategy (e.g. concatenation of vector representations). We show that these interactions between certain similarity measurements and early fusion strategies result in combinations of representations at the decision level (late fusion strategy). In other words, we theoretically prove that certain combinations of early fusion strategies and certain similarity measurements are equivalent to particular combinations of measurements (i.e. relevance scores) computed on individual feature spaces.

Our findings are important from both theoretical and practical perspectives. First, we should be careful when comparing early and late fusion strategies as they may represent equivalent approaches. Second, specific combinations of individual relevance scores can have a sound theoretical interpretation which would be easier to analyze as an early fusion. Finally, knowing how to represent early fusion as a late one can help us avoid the curse of dimensionality.

Thus, an interesting open question arises - the most important, in our opinion, consequence of our observations. Does late fusion strategy, contrary to current belief, capture the relationships between feature spaces or does the interaction between the similarity measurement and early fusion operators decorrelate features? This question should make us look at feature combination methods from different perspective.

### 9.1.8 Query Modification as a Late Fusion (Objective 6)

Query modification strategies utilize relevance feedback to modify the query in order to narrow down the search. We show that query modification can be represented as a late fusion strategy. This observation has a few important implications. First, some complex combinations of measurements performed on individual feature spaces may be regarded as a query modification technique. Second, query modification represented as a late fusion does not require an actual modification of query representation. The actual query modification would often lead to query renormalization. While renormalization in text IR is often desirable, renormalization of mid-level visual representations may even hamper the retrieval performance. Third, it is often easier to implement and work with relevance scores. Finally, knowing how to represent query modification as a late fusion can help us develop a family of hybrid relevance models (combinations of scores computed on lower dimensional feature spaces as opposed to high dimensional hybrid representations).

### 9.1.9 Prototype Hybrid System with Interactive User Interface

We present an interactive user interface which has been synchronized with our prototype system. The user interface is the communication platform between the system and the user. In order to

show a working demo of the prototype system, we have integrated it with the interactive user interface. Thus, we present a unified working framework (demo of our prototype system) for content-based image retrieval comprising: novel visual features for generic image retrieval and their combinations with existing visual features (if selected), combination of visual and textual features in the first round retrieval, combination of visual and textual features in the context of relevance feedback (search refinement), interactive user interface. To the best of our knowledge, this is the first prototype system that utilizes hybrid model for combination of visual and textual features as well as a hybrid relevance feedback model, and allows for a full interaction with the system.

### 9.1.10 Other Contributions

Other contributions include a proposal of a family of hybrid models for visual and textual features combination, as well as a family of hybrid relevance feedback models. We also provide a theoretical model for the combination of visual and textual features in the context of user feedback with various degrees of relevance (discrete and continuous). Thus, a user would drag and drop feedback images in a natural way onto a relevance bar. We also introduce a number of variations of the experimentally tested models.

#### 9.1.10.1 Application of Novel Local Features to Music Genre Classification

We introduce a novel approach to MIR. Having represented the music tracks in the form of two dimensional images, we apply the "bag of visual words" method from visual IR in order to classify the songs into 19 genres. The motivation behind this work was the hypothesis that 2D images of music tracks (spectrograms) perceived as similar would correspond to the same music genres (perhaps even similar music tracks). Conversely, we can treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This would point to an interesting interchangeability between visual and music information retrieval.

We obtained classification accuracy of 46% (with a 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.

The main advantages of our method are: it is a more intuitive, easy way to automatic music classification, has classification accuracy comparable with state-of-the-art and is a promising new research direction.

Some of the work presented in this thesis has been published and presented at various conferences. For the list of publications see the "Published Papers" appendix.

## 9.2 Future Work

Here, we present a list of potential future research. We hope that the research propositions will inspire the reader to follow our work.

### 9.2.1 Visual Features

We are planning to extend our evaluation of different sampling techniques to other various detectors and descriptors such as Scale Invariant Feature Transform. We will extend the experiments to different numbers of sample points. The influence of the size of the codebook on the retrieval performance will also be measured. We believe that the optimal number of "visual words" may depend on the sort of detector and descriptor used and their representative ability. To the best of our knowledge, the existing evaluation assumes that the size of the codebook is independent of the descriptors and detectors or tests the influence on one specific method. This is an open issue that remains to be tested experimentally.

### 9.2.2 Hybrid Tensor-based Model

In our experiments, we assumed that all dimensions are orthogonal to each other. This assumption, however, can be relaxed. We can either replace the synonyms with one representative word, or apply dimensionality reduction techniques in our tensor model. This is also sensible from a practical point of view, as high dimensional textual feature space can make the computation of the ranking scores difficult on a large scale collection. Moreover, we would like to test the tensor product model on a wide selection of visual content features.

### 9.2.3 Image Auto-Annotation

The correlations at the image-level may be stronger that the correlations based on the proximity between visual terms (pixels, local patches). Recent works build the correlations based on the proximity between image patches to capture the spatial information, as researchers believe that the relative distance between them is important. Thus, we need to test this alternative method for correlation matrix generation and perform large scale experiments.

### 9.2.4 Hybrid Relevance Feedback Model

The future work will involve testing different notions of correlation within the proposed framework (we can construct correlation matrices in such a way that they can be regarded as density matrices). In this experiment, we incorporate document/image level correlations only. However, in the case of textual representations, we can also experiment with Hyperspace Analogue to Language (HAL). In the this approach, the context is represented by a sliding window of a fixed size (while in document level correlation the context is represented by the whole document). We can also consider a visual counterpart to HAL, where a window of a fixed size (e.g. square, circular) is shifted from one

instance of a visual word to another. Then, the number of instances of visual words that appear in the proximity of the visual word on which the window is centered can be calculated. In case of a dense sampling, the window would be shifted analogously to HAL in text IR. If the sparse sampling was utilized, however, the window would shift from one instance of a visual word to another.

The future work may involve an adaptation of weights associated with the importance of visual and textual representations in the context of user feedback. Neither the original or our enhanced models allow for adaptive weighting of visual and textual features. However, it could be advantageous to be able to infer from the user feedback, the subjective importance of the visual and textual parts of the system.

The original model (represented in the following form, not the end product) can be interpreted as a fusion-like strategy similar to combPROD late fusion

$$\left\langle M_1 \middle| a^T \cdot a \right\rangle \left\langle M_2 \middle| b^T \cdot b \right\rangle \tag{9.1}$$

However, the weighted linear combination with adaptive weights corresponding to the importance of visual and textual features can have an even better capability for performance improvement (model personalization). Thus, we can treat our model as a kind of a weighted linear combination

$$\begin{aligned} r_v \left\langle M_1 \middle| a^T \cdot a \right\rangle + r_t \left\langle M_2 \middle| b^T \cdot b \right\rangle = \\ r_v \left( str_v \left\langle q_v | a \right\rangle^2 + (1 - str_v) \frac{1}{n} \sum_i \left\langle c^i \middle| a \right\rangle^2 \right) + \\ r_t \left( str_t \left\langle q_t | b \right\rangle^2 + (1 - str_t) \frac{1}{n} \sum_i \left\langle d^i \middle| b \right\rangle^2 \right) \end{aligned} \tag{9.2}$$

$$str_v = \frac{\left\langle D_q^v \middle| D_f^v \right\rangle}{\left\| D_q^v \right\| \left\| D_f^v \right\|} \tag{9.3}$$

$$str_t = \frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\| D_q^t \right\| \left\| D_f^t \right\|} \tag{9.4}$$

where the weights corresponding to the importance of visual and textual features, $r_v$ and $r_t$, could be somehow automatically modified depending on the inferred user preferences and interests.

### 9.2.5 Other Applications of Novel Visual Features. Music Genre Classification

Our future work may include incorporation of the spatial information for local image patches, experimentation with different sampling techniques and incorporation of temporal information (short time Fourier transform, wavelets), which should further improve the classification accuracy. Additionally, interesting future work would be to investigate whether image patches identified by

corner detectors and "visual words" correspond to some important characteristics of the audio signal. In other words, new specialized visual features can be developed for this particular task.

### 9.2.6 On the Duality of Fusion Strategies and Query Modification as a Combination of Relevance Scores

For future work we plan to search for other combinations of various operators and similarity measures that could interact in such a way as to represent late fusion. We will also try to answer the question regarding the late fusion and the relationships between the feature spaces: does the interchangeability between specific fusion schemes mean that the late fusion can capture the relationships between the features (contrary to current belief) or do the effects of particular combinations of similarity measures, notions of correlation and operators neutralize the relationship properties?

In addition, the observations on the duality of fusion strategies may have other interesting applications. In many research fields the concatenated or tensored vectors need to be clustered

$$\min \sum_{i=1}^{k} \sum_{x_j \otimes y_j \in s_i} d_e^2(x_j \otimes y_j, \frac{1}{\left|s_i^{(t)}\right|} \sum_{x_i \otimes y_i \in s_i^{(t)}} x_i \otimes y_i) \tag{9.5}$$

$$\min \sum_{i=1}^{k} \sum_{x_j \oplus y_j \in s_i} d_e^2(x_j \oplus y_j, \frac{1}{\left|s_i^{(t)}\right|} \sum_{x_i \oplus y_i \in s_i^{(t)}} x_i \oplus y_i) \tag{9.6}$$

The squared distances (often Euclidean) between the given tensored or concatenated vectors and the means of tensored or concatenated vectors can also be represented as combinations of distances computed on individual vector spaces. Therefore, there is no need to work with high dimensional vectors and the particular combination of distances computed on individual vector spaces will produce the same effect as clustering of concatenated or tensored vectors.

# Bibliography


Aharonov, Y., Albert, D. Z. & Au, C. (1981). New interpretation of the scalar product in hilbert space, *Physical Review Letters* **47**(15): 1029–1031.

Albanese, M., Chianese, A., D'Acierno, A., Moscato, V. & Picariello, A. (2010). A multimedia recommender integrating object features and user behavior, *Multimedia Tools Appl.* **50**(3): 563–585.

ao Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S. & de Albuquerque Araújo, A. (2009). A bag-of-features approach based on hue-sift descriptor for nude detection, *Proceedings of the XVII European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland.

Athanasakos, K., Stathopoulos, V. & Jose, J. (2010). A framework for evaluating automatic image annotation algorithms, *ECIR*, 5993, pp. 217–228.

Author, N. (n.d.a). Imageclef website. `www.imageclef.org`.

Author, N. (n.d.b). Mirflickr website. `http://press.liacs.nl/mirflickr/`.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. & Sandler, M. B. (2005). A tutorial on onset detection in music signals, *Speech and Audio Processing, IEEE Transactions on* **13**(5): 1035–1047.

Berger, H., Denk, M., Dittenbach, M., Pesenhofer, A. & Merkl, D. (2007). Photo-based user profiling for tourism recommender systems, *Proceedings of the 8th International Conference on E-commerce and Web Technologies*, EC-Web'07, Springer-Verlag, pp. 46–55.

Biancalana, C., Lapolla, A. & Micarelli, A. (2009). Personalized web search using correlation matrix for query expansion, *Web Information Systems and Technologies* pp. 186–198.

Bui, T. H., Lenz, R. & Kruse, B. (2005). Abstract color and texture induced structures of keyword space, *Proceedings of the 10th Congress of the International Colour Association* . `http://webstaff.itn.liu.se/~reile/csp-pages/publications/reprints/aic05-319-2.pdf`.

Castellano, G., Fanelli, A., Mencar, C. & Torsello, M. (2007). Similarity-based fuzzy clustering for user profiling, *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences on*, pp. 75–78.

Celebi, E. & Alpkocak, A. (2000). Clustering of texture features for content-based image retrieval, *Proceedings of the First International Conference on Advances in Information Systems*, ADVIS '00, Springer-Verlag, pp. 216–225.

Chang, E. Y., Cheng, K. T., Lai, W. C., Wu, C. T., Chang, C. & Wu, Y. L. (2001). Pbir: Perception-based image retrieval - a system that can quickly capture subjective image query concepts, *ACM Multimedia*, pp. 611–614.

Chang, Y.-C. & Chen, H.-H. (2009). Increasing precision and diversity in photo retrieval by result fusion, *Proceedings of the 9th Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*, CLEF'08, Springer-Verlag, pp. 612–619.

Chen, Z., Wenyin, L., Zhang, F., Li, M. & Zhang, H. (2001). Web mining for web image retrieval, *Journal of the American Society for Information Science and Technology* **52**(10): 831–839.

Chu, W. & Park, S.-T. (2009). Personalized recommendation on dynamic content using predictive bilinear models, *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, ACM, pp. 691–700.

Clinchant, S., Ah-Pine, J. & Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval, *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, number 44 in *ICMR '11*, ACM, pp. 1–8.

Collins, N. (2005). Using a pitch detector for onset detection, *Proc. of ISMIR2005* pp. 100–106.

Collins, R. (2011). Lecture 06: Harris corner detector. `http://www.cse.psu.edu/~rcollins/CSE486/lecture06_6pp.pdf`.

de Ves, E., Ruedin, A., Acevedo, D., Benavent, X. & Seijas, L. (2007). A new wavelet-based texture descriptor for image retrieval, *Proceedings of the 12th international conference on Computer analysis of images and patterns*, CAIP'07, Springer-Verlag, pp. 895–902.

del Bimbo, A. (1999). *Visual Information Retrieval*, The Morgan Kaufmann Series in Multimedia Information and Systems Series, Morgan Kaufmann. `http://books.google.co.uk/books?id=NeoevqTFtikC`.

Depeursinge, A. & Muller, H. (2010). Fusion techniques for combining textual and visual information retrieval, *ImageCLEF*, Vol. 32 of *The Information Retrieval Series*, Springer Berlin Heidelberg, pp. 95–114.

Deselaers, T., Keysers, D. & Ney, H. (2005). Fire - flexible image retrieval engine: Imageclef 2004 evaluation, *Proceedings of the 5th Conference on Cross-Language Evaluation Forum: Multilingual Information Access for Text, Speech and Images*, CLEF'04, Springer-Verlag, pp. 688–698.

Di Buccio, E., Melucci, M. & Song, D. (2011). Towards predicting relevance using a quantum-like framework, *Advances in Information Retrieval* pp. 755–758.

Dobrescu, R., Dobrescu, M. & Ichim, L. (2006). Adding fractal dimension as textural feature for content based image retrieval, *Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*, SMO'06, World Scientific and Engineering Academy and Society (WSEAS), pp. 648–653.

Duygulu, P., Barnard, K., Freitas, J. F. G. d. & Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, Springer-Verlag, pp. 97–112.

Felden, C. & Linden, M. (2007). Ontology-based user profiling, *Proceedings of the 10th International Conference on Business Information Systems*, BIS'07, Springer-Verlag, pp. 314–327.

Foote, J. & Cooper, M. (2001). Visualizing musical structure and rhythm via self-similarity, *Proceedings of the 2001 International Computer Music Conference*, pp. 419–422.

Gotlieb, C. C. & Kreyszig, H. E. (1990). Texture descriptors based on co-occurrence matrices, *Computer Vision, Graphics, and Image Processing* **51**(1): 70–86.

Griffiths, R. B. (2003). *Consistent Quantum Theory*, Cambridge University Press.

Grubinger, M., Clough, P., Hanbury, A. & Müller, H. (2008). Overview of the imageclefphoto 2007 photographic retrieval task, *Advances in Multilingual and Multimodal Information Retrieval* pp. 433–444.

Heesch, D., Yavlinsky, A. & Rüger, S. (2003). Performance comparison of different similarity models for cbir with relevance feedback, *Proceedings of the 2nd International Conference on Image and Video Retrieval*, CIVR'03, Springer-Verlag, pp. 456–466.

Hu, N., Dannenberg, R. & Tzanetakis, G. (2003). Polyphonic audio matching and alignment for music retrieval, *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on*, pp. 185–188.

Huiskes, M. J. & Lew, M. S. (2008). The mir flickr retrieval evaluation, *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ACM, pp. 39–43.

Jacobs, K. (2011). Combining systems: The tensor product and partial trace. `www.quantum.umb.edu/Jacobs/QMT/QMT-AppendixA.pdf`.

Jamieson, M., Dickinson, S., Stevenson, S. & Wachsmuth, S. (2006). Using language to drive the perceptual grouping of local image features, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, IEEE Computer Society, pp. 2102–2109.

Jin, H., He, R. & Tao, W. (2008). Multi-relationship based relevance feedback scheme in web image retrieval, *International Journal of Innovative Computing, Information and Control* **4**(6): 1315–1324.

Kherfi, M. L., Ziou, D. & Bernardi, A. (2004). Image retrieval from the world wide web: Issues, techniques, and systems, *ACM Comput. Surv.* **36**(1): 35–67.

Landy, M. S. & Graham, N. (2004). Visual perception of texture, *The Visual Neurosciences*, MIT Press, pp. 1106–1118.

Li, J. & Wang, J. (2008). Real-time computerized annotation of pictures, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **30**(6): 985–1002.

Li, Y. & Cunningham, H. (2008). Geometric and quantum methods for information retrieval, **42**(2): 22–32.

Lindeberg, T. (1994). Scale-space theory in computer vision.

Lingenfelser, F., Wagner, J. & André, E. (2011). A systematic discussion of fusion techniques for multi-modal affect recognition tasks, *Proceedings of the 13th International Conference on Multimodal Interfaces*, ACM, pp. 19–26.

Liu, H. (2010). *A Framework for Understanding User Interaction with Content-based Image Retrieval: Model, Interface and Users*, PhD thesis, The Open University.

Liu, H., Song, D., Rüger, S., Hu, R. & Uren, V. (2008). Comparing dissimilarity measures for content-based image retrieval, *Information Retrieval Technology* pp. 44–50.

Liu, H., Zagorac, S., Uren, V., Song, D. & Rüger, S. (2009). Enabling effective user interactions in content-based image retrieval, *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, AIRS '09, Springer-Verlag, pp. 265–276.

Liu, T., Liu, J., Liu, Q. & Lu, H. (2009). Expanded bag of words representation for object classification, *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 297–300.

Lowe, D. (1999). Object recognition from local scale-invariant features, *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, Vol. 2, pp. 1150–1157.

Lu, Y., Zhang, H., Wenyin, L. & Hu, C. (2003). Joint semantics and feature based image retrieval using relevance feedback, *Trans. Multi.* **5**(3): 339–347.

Maillot, N., Chevallet, J.-P. & Lim, J. H. (2007). Inter-media pseudo-relevance feedback application to imageclef 2006 photo retrieval, *Proceedings of the 7th International Conference on Cross-Language Evaluation Forum: Evaluation of Multilingual and Multi-modal Information Retrieval*, CLEF'06, Springer-Verlag, pp. 735–738.

Melucci, M. (2005). Context modeling and discovery using vector space bases, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ACM, pp. 808–815.

Meng, A., Ahrendt, P., Larsen, J. & Hansen, L. K. (2007). Temporal feature integration for music genre classification, *Audio, Speech, and Language Processing, IEEE Transactions on* **15**(5): 1654–1664.

Mensink, T., Csurka, G., Perronnin, F., Sánchez, J. & Verbeek, J. J. (2010). Lear and xrce's participation to visual concept detection task - imageclef 2010, *CLEF (Notebook Papers/LABs/Workshops)*. `http://clef2010.org/resources/proceedings/clef2010labs_submission_65.pdf`.

Mensink, T., Verbeek, J. & Csurka, G. (2011). Weighted transmedia relevance feedback for image retrieval and auto-annotation, *Technical report*, INRIA. `http://hal.inria.fr/hal-00645608/PDF/RT-0415.pdf`.

Middleton, S. E., Shadbolt, N. R. & De Roure, D. C. (2004). Ontological user profiling in recommender systems, *ACM Trans. Inf. Syst.* **22**(1): 54–88.

Mikolajczyk, K. & Schmid, C. (2004). Comparison of affine-invariant local detectors and descriptors, *12th European Signal Processing Conference (EUSIPCO '04)*, pp. 1729–1732.

Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(10): 1615–1630.

Mokji, M. & Abu Bakar, S. (2007). Gray level co-occurrence matrix computation based on haar wavelet, *Computer Graphics, Imaging and Visualisation, 2007. CGIV '07*, pp. 273–279.

Moosmann, F., Triggs, B. & Jurie, F. (2007). Fast discriminative visual codebooks using randomized clustering forests, *In NIPS*.

Muhling, M., Ewerth, R., Stadelmann, T., Shi, B. & Freisleben, B. (2009). University of marburg at trecvid 2009: High-level feature extraction, *Online Proceedings of TRECVID Conference Series 2008* . `http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/marburg.pdf`.

Muller, H., Muller, W., Marchand-Maillet, S., Pun, T. & Squire, D. (2000). Strategies for positive and negative relevance feedback in image retrieval, *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, Vol. 1, pp. 1043–1046.

Müller, H., Pun, T. & Squire, D. (2004). Learning from user behavior in image retrieval: Application of market basket analysis, *Int. J. Comput. Vision* **56**(1-2): 65–77.

Nanas, N., Vavalis, M. & De Roeck, A. (2010). A network-based model for high-dimensional information filtering, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, ACM, pp. 202–209.

Nowak, E. & Jurie, F. (2007). Learning visual similarity measures for comparing never seen objects, *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8.

Nowak, E., Jurie, F. & Triggs, B. (2006a). Sampling strategies for bag-of-features image classification, *European Conference on Computer Vision*, Springer. `http://lear.inrialpes.fr/pubs/2006/NJT06`.

Nowak, E., Jurie, F. & Triggs, B. (2006b). Sampling strategies for bag-of-features image classification, *European Conference on Computer Vision*, Springer. `http://lear.inrialpes.fr/pubs/2006/NJT06`.

Ohbuchi, R. & Furuya, T. (2008). Accelerating bag-of-features sift algorithm for 3d model retrieval.

Olshausen, B. A. et al. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature* **381**(6583): 607–609.

Ortega-Binderberger, M., Mehrotra, S., Chakrabarti, K. & Porkaew, K. (1999). Webmars: a multimedia search engine, pp. 314–321.

Overell, S., Llorente, A., Liu, H., Hu, R., Rae, A., Zhu, J., Song, D. & Rüger, S. (2008). Mmis at imageclef 2008: Experiments combining different evidence sources, *Working Notes from the Cross Language Evaluation Forum*.

Perronnin, F., Dance, C., Csurka, G. & Bressan, M. (2006). Adapted vocabularies for generic visual categorization, *In ECCV*, pp. 464–475.

Perrot, D. & Gjerdigen, R. (1999). Scanning the dial: An exploration of factors in the identification of musical style, *Proceedings of the 1999 Society for Music Perception and Cognition*, p. 88.

Pickering, M. J. & Rüger, S. (2003). Evaluation of key frame-based retrieval techniques for video, *Comput. Vis. Image Underst.* **92**(2-3): 217–235.

Piwowarski, B., Frommholz, I., Lalmas, M. & van Rijsbergen, K. (2010). What can quantum theory bring to information retrieval, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, ACM, pp. 59–68.

Quack, T., Ferrari, V., Leibe, B. & Van Gool, L. (2007). Efficient mining of frequent and distinctive feature configurations, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8.

Quack, T., Mönich, U., Thiele, L. & Manjunath, B. S. (2004). Cortina: a system for large-scale, content-based web image retrieval, *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, ACM, pp. 508–511.

Rahman, M. M., Bhattacharya, P. & Desai, B. C. (2009). A unified image retrieval framework on local visual and semantic concept-based feature spaces, *J. Vis. Comun. Image Represent.* **20**(7): 450–462.

Rijsbergen, C. J. V. (1979). *Information Retrieval*, 2nd edn, Butterworth-Heinemann.

Rocchio, J. (1971). *Relevance Feedback in Information Retrieval*, pp. 313–323.

Rui, Y., Huang, T. S., Ortega, M. & Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval.

Salton, G. & Buckley, C. (1997). Readings in information retrieval, Morgan Kaufmann Publishers Inc., chapter Improving Retrieval Performance by Relevance Feedback, pp. 355–364.

Savarese, S., Winn, J. & Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, IEEE Computer Society, pp. 2033–2040.

Sclaroff, S., La Cascia, M. & Sethi, S. (1999). Unifying textual and visual cues for content-based image retrieval on the world wide web, *Comput. Vis. Image Underst.* **75**(1-2): 86–98.

Serra, X. (2009). *Audio Content Processing for Automatic Music Genre Classification: Descriptors, Databases, and Classifiers*, PhD thesis. `www.tesisenred.net/bitstream/handle/10803/7559/tegt.pdf?sequence=1`.

Shah-Hosseini, A. & Knapp, G. M. (2004). Learning image semantics from users relevance feedback, *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, ACM, pp. 452–455.

Shi, J. & Tomasi, C. (1994). Good features to track, *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, IEEE, pp. 593–600.

Simpson, M. S., Rahman, M. M., Singhal, S., Demner-Fushman, D., Antani, S. & Thoma, G. R. (2010). Text- and content-based approaches to image modality detection and retrieval for the imageclef 2010 medical retrieval track, *CLEF (Notebook*

*Papers/LABs/Workshops)*. `http://clef2010.org/resources/proceedings/clef2010labs_submission_84.pdf`.

Sivic, J. & Zisserman, A. (2003). Video google: a text retrieval approach to object matching in videos, *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Vol. 2, pp. 1470–1477.

Song, D., Lalmas, M., van Rijsbergen, K., Frommholz, I., Piwowarski, B., Wang, J., Zhang, P., Zuccon, G., Bruza, P. D., Arafat, S., Azzopardi, L., Buccio, E. D., Huertas-Rosero, A., Hou, Y., Melucci, M. & Ruger, S. (2010). How quantum theory is developing the field of information retrieval, *AAAI Fall Symposium on Quantum Informatics for Cognitive, Social and Semantic Processes 2010*, AAAI Press, pp. 105–108.

Stricker, M. & Orengo, M. (1995). Similarity of color images, pp. 381–392.

Suyoto, I. S., Uitdenbogerd, A. L. & Scholer, F. (2008). Searching musical audio using symbolic queries, *Trans. Audio, Speech and Lang. Proc.* **16**(2): 372–381.

Tamura, H., Mori, S. & Yamawaki, T. (1978). Textural features corresponding to visual perception, *Systems, Man and Cybernetics, IEEE Transactions on* **8**(6): 460–473.

Teevan, J., Dumais, S. T. & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 449–456.

Tjondronegoro, D., Zhang, J., Gu, J., Nguyen, A. & Geva, S. (2006). Integrating text retrieval and image retrieval in xml document searching, *Proceedings of the 4th International Conference on Initiative for the Evaluation of XML Retrieval*, INEX'05, Springer-Verlag, pp. 511–524.

van Rijsbergen, C. (2004). *The Geometry of Information Retrieval*, Cambridge University Press.

Wang, G., Hoiem, D. & Forsyth, D. (2009). Building text features for object image classification, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, pp. 1367–1374.

Wang, X.-J., Yu, M., Zhang, L., Cai, R. & Ma, W.-Y. (2009). Argo: Intelligent advertising by mining a user's interest from his photo collections, *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '09, ACM, pp. 18–26.

Wang, X., Yang, M., Qi, H., Li, S. & Zhao, T. (2012). Adaptive weighting approach to context-sensitive retrieval model, *8th Asia Information Retrieval Societies Conference* **7675**: 417–426.

Wickerhauser, M. V. & Czaja, W. (2004). A simple nonlinear filter for edge detection in images, *def* **1000**: 2.

Wilson, C. & Srinivasan, B. (2005). Multiple feature relevance feedback in content- based image retrieval using probabilistic inference networks, *in* S. Halgamuge & L. Wang (eds), *Computational Intelligence for Modelling and Prediction*, Vol. 2 of *Studies in Computational Intelligence*, Springer Berlin Heidelberg, pp. 197–208.

Wu, L., Li, M., Li, Z., Ma, W.-Y. & Yu, N. (2007). Visual language modeling for image classification, *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, ACM, pp. 115–124.

Wu, S., Xing, Y., Li, J. & Bi, J. (2012). Adaptive data fusion methods for dynamic search environments, *8th Asia Information Retrieval Societies Conference* **7675**: 336–345.

Yanai, K. (2003). Generic image classification using visual knowledge on the web, *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, ACM, pp. 167–176.

Yang, J., Jiang, Y.-G., Hauptmann, A. G. & Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification, *Proceedings of the international workshop on Workshop on multimedia information retrieval*, MIR '07, ACM, pp. 197–206.

Yu, K., Ma, W.-Y., Tresp, V., Xu, Z., He, X., Zhang, H. & Kriegel, H.-P. (2003). Knowing a tree from the forest: Art image retrieval using a society of profiles, *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, ACM, pp. 622–631.

Yuan, J., Wu, Y. & Yang, M. (2007). Discovery of collocation patterns: from visual words to visual phrases, *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8.

Zhang, P., Hou, Y. & Song, D. (2009). Approximating true relevance distribution from a mixture model based on irrelevance data, *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, ACM, pp. 107–114.

Zhang, S., Huang, Q., Lu, Y., Wen, G. & Tian, Q. (2010). Building pair-wise visual word tree for efficient image re-ranking, *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 794–797.

Zheng, Q.-F., Wang, W.-Q. & Gao, W. (2006). Effective and efficient object-based image retrieval using visual phrases, *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, ACM, pp. 77–80.

# Publications

- Kaliciak, L., Song, D., Wiratunga N., Pan, J., (2013) On the Duality of Fusion Strategies and Query Modification as a Combination of Scores. *Quantum Interactions* (**QI2013**), Leicester, UK.

- Kaliciak, L., Song, D., Wiratunga, N., Pan, J., (2013) Combining Visual and Textual Systems within the Context of User Feedback. *The 19th International Conference on Multimedia Modeling* (**MMM2013**), Springer, LNCS 7732, pp. 445-455, Huangshan, China.

- Kaliciak, L., Song, D., Wiratunga N., Pan, J., (2012) Improving Content Based Image Retrieval by Identifying Least and Most Correlated Visual Words. *The 8th Asia Information Retrieval Societies Conference* (**AIRS2012**), Springer, LNCS 7675, Tianjin, China.

- Kaliciak, L., Horsburgh, B., Song, D., Wiratunga N., Pan, J., (2012) Enhancing Music Information Retrieval by Incorporating Image-Based Local Features. *The 8th Asia Information Retrieval Societies Conference* (**AIRS2012**), Springer, LNCS 7675, Tianjin, China.

- Kaliciak, L., Wang, J., Song, D., Zhang, P., Hou, Y., (2011) Contextual Image Annotation via Projection and Quantum theory Inspired Measurement for Integrating Text and Visual Features. *The 5th International Symposium on Quantum Interactions* (**QI2011**), Springer, pp. 217-222, Aberdeen, UK.

- Kaliciak, L., Song, D., Wiratunga, N., Pan, J., (2010) Novel Local Features with Hybrid Sampling Technique for Image Retrieval. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (**CIKM2010**), pp. 1557-1560, Toronto, Canada.

- Wang, J., Song, D., Kaliciak, L., (2010) Tensor Product of Correlated Text and Visual Features: A Quantum Theory Inspired Image Retrieval Framework. *AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes* (**QI2010**), pp. 109-116, Washington DC, USA.

- Wang, J., Song, D., Kaliciak, L., (2010) RGU at ImageCLEF2010 Wikipedia Retrieval Task. *CLEF (Notebook Papers/LABs/Workshops) 2010*, Padua, Italy.