UNIVERSIDAD NACIONAL DE COLOMBIA

A Strategy for Interactive Exploration of Multimodal Image Collections

Jorge Eliécer Camargo Mendoza

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., Colombia
2014

A Strategy for Interactive Exploration of Multimodal Image Collections

Jorge Eliécer Camargo Mendoza

In fulfillment of the requirements for the degree of:
Doctor en Ingeniería - Ingeniería de Sistemas y Computación

Advisor:
Fabio Augusto González Osorio, Ph.D.

Research Field:
Machine Learning - Image Retrieval
Research Group:
MindLab

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., Colombia
2014

To my wife Fanny
To my daughters Laura y Camila
To my parents and brothers
To all people that support me during this process

Acknowledgements

I want to express my thanks to my advisor professor Fabio González, for his support, confidence, and help. Professor Fabio always motivated me to address each stage of the doctorate process. As a result of this process I developed invaluable abilities and skills, which I will expect to contribute to my country and in general, to the society.

My thanks also to professors Luc Patiny at the École Politechnique Fédérale de Laussane and Pierre Esseiva at the Scientifique Police of Université de Laussane in Switzerland. The doctoral internship was very important in the results of this thesis. Thanks also to professor Eduardo Romero for his comments, which were very important to address some research questions. Many thanks to Mauricio Vargas, his support was fundamental when I was working in the industry.

Thanks to my partners of MindLab, BioIngenium, and CIM@LAB research groups. Special thanks to Juan Caicedo, Angel Cruz, Jorge Vanegas, John Arévalo, Fabio Martínez, Andrea Rueda, Gloria Díaz, Hugo Franco, Francisco Gómez, Fabián Narváez, Raúl Ramos, Cristina Lasso, Marcela Cardona, for their support, suggestions, comments, beers and food, and for their invaluable friendship. Thanks also to Anyela Chavarro, Camilo Salomón, and Oscar Paruma for allowing me to contribute in their own research problems.

Resumen

La producción de contenido multimedia incluyendo documentos de texto, imágenes, videos y audio, ha experimentado un crecimiento exponencial producto del desarrollo de los sistemas de computación y comunicaciones. El proceso de encontrar y acceder este gran volumen de información requiere de métodos computacionales efectivos y eficientes. En el caso de texto, una gran cantidad de trabajo se ha realizado por parte de la comunidad de recuperación de información, y gracias a ese esfuerzo, hoy contamos con motores de búsqueda de documentos de texto apropiados, los cuales nos permiten fácilmente encontrar información. Sin embargo, para el caso de otros tipos de información multimedia, como es el caso de las imágenes, los resultados no son todavía satisfactorios. Nuevos mecanismos para explorar grandes colecciones de imágenes son necesarios de tal forma que se le ofrezca al usuario diferentes alternativas para acceder y encontrar información.

La exploración de colecciones de imágenes es un nuevo mecanismo para acceder grandes repositorios de imágenes de una manera más eficiente y intuitiva. Este mecanismo está siendo activamente investigado por la comunidad científica. El proceso de exploración de colecciones de imágenes se compone de los siguientes componentes: (1) Representación de la colección de imágenes; (2) Construcción de resúmenes; (3) Visualización de colecciones de imágenes; y (4) Interacción con la colección de imágenes. La mayoría de estrategias de exploración que se encuentran en la literatura usan principalmente contenido visual en cada uno de estos componentes, ignorando otras posibles fuentes de información (modalidades) tales como texto, el cual puede complementar cada uno de estos componentes de un sistema de exploración de imágenes.

En esta tesis se estudia cómo conjuntamente utilizar información visual y textual con el fin de modelar de una mejor manera cada uno de los componentes de los sistemas de exploración. Para alcanzar este objetivo, se propone una familia de algoritmos que fusionan ambas modalidades de diferentes formas utilizando métodos de kernel y análisis de semántica latente.

Palabras clave: Exploración de colecciones de imágenes, Aprendizaje de máquina, Procesamiento de imágenes, Construcción de resúmenes, Funcionaes de kernel, Análisis de temas latentes.

Abstract

Multimedia content production, including documents with text, images, videos and audio, has experienced an exponential growth thanks to the development of computer and communication systems. The process of finding and accessing this vast volume of information requires effective and efficient computational methods. In the case of text, a lot of work has been done by the information retrieval community, and thanks to this effort, today we have suitable text document search engines that allow us to easily find information. However, in the case of other type of multimedia content, such as images, the results are not as satisfactory yet. New mechanisms to explore large image collections are necessary to offer the user different alternatives for accessing and finding information.

Image collection exploration is a new mechanism to access large image repositories in a more efficient and intuitive way. This mechanism is being actively studied by the research community. Image collection exploration consists of the following stages: (1) Image collection representation; (2) Image collection summarization; (3) Image collection visualization; and (4) Image collection interaction. Most of the image collection exploration strategies found in the literature mainly use visual content to model each one of these stages, ignoring other possible information sources (modalities) such as text, which may complement each stage of an image collection exploration system.

In this thesis we investigated how to jointly use visual and textual modalities to better model each stage of an image collection exploration system. To reach this goal, we proposed a family of algorithms that fuse both modalities in different ways such as kernel-based methods and latent semantic analysis.

Systematic experiments were conducted on different data sets to evaluate the proposed image collection exploration algorithms in a qualitative and quantitative way. The experimental results showed that the proposed strategy is an effective mechanism for designing image collection exploration systems.

Keywords: Image collection exploration, Machine Learning, Image processing, Summarization, Kernel functions, Latent topic analysis.

Contents

List of Figures

List of Tables

1. Introduction

This thesis addresses the problem of involving multimodal information in the construction of image collection exploration systems. Image collection exploration is a research area in which researchers are actively working to find efficient, effective and intuitive mechanisms to improve the user experience.

This work defines a strategy for interactive exploration of multimodal image collections by proposing a family of algorithms that involves visual and text information in each stage of an image collection exploration system.

1.1. Motivation

Multimedia content, such as text, images, videos and audio, has experienced an exponential growth thanks to the development of more powerful computer systems and communications infrastructure. The process of finding and accessing this vast volume of information requires effective and efficient computational methods. In the case of text, a lot of work has been done by information retrieval researchers, and thanks to its development, today we have suitable systems, such as search engines, that allow us to easily find information. However in the case of other type of multimedia content, such as images, the results are not as satisfactory. Content-Based Image Retrieval (CBIR) is an active research area that investigates the problem of how to efficiently and effectively retrieve images based on their content. CBIR systems require that users express their information needs following one of the following approaches: keyword-based retrieval, query by visual example, and query by semantic image example. In the keyword-based retrieval approach, users describe what they need by means of keywords, so that the system retrieves images with similar annotations associated to them. In query by image example, the user selects an image as query, and the system searches for images with similar visual content. In query by semantic image example, the query is semantically similar to retrieved images, but they are not necessarily visually similar. Nevertheless, users do not always have an image to query the system; in some cases, users do not have a clear idea of what they are looking for.

New mechanisms to explore large information repositories are being proposed by the research community in order to offer the user new alternatives for accessing and finding information. *Image collection exploration* has been shown to be a good strategy for image retrieval, reaching good performance. The success of this strategy is thanks to the inclusion of the user in the loop. In the exploration process, the user interacts with the system whilst it learns from this interaction to deliver more precise results. Typically, a user interacts with the image collection by selecting an image or group of them, and the system retrieves similar images to the selected ones. This process is repeated until the user reaches a target image or a set of relevant images.

Most of the systems visualize only images, ignoring other information sources related to images. Traditionally, images have non-visual information associated to them that complement the visual content. These information sources such as text (explicit) and user's feedback (implicit) are very important since they enrich the visual information to better model the image semantics. These information sources are called *modalities* in this research.

An image collection exploration system is composed of representation, summarization, visualization and interaction stages. *Representation* consists in indexing image content in a computer structure, which captures image characteristics typically in a feature vector. *Summarization* consists in selecting a portion of the repository that faithfully represents the complete collection. Since it is not possible to show the user the complete image data set, because of the computer screen limitations, it is necessary to build an overview that allows the user to see a portion of the image collection at each step of the exploration process. *Visualization* consists in projecting an image collection into a low-dimensional space using a metaphor that represents image relationships. This projection is typically reached by reducing the original high dimensionality of the image representation into 2 or 3 coordinates that better represent the inter-image relationship. *Interaction* is the stage in which the user interacts with the system. At each step of the exploration process the user performs some actions to feedback the system. Typical users' actions are selection of relevant/non-relevant images, image discarding, group selection, zoom, layout reorganization, etc. Figure 1.1 illustrates these image collection exploration stages.

Each of these stages can be benefited from involving complementary visual image information. The main hypothesis is that including this information, the image collection exploration process will be improved. That is, the quality of the summary should be higher, visualization should provide more useful information to the user, and interaction should provide more precise queries to the system.

1.2. Problem Statement

The problem studied in this dissertation is the design of an effective and efficient strategy that allows to explore image collections. In particular, this work focuses on involving visual and text modalities in each stage of an image collection exploration system. The main research question of this work is: *How to exploit multimodal information for improving the user exploration process?*

1.2.1. Representation

CBIR is an active research area that investigates the problem of how to efficiently and effectively retrieve images based on their content. CBIR systems require that users express their information needs following one of the following approaches: keyword-based retrieval, query by visual example, and query by semantic image example. In the keyword-based retrieval approach, users describe what they need by means of keywords, so that the system retrieves images with similar annotations associated to them. In query by image example, the user selects an image as query, and the system searches for images with similar visual content. In query by semantic

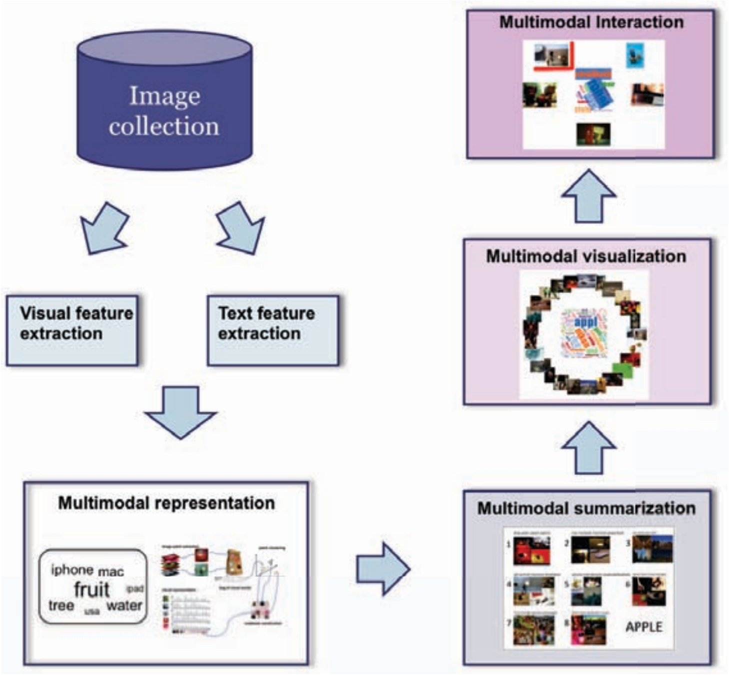


Figure 1.1.: Image collection exploration stages.

example, the query is semantically similar to retrieved images, but these images are not necessarily visually similar to the query. Nevertheless, users do not always have an image to query the system; in some cases, users do not have a clear idea of what they are looking for. The challenge that arises is then *how to reduce the semantic gap between high level human perception of the images and the computer-based image representation.*

1.2.2. Visualization

The visualization problem consists in finding an image representation in a low-dimensional space where similar images are mapped to neighboring regions. Since there are not only images but text, the challenging problem is how to visualize visual and text modalities. Three challenges arise in this stage: *(1) How to visualize each information modality*; *(2) Which modalities should be visualized to the user; and (3) Which visualization metaphor is suitable to visualize both modalities: text and visual content.*

1.2.3. Summarization

The summarization problem is related to the selection of a portion of the collection that best represents the complete data set. This portion is used to see the repository using a ”window”. That is, since it is not possible to see the complete results in the exploration process due to screen limitations, this window allows the user to see a representative subset to interact with. Herein we are interested in building summaries that take into account multimodal information. The challenge is: *How to use multimodal information to build better summaries? How to objectively evaluate the quality of a summary.*

1.2.4. Interaction

In the interaction process, a user can interact with the collection doing different operations such as selection of relevant/non-relevant images, image discarding, group selection, zoom, layout reorganization, etc. So, here the challenge is: *How to interact with the image collection using both modalities to improve the exploration process?*

1.3. Contributions

This thesis presents several contributions to solve each of the problems described above. They are described briefly in the following subsections.

1.3.1. Algorithms for visualization

Latent topic visualization This work presents an algorithm to build multimodal visualizations by fusing tags and visual content in the same latent space. This latent space allows to represent at the same time images and tags, which is used to project both modalities in a 2D visualization space. A quantitative and qualitative evaluation was performed to evaluate the quality of the visualization.

- **Camargo, J.**, Caicedo, J. and González F. *Multimodal Image Collection Visualization using On Non-negative Matrix Factorization.* 14th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2010. Lecture Notes in Computer Science, 2010, Volume 6273/2010, 429-432, 2010.

Kernel-based visualization A system prototype was implemented for visualizing image results using a 2D metaphor. Kernel functions were used to combine different image representations. The system prototype was evaluated in a task-oriented scenario. Results were published in a journal and conference as follows:

- **Camargo**, **J.**, Caicedo, J., González, F. *A Kernel-based Framework for Image Collection Exploration.* Journal of Visual Languages & Computing, Volume 24, Issue 1, February 2013, 53-57. ISSN 10-45-926X.

- **Camargo, J.**, Chavarro, A. and González F. *MedViz: A System for Medical Image Collection Visualization and Exploration.* V Seminario Internacional Procesamiento y Análisis de Imágenes Médicas (SIPAIM), 2009.

The proposed kernel based visualization framework was applied to a medical image collection and presented in human computer interaction conference:

- **Camargo, J.**, Caicedo, J. and González, F. *Kernel-Based Visualization of Large Collections of Medical Images Involving Domain Knowledge.* X Congreso Internacional de Interacción Persona-Ordenador, 2009.

The following work presented a comparison of different aspects of an image collection visualization such as visualization metaphor, distance function, and image representation. This work allowed to understand the importance of such aspects in the construction of an image collection exploration system:

- **Camargo, J.** and González, F. *Visualization of Large Collections of Medical Images.* IV Congreso Colombiano de Computación (4CCC), Universidad Autónoma de Bucaramanga, 2009.

1.3.2. Algorithms for summarization

The main contributions on this topic are the models associated to the construction of multimodal summaries. The summarization process was addressed from two different perspectives: latent topic based summarization and clustering based summarization.

Latent topic summarization algorithms The following works were focused on the construction of multimodal summaries by fusing text and visual modalities in a latent space. This latent space is used to perform a topic analysis to select a set of representative images of the collection. A set of performance measures were proposed to evaluate the quality of the obtained summaries:

- **Camargo, J.** and González F. *Multimodal Image Collection Summarization using Non-negative Matrix Factorization.* 6th Colombian Computing Congress, IEEE 6CCC, 2011.
- **Camargo, J.**, González, F. *MICS: Multimodal Image Collection Summarization by Optimal Reconstruction Subset Selection.* 8ht Congreso Colombiano de Computación, IEEE 8CCC, 2013.

An extended version of this work is being prepared for a Journal submission:

- **Camargo, J.**, González, F. *Latent Topic Analysis for Multimodal Image Collection Summarization. To be submitted*, 2014.

Cluster-based summarization algorithms A second family of algorithms was proposed to summarize image collections. This family of algorithms involves semantic information from a supervised perspective to better model image semantics.

- **Camargo, J**. and González F. *A Multi-Class Kernel Alignment Method for Image Collection Summarization.* 14th Iberoamerican Congress on Pattern Recognition (CIARP2009), LCNS 5856, pp. 545-552, 2009.

An extended version of this work was published in the Journal of Visual Languages & Computing:

- **Camargo**, **J.**, Caicedo, J., González, F. *A Kernel-based Framework for Image Collection Exploration.* Journal of Visual Languages & Computing, Volume 24, Issue 1, February 2013, 53-57. ISSN 10-45-926X.

In this work a summarization algorithm was proposed to analyze illicit pill distribution networks. Ecstasy image pills were clustered to visually and quantitatively evaluate relationships between distribution batches. A tool that visualizes clusters was developed to allow scientific police to explore a hierarchy of clusters in a visual way:

- **Camargo**, **J.**, Esseiva, P., González, F., Wist, J., Patiny, L. *Monitoring of illicit pill distribution networks using an image collection exploration framework*. Forensic Science International, Volume 223, Issue 1, November 2012, 298-305. ISSN 0379-0738, 2012.

1.3.3. Algorithms for interaction

Kernel-based framework for exploratory search

The following publication presents an image collection interaction model to learn from the user's feedback. The proposed model uses kernel functions to involve semantic information in an image collection search scenario, specifically a category search task. A user-centered experimentation was conducted to evaluate the system's performance:

- **Camargo, J.**, Chavarro, A. and González F. *A Kernel-Based Strategy for Exploratory Image Collection Search.* IEEE CBMI 2010.

An extended version of this work was published in the Journal of Visual Languages & Computing:

- **Camargo**, **J.**, Caicedo, J., González, F. *A Kernel-based Framework for Image Collection Exploration.* Journal of Visual Languages & Computing, Volume 24, Issue 1, February 2013, 53-57. ISSN 10-45-926X.

Drug intelligence interaction framework

A real image collection exploration system was developed at the École Polytechnique Fédérale de Laussane (EPFL), Switzerland. The department of scientific Police of this university actively works in an area named *drug intelligence*, which studies the production and distribution of illegal drugs such as Ecstasy. In this work, a concrete system was developed to help the scientific Police to analyze the distribution of Ecstasy pills. Visualization, summarization and interaction models were used to implement such as system. Results of this work were published at:

- **Camargo**, **J.**, Esseiva, P., González, F., Wist, J., Patiny, L. *Monitoring of illicit pill distribution networks using an image collection exploration framework*. Forensic Science International, Volume 223, Issue 1, November 2012, 298-305. ISSN 0379-0738, 2012.

1.3.4. Other contributions

Additional papers were published as a result of preliminary work and collaborations performed with other institutions.

State of the art reviews

One of the main research tasks in this thesis was to build an up-to-date state of the art in image collection exploration:

- **Camargo, J.** and González, F. *Visualization, Summarization and Exploration of Large Collections of Images: State of the Art.* LatinAmerican Conference On Networked and Electronic Media (LACNEM), 2009.

- Chavarro, A., **Camargo, J.**, and González F. *Exploration and Retrieval in Large Biomedical Image Collections: A State of the Art.* 6th International Seminar on Medical Image Processing and Analysis (SIPAIM), 2010.

Additional publications

- Caicedo, J., **Camargo**, **J.**, and González, F. *Content-Based Access to Medical Image Collections.* Book chapter in Biomedical Image Analysis and Machine Learning Technologies: Applications and Techniques. Idea Group Inc, 2009.

- González, F., Caicedo, J., **Camargo**, **J.**, Cruz, A., Spinel, C., Romero, E. *Sistema para Acceder una Colección de Imágenes Biomédicas Mediante Estrategias Basadas en el Contenido.* V Seminario Internacional Procesamiento y Análisis de Imágenes Médicas, SIB-SIPAIM, 2009.

- González F., Caicedo J., Cruz Roa A., **Camargo J.**, Spinel C. *A system for accessing a collection of histology images using content-based strategies.* Acta Biológica Colombiana. Vol 15, No 3, 221-234, 2010.

- Pérez S., Otálora, S., **Camargo**, J., González, F. *Explorando Grandes Colecciones de Imágenes de Histología a través de Factores Latentes.* 7th International Seminar on Medical Information Processing and Analysis (SIPAIM), 2011.

- Vanegas J., Caicedo J., **Camargo J.**, Ramos-Pollán R., González F. Bioingenium at ImageCLEF 2012: Textual and Visual Indexing for Medical Images. CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, 2012.

- González F.A, Caicedo J.C**.**, Cruz-Roa A., **Camargo J.**, Romero E., Spinel C., Seligmann D., Forero J. *A Web-Based System for Biomedical Image Storage, Annotation, Content-Based Retrieval and Exploration.* Microsoft eScience, 2009.

- Ramos-Pollan, R., Gonzalez, F.A., Caicedo, J.C., Cruz-Roa, A., **Camargo, J**., Vanegas, J.A., Perez, S.A., Bermeo, J.D., Otálora, J.S., Rozo, P.K., Arévalo, J.E. *BIGS: A framework for large-scale image processing and analysis over distributed and heterogeneous computing resources.* E-Science (e-Science), 2012 IEEE 8th International Conference on, 2012.

- Chavarro, A., **Camargo, J.**, González F. *Multimodal Image Collection Visualization.* XVIII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial, STSIVA, 2013.

1.4. Thesis Organization

The remaining chapters of the thesis are organized as follows:

- *Chapter 2: A Review of Image Collection Exploration Methodologies.* This chapter discusses previous works that address each one of the stages of an image collection exploration process. This state of the art allows to understand different approaches used in the literature.

- *Chapter 3: Multimodal Latent Topic Analysis for Image Collection Summarization.* This chapter presents a multimodal latent topic analysis method for the construction of image collection summaries. The method automatically selects a set of prototypical images from a large set of retrieved images for a given query. To build such a summary a new method named MICS is proposed, which combines textual and visual modalities in a common latent space that allows to find a subset of images from which the whole collection can be reconstructed. Experiments were conducted on a collection of tagged images and demonstrate the ability of the approach to build summaries with representative visual and semantic content. The proposed method was compared to other state-of-the-art image collection summarization approaches by using objective measures such as reconstruction error ability and diversity of the summary.

- *Chapter 4: Multimodal Image Collection Visualization Based on Non-negative Matrix Factorization.* A new multimodal visualization method is presented in this chapter, in which visual and textual content are projected in the same visualization space. The resulting image visualization allows to identify the relationships between images and text terms, allowing to understand the distribution of the collection in a more semantic way. The resulting visualization scheme can be used as the core metaphor in an interactive image exploration system.

- *Chapter 5: A Kernel-based Framework for Image Collection Exploration.* This chapter proposes a framework for the construction of an image collection exploration system based on kernel methods, which offers a mathematically strong basis to address each stage of an image collection exploration system: image representation, summarization, visualization and interaction. In particular, the proposed approach emphasizes a semantic representation of images using kernel functions, which can be seamlessly harnessed across all system components. Experiments were conducted with real users to verify the effectiveness and efficiency of the proposed strategy.

- *Chapter 6: Monitoring of Illicit Pill Distribution Networks Using an Image Collection Exploration Framework .* This chapter proposes a novel approach for the analysis of illicit tablets based on their visual characteristics. In particular, the chapter concentrates on the problem of ecstasy pill seizure profiling and monitoring. The presented method extracts the visual information from pill images and builds a representation of them. Different visual features were used to build different image similarity measures. The discriminative proposed model permits to infer whether two pills come from a same seizure, while a clustering model groups of pills that share similar visual characteristics. The resulting clustering structure allows to perform a visual identification of the relationships between different seizures. The proposed approach was evaluated using a data set of Ecstasy pill pictures. The results demonstrate that this is a feasible and cost effective image collection exploration method for performing pill profiling and monitoring.

- *Chapter 7: Conclusions.* The final chapter presents the main conclusions and discussions of the dissertation, summarizes the main contributions, and highlights the most important findings. Also, some future research directions are presented and discussed.

2. A Review of Image Collection Exploration Methodologies

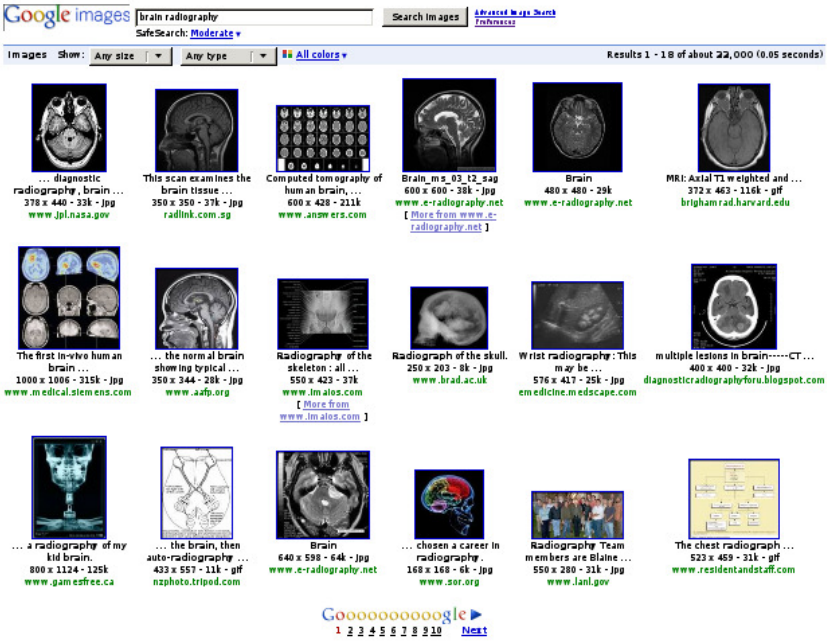
Due to the amount of multimedia content generated with different kind of devices and to the ease of publishing in the web, it is necessary to build suitable tools that allow us to manage this information. This generates problems like how to find efficiently and effectively the information needed, and how to extract knowledge from the data. These issues have been extensively studied by Information Retrieval (IR) researchers, but the main focus has been on textual data [5]. Therefore, there are still a huge amount of work to do on other kind of non-textual data such as images.

This chapter provides a brief but comprehensive state of the art of the recent technical achievements in image collection exploration techniques. Major recent publications are included in this chapter covering different aspects of the research in this area, including visualization, summarization and exploration. In addition, some other related issues such as performance measures and experimental setups are also discussed. Finally, based on existing technology and the demand from real-world applications, some promising future research directions are discussed.

2.1. Visualization techniques

Information visualization offers ways to reveal hidden information (complex relationships) in a visual representation and allows the users to seek information in a more efficient way [6]. Thanks to the human visual capacity for learning and identifying patterns, visualization is a good alternative to deal with this kind of problems [7]. However, visualization itself is a hard problem; one of the main challenges is how to find low-dimensional, simple representations that faithfully represent the complete dataset and the relationships among data objects [8]. The majority of existent approaches use a 2D grid layout for visualizing results. Figure 2.1 shows a screenshot of the result for a query in Google Similar Images. The main problem of this kind of visualization is that it does not make explicit the relationships among the presented images and only a portion of the results is shown to the user.

Due to the large amount of visual and multimedia data generated in the Internet, health centers, enterprises, research community, and others, it is necessary to build new mechanisms that allow us to access multimedia data sets in an effective and efficient way. We are interested in providing to the user new ways to navigate collections of multimedia data, specifically images, such that user can visualize and explore it in an intuitive way. The first natural question is how to visualize an image collection? In the original space images are represented by many dimensions, so how to reduce the dimensionality such that users can visualize an image in a two



brain radiography

SafeSearch: Moderate

Show: Any size, Any type, All colors

Results 1 - 18 of about 22,000 (0.05 seconds)

Figure 2.1.: Typical visualization grid layout using Google Images

dimension space? Assume that we have a way for visualizing the image collection: how to display a summary of the entire collection in a computer screen? Once we have a way to visualize and summarize the collection, how we allow users to explore the images in an intuitive way taking into account the similarity among images? Finally, how to evaluate the performance of the techniques used to solve the mentioned issues? These questions are open and are being addressed in some recent works. In general, a document (e.g. an image) is represented by a large set of features, this implies a high-dimensional representation space. The visualization of this space requires its projection into a low-dimensional space, typically 2D or 3D, without losing much information. This general problem has been tackled using different approaches, which are briefly discussed in the following paragraphs.

Multidimensional Scaling (MDS) MDS [9] techniques are a family of methods that focus on finding the subspace that best preserves the inter-point distances using linear algebra operations. The input is an image similarity matrix corresponds to the high-dimensional space and the result is a set of coordinates that represent these images in a low dimensional space [6]. Figure 2.2 shows a visualization of Corel dataset using this method.

Principal Component Analysis (PCA) PCA [10] is an Eigenvector method designed to model linear variabilities in high-dimensional data. The method computes the linear projections of greatest variance from the top Eigenvectors of the data covariance matrix. In classical MDS, the low dimensional embedding is computed such that best preserves pair wise distances among objects. If these distances correspond

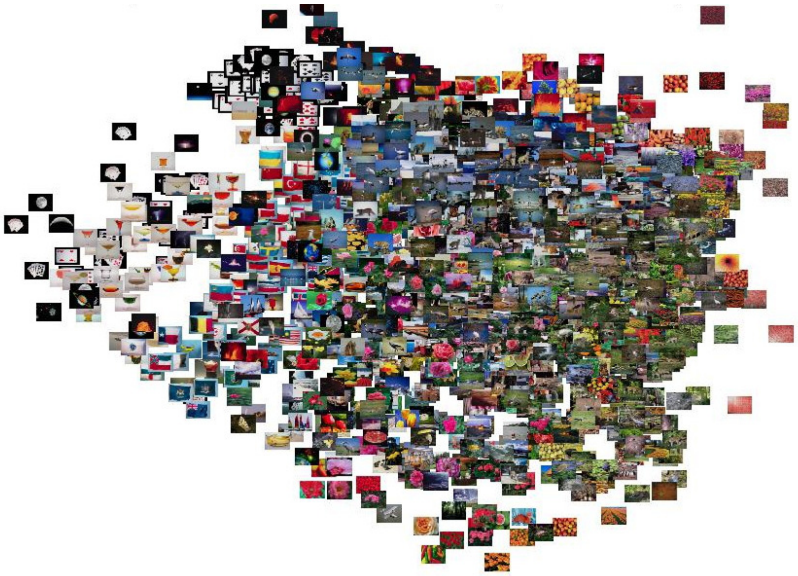


Figure 2.2.: Visualization of Corel dataset using MDS method [1]

to Euclidean distances, the results of metric MDS are equivalent to PCA [11].

Isometric Mapping (Isomap) Isomap [12] uses graph-based distance computation in order to measure the distance along local structures. The technique builds the neighborhood graph using k-nearest neighbors, then uses Dijkstra's algorithm to find shortest paths between every pair of points in the graph, then the distance for each pair is assigned the length of this shortest path and finally, when the distances are recomputed, MDS is applied to the new distance matrix [8].

Self-Organizing Maps (SOM) SOMs [13] are a family of techniques based on artificial neural networks and unsupervised learning. These methods are designed for data clustering, information visualization, data mining and data abstraction. SOMs are a special topology-preserving map because intrinsic topological structures and important features of input data are revealed and kept in the resulting output grid. Basically, the method consists of finding the best match between the input signal and all neurons in the output grid, it means, all neurons are competing for the input signal [6].

Curvilinear Component Analysis (CCA) CCA [14] is a method based on SOMs and belongs to the class of distance-preserving methods closely related to Sammon's NLM. CCA is also known as VQP (Vector Quantization and Projection). CCA and SOM methods work with the same kind of optimization techniques, the algorithm performs simultaneously the vector quantization and the nonlinear dimensionality reduction, exactly like an SOM [15].

Laplacian Eigenmaps (LE) LE [16] corresponds to a family of techniques based on spectral decomposition. This kind of methods try to remedy some problems of other spectral methods like Isomap [12] and LLE [11]. LE develops a local approach to the problem of nonlinear dimensionality reduction and it is closely related to LLE. This method, instead of reproducing small linear patches around each datum, relies on graph-theoretic concepts like the Laplacian operator on a graph. LE is based on the minimization of local distances and to avoid the trivial solution where all points are mapped to a single points, the minimization is constrained [15].

Isotop Isotop [17] is a method that is based on graphs, which focuses on addressing the limitations of SOMs when they are used for nonlinear dimensionality reduction. Isotop reduces the dimensionality in three steps: vector quantization (reduction of the number of points), graph building and low-dimensional embedding [15].

Samnon's Mapping (NML) Sammon [18] proposed the method NML (Standing for Nonlinear Mapping), which establishes a mapping between a high-dimensional space and a lower-dimensional one. Author proposes to reduce the dimensionality of a finite set of data points. NLM is closely related to metric MDS, where no generative model of data is assumed: only a stress function is defined. In this method, the low-dimensional representation obtained can be totally different from the distribution of the true latent variables [15].

Locally Linear Embedding (LLE) LLE [11] is an unsupervised learning algorithm that computes low-dimensional neighborhood preserving embeddings of high dimensional data. LLE proposes an approach based on conformal mappings, which is a transformation that preserves local angles. To some extent, the preservation of local angles and that of local distances are related and may be interpreted as two different ways to preserve local scalar products.

Stochastic Neighbor Embedding (SNE) SNE [19] is a method based on the computation of probabilities of neighborhood assuming a Gaussian distribution in both, the high dimensional and the 2D space. Basically, this method tries to match both probability distributions.

Kernel PCA Kernel PCA is the application of PCA in a kernel-defined feature space [20], using a kernel, the original linear operations of PCA are done in a reproducing kernel Hilbert space with a non-linear mapping. The main idea of KPCA consists of reformulating the PCA into its metric MDS equivalent, or dual form [15].

Kernel Isomap In [21], authors proposed *kernel Isomap,* a modification of the original Isomap method inspired in kernel PCA, where generalization and topological stability problems found in the original Isomap method are addressed.

2.2. Summarization techniques

Automatic summarization has been broadly studied in areas such as document summarization, which aims to create a summary from a document retaining the most important points of the original document. In a broader sense, multi-document summarization tries to build a text summary from a set of documents [22].

Image collection summarization consists in finding a representative set of images from a larger set of images. Different methods have been proposed to build such a summary. One of the main problems when search engines return results is that it is not possible to display all images to the user. Therefore, only a small set of images can be displayed to users. A summary also is useful to show an *overview* of the complete collection in order to allow the user to begin the exploration process.

Clustering In [23], authors propose an exploration system with visualization and summarization capabilities. They extract image features, summarize the collection with k-means (for building a hierarchy of clusters), and project the clusters with MDS. It is possible to annotate the collection in a semi-automatic way thanks to the cluster hierarchy (perceptual concepts). For each cluster in all hierarchy levels, it is selected the most similar image (eID) to the others in the cluster and then it is used to give access to its IDs on a lower level. With this method, it is possible to explore the collection in a detailed way according to the level of the hierarchy (details on demand).

In [24], authors select a set of images that represents the visual content of a given scene. They examine the distribution of images in the collection to select a set of canonical views to form the summary, using clustering techniques on visual features. The summary is improved with the use of textual tags in a learning phase.

Tree structure This kind of methods apply successively the clustering method in order to break the collection in a hierarchy of clusters. The first overview is obtained applying for example *k-means* for selecting the k most representatives images. Then, it is applied again the clustering algorithm to images belong to each cluster represented by the medoid. It process is repeated until the collection be totally divided.

In [25], authors build a cluster hierarchy of images based on keywords and pixel values, and representative images are selected for each cluster. This task is performed in a pre-processing step. Authors propose a hierarchical data visualization technique to visualize the tree structure of images using nested rectangular region.

Similarity pyramids Similarity pyramids focuses on the use of hierarchical tree structures. In [26], authors develop a search algorithm based on best-first branch and bound search. The tree structure is used to perform an approximate search. In this work it is proposed a hierarchical browsing environment called similarity pyramid. The similarity pyramid groups similar images together while allowing users to view the database at varying levels of resolution. The similarity pyramid is best constructed using agglomerative clustering methods, and presents a fast sparse clustering method which reduces both memory and computation over conventional methods.

Nearest neighbors [27] explores the use of a nearest neighbor network. Authors created a prototype that visualizes the network of images that are connected by similarity in a nearest neighbor network. They assume that if an image is selected or deselected, the same action can be performed on its neighbors. As result, following actions are possible: selecting an image, selecting an image with its nearest neighbors, deselecting an image, deselecting an image with its nearest neighbors, and growing the selection with all the nearest neighbors. Authors tested the proposed method with some interaction scenarios in experimentation phase. Results show that the nearest neighbor network can have a positive effect on the interaction effort needed to select images compared to the baseline of sequentially selecting images.

Graphs Graph methods build a graph representation of the image collection. The graph is constructed such that vertices are the images and edges represent the similarity among images. In [28], authors address the problem of clustering web image search results. The method proposed is based on organizing the images into semantic clusters. Authors propose a hierarchical clustering method using visual, textual and link analysis. The method uses a vision-based page segmentation algorithm and the textual and link information of an image can be extracted from the block containing that image. By using block-level link analysis techniques, an image graph can be constructed. Then, they apply spectral techniques to find a Euclidean embedding of the images respecting the graph structure. For each image, they have three kinds of representations, i.e. visual feature based representation, textual feature based representation and graph based representation. Using spectral clustering techniques, authors can cluster the search results into different semantic clusters.

In [29], authors propose to exploit both low visual features and text for clustering. They called this method *consistent bipartite graph co-partitioning,* which clusters web images based on visual features and text fusion. They formulate a multi-objective optimization problem, which can be solved by semi-definite programming (SDP). Authors base their propose on the use of spectral clustering and bipartite spectral clustering partition. The algorithm proposed is called F-I-T (low-level Features, Images, Terms in surrounding texts). In the experimentation, they crawled the Photography Museums and Galleries of the Yahoo Directory.

Lattices In [30], authors use a structure visualization method by formal concept analysis. They build a image summary based on the lattice structure and propose an algorithm that generates predictive frames from the original frames and divides them to blocks with suitable size. Authors calculate standard deviation respect to each block, and construct information table, where the objects and attributes correspond to frames and the absolute mean of pixels in the block, respectively. A concept lattice with respect to the information table is obtained by the formal concept analysis, and it is helpful to understand the overview of the image databases.

Attribute partition In [31], authors propose a automatic organization method based on analysis of time stamps and image content, in order to allow user navigation and summarization of images (photos). They use attributes like time and image content in order to partition related images in two stages. From the partitions, key photos are selected to represent the partition based on content and then are used for building the summary. Authors are focused on building image summaries for camera phones.

Ontologies Concept ontology has recently been used to summarize and index large-scale image collections at the concept level by using some hierarchical inter-concept relationships [32]. Some works like [33, 34] have used this method. In [33], authors developed a new scheme for achieving multilevel annotations of large-scale images automatically. Global visual features and the local visual features are extracted for image representation. Authors used kernels to characterize the diverse visual similarity relationships between the images, and a multiple kernel learning algorithm is developed for SVM image classifier training. Authors used ontologies and a hierarchical boosting algorithm in order to learn the classifiers hierarchically. They developed a hyperbolic framework for visualizing and summarizing the image collection.

Kernel-based methods Other kind of approaches are based on kernel methods in order to build the image collection summary. In [32], authors use kernel functions and combinations of them (mixture-of-kernels) for involving semantic in visual summarization. They experiment with *Flickr* dataset which has 1.5 million of images approximately. Authors propose a clustering algorithm for summarizing the collection and it is possible to select a number of images to be displayed in the summarization.

2.3. Interaction techniques

Exploration plays an important role to users to interact with the visualization, assess the relevance between the returned images and their real query intentions, and direct the system to find more relevant images adaptively [32]. Exploration allows the user to navigate the collection in an intuitive way through visual controls. In this section, we present some of the most used techniques to interact with collections of images.

Ranked-based list This is the conventional method used in search engines, where a ranked list of images is shown to the user. Usually, user interacts with the list by clicking the links at the end of the page that jump among pages partitioned by a fixed number.

Clustering-based exploration In clustering-based exploration, it is offered a panel to browse the search results by looking at the preview of each cluster and a visualization area where images belong to the cluster.

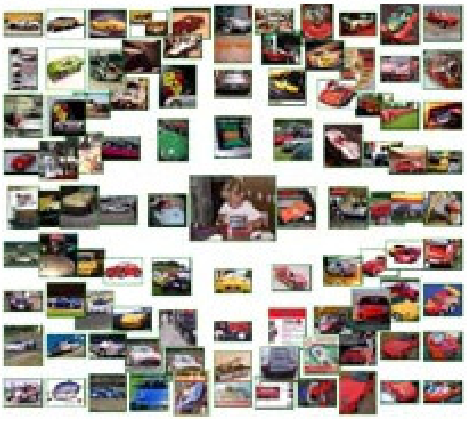


Figure 2.3.: Fisheye view. A distorted polar coordinate system that modifies the spatial relationship of images on the presentation view [2].

Fisheye view Usually, users are interested in a small part of the image collection, so they will feel more convenient if part of the image collection is presented instead of showing the entire collection. Fish view is a suitable tool to allow users to see local details and global perspective simultaneously. This technique is based on a distorted polar coordinate system that modifies the spatial relationship of images on the presentation view. An example of a fisheye view is shown in Figure 2.3.

In [2], authors propose different mechanisms to explore an image collection: ranking-based list, cluster-based and fisheye view. They carry out a user study to compare the approaches and they use in their prototype a slider to adjust the image overview. The experimentation phase is performed with real users that search objective images while time is measured.

Tree-maps Tree-maps are a rectangular, space-filling approach for visualizing hierarchical data. In [35], authors use a 2D tree visualization where the tree nodes are encapsulated into the area of its parent node. The size of the single nodes is determined proportionally in relation to all other nodes of the hierarchy by an attribute of the node. PhotoMesa [36] is an example of system that uses this visualization technique.

Hyperbolic and cone trees In [37], authors present a framework for the visualization of hierarchies: hierarchical visualization system (HVS). HVS provides a synchronized, multiple view environment for visualizing, exploring and managing large hierarchies. HVS includes tree views, a walker tree layout, information pyramids, tree-maps, hyperbolic, sunburst, and cone trees.

MoireGraphs MoireGraphs [3] combine a focus+context radial graph layout with a family of interaction techniques to help in the exploration of graphs. A MoireGraph displays a spanning tree induced upon a visual node graph using a radial

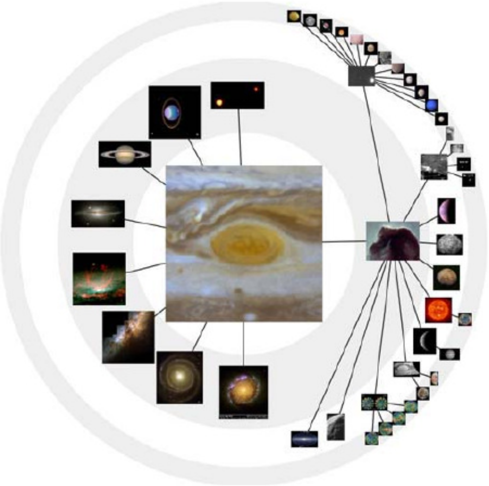


Figure 2.4.: A MoireGraph for a subset of the NASA Planetary Photo journal image collection [3].

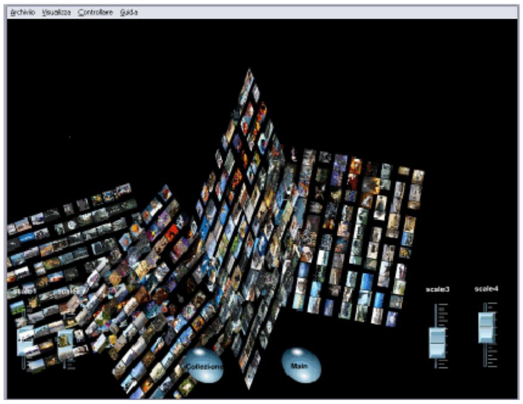


Figure 2.5.: Tornado of planes method [4].

focus+context graph layout allowing interactive exploration of the graph. Figure 2.4 shows an example of this method.

Other non-conventional methods In [4], authors developed various methods for visualizing and exploring large collection of images (cube, snow, snake, volcano, funnel, elastic image browsing, shot display, spot display, cylinder display, rotor display, tornado display and tornado of planes display). These methods are non-conventional and are new interesting ways to offer to user navigating mechanisms. An example of one of the methods proposed is illustrated in Figure 2.5.

2.4. Performance measures

The development of image collection exploration systems requires measures that provide information about how good visualization, summarization and interaction methods are. In this section, we describe some objective and subjective performance measures used in particular tasks.

2.4.1. Visualization measures

Kruskal stress Stress [38] is a measure used in multidimensional scaling, which expresses the difference between the distances d in the original space and the distances D in the projected space for all the images. A small value of stress indicates that the original distances have been preserved in the projected space [39, 40]. Stress is calculated as follows,

$$Stress = \frac{\sum_{i,j}(d_{i,j} \quad D_{i,j})^2}{\sum_{i,j}(D^2{}_{i,j})} \tag{2.1}$$

, where $d_{i,j}$ is the image distance in the original space and $D_{i,j}$ is the distance in the projected space.

Kullback-Leibler Kullback-Leibler [41] is a measure that calculates the difference between the distribution probabilities of the original and projected spaces. In [8], authors try to match the two probability distributions for finding the optimal projected positions by minimizing a cost function. This cost optimization is used in order to preserve the *structure*. The distance between these distributions is calculated using Equation 2.2. The lower the cost, the better the projection.

$$C_s = \sum_i \sum_j P_{i,j} log \frac{P_{i,j}}{Q_{i,j}}, \tag{2.2}$$

where $P_{i,j}$ is the probability that an image i would pick j as its neighbor in the high dimensional space, and $Q_{i,j}$ is a target probability. For details see [8].

2.4.2. Summarization measures

Cluster-based Clustering methods are commonly used in image collection summarization tasks. Therefore, measures such as purity, entropy, intra-cluster/inter-cluster similarity, mutual information, Hubert statistic [42] [8] , and confusion matrix, are used to evaluate cluster quality.

Retrieval-based Performance measures from image retrieval are also used to evaluate the quality of summaries. Measures such as recall [43] and precision [44] are used to reach this goal.

Result diversification Image result diversification aims to improve the diversity of images returned by a search engine. Measures such as diversity score [45], intra-list similarity [46] and Folwes-Mallows index [47] are examples of these measures.

2.4.3. Interaction measures

Searching time Searching time [4] is measure used to determine the time taken by users to find a target image.

Searching efficiency Searching efficiency [4] is measure used to determine the ratio between the percentage of correct images selected by a user and the navigation time.

2.4.4. Subjective evaluation

Other works evaluate the performance of an image collection exploration system from a user-center perspective. The objective of these user-center evaluation is to measure aspects such as usability, easiness, efficiency and effectivity of the system. Questionnaires are commonly used to ask the users in a quantitative and qualitative way how the system's performance is [48][49].

2.5. Software

In this section, we relate some tools used in the information visualization area. This list does not pretend to be exhaustive, therefore it only provides illustrative examples of real systems implemented by the research community.

Xcavator [50] is a stock photo search portal for the community. This tool allows to browse visually through millions of stock images, vector illustrations, flash files, and videos. Xcavator supports searching by image examples and searching by text.

Automatic Photo Tagging and Visual Image Search (ALIPR) [51], is an on line search engine that supports searching by image examples. User can select an image from his/her computer and the system searches similar images.

Caliph & Emir [52] are MPEG-7 based Java prototypes for digital photo, image annotation and retrieval supporting graph like annotation for semantic meta data and content based image retrieval using MPEG-7 descriptors.

GGobi [53] is an open source visualization program for exploring high-dimensional data. It provides highly dynamic and interactive graphics such as tours, as well as familiar graphics such as the scatter plot, bar chart and parallel coordinates plots. Plots are interactive and linked with brushing and identification.

Google Image Visualization and Filtering [54] is a demo mainly to demonstrate how machine leaning, image analysis and visualization techniques can work together to enhance content-based image retrieval and junk image filtering. It uses Treebolic a free software (Hyperbolic tree Engine, Generator and Browser).

PhotoMesa[1] is a desktop tool that incorporates a zoomable image browser and it allows the user to view multiple directories of images in a zoomable environment using a set of simple navigation mechanisms to move through the space of images.

Google Image Swirl [55] is tool in which visualization, summarization and interaction techniques are incorporated.

[1] http://www.photomesa.com

2.6. Applications

Search engines Search engines are currently powerful tools that allow the user to find information easily, so image collection exploration can offer new mechanisms in order to improve the user experience. Systems like Google Images[2] and Flickr[3] are search engines that can be improved in this sense. The following are some examples in which image collection exploration can improve the user experience:

Personal photo collections People generate a lot of pictures that are captured with cameras and cellular phones. These images are generally organized by meta information such as date and user tags. However this organization is not enough when users want to find pictures by other characteristics such as place, people, event, etc. An image collection exploration system could automatically organizes the picture collection through summarization methods and visualize image results when similar images are returned to users [56].

E-commerce On-line shopping web sites can also be benefited of image collection exploration. Conventionally, images are organized by predefined categories, which are not the best navigation scenario when users do not know what they are searching. These categories are represented by some images, which are selected by administrators. However, automatic selection of a summary of each product category could allow users to see representative image products of each category [57].

Biomedical image collections A huge amount of medical images are produced routinely in health centers that demand effective and efficient techniques for searching, exploration and retrieval. These images have a good amount of semantic, domain-specific content that has to be modeled in order to build effective medical decision support systems. Scientific medical papers could be retrieved by visual content and then displayed to user. Diagnostic image collection exploration could allow to physicians in an intuitive way finding medical images. Pattern finding in medical image collections is other kind of application, where medical experts could find patterns that are hidden thanks to the visualization. Dynamic visualization based on online user relevance feedback could allow to the system to learn from the user actions in order to dynamically improve the exploration process.

2.7. Conclusion

Image collection exploration is an interesting area that has challenging problems that have being addressed by the research community using different approaches. In this chapter, we survey different techniques used for visualizing, summarizing and exploring large collection of images. This is an area that is actively studied due to current solutions are not satisfactory at all. We also review the main performance measures used to measure how good are the methods used in the area. Current

[2] http://images.google.com
[3] http://www.flickr.com

performance measures are not enough, it is necessary to define new formal measures that allow researchers measure how good is certain method with respect to others. Machine learning offers interesting methods that learn from the user interaction to improve the exploration process. Search engines can improve the user searching experience, involving exploration techniques that reduce the search time. Users want to find efficiently and effectively images, but in many cases, they do not know how to start the search. With an overview of the collection, users can start to explore the images and thus they can define their needs.

Part I.

Summarization

3. Multimodal Latent Topic Analysis for Image Collection Summarization

This work is being prepared to be submitted to the Pattern Recognition Journal.

This chapter presents a multimodal latent topic analysis method for the construction of image collection summaries. The method automatically selects a set of prototypical images from a large set of retrieved images for a given query. We define an image collection summary as a subset of images from a collection, which is visually and semantically representative. To build such a summary we propose MICS, a method that combines textual and visual modalities in a common latent space, which allows to find a subset of images from which the whole collection can be reconstructed. We conducted experiments on a collection of tagged images and demonstrate the ability of our approach to build summaries with representative visual and semantic content. Our method was compared to other state-of-the-art image collection summarization approaches by using objective measures such as reconstruction error ability and diversity of the summary, showing competitive results.

3.1. Introduction

The large amount of images produced every day requires of suitable systems to efficiently and effectively manage them. Photo-sharing systems like Flickr[1] pose important challenges to organize, browse and query large image collections. The typical scenario to search images within Flickr consists in providing a query by means of keywords, which is processed by the system to return a set of similar images according to a similarity criteria. Although this paradigm has been satisfactorily used in search engines for searching textual content, it is not necessarily the most suitable way to interact with large image collections. One of the problems of this approach is that, in general, textual queries are not enough to express the visual richness of images and therefore the most relevant images are not necessarily at the top of the search results. Conventionally the user only explore the first result pages, so if the user does not navigate the other pages, (s)he will not see other images that could be of interest. Figure 3.1 shows the top 24 images returned by Flicker for the query *apple*. Note that the returned images have some relation with the *apple*

[1] http://www.flickr.com

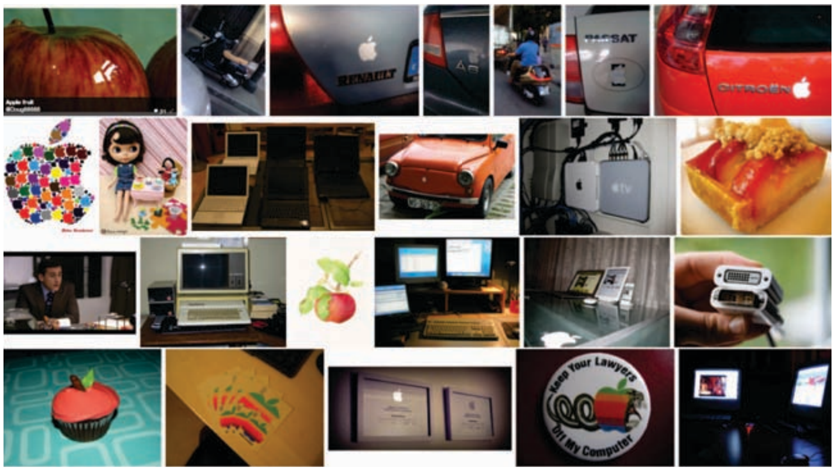


Figure 3.1.: The first 24 images retrieved by Flickr for the query *apple* (retrieved on October 1, 2013).

term since the associated tags contain such term. However this term can be used to describe different semantic concepts such as fruit, computers, food, cake, etc. The returned images in this example are not representative (iconic) of the complete set of results, so the user only has access to a small portion of them in a first view, and some relevant images may be missed by the user.

Automatic image collection summarization is the process of selecting a small set of representative images that allows to highlight larger amounts of images. This process becomes very important to enable interactive navigation and exploration of large-scale image collections [58, 59].

This chapter proposes a new method to automatically build *multimodal image collection summaries* in which both text and visual content are combined in the same latent semantic space. Most of the proposed summarization methods found in the literature extract visual features such as color, texture and edges to represent image content, which use clustering algorithms to perform the summarization process. However, images are commonly accompanied of other information sources such as text, audio, links, etc. Therefore, in the same way that visual content is used to represent images, these additional information sources (modalities) can be also used to better represent the image semantics. The proposed method also provides a mechanism to project images that do not have associated text, which addresses the problem of images that are not accessible because of the lack of text information associated to them. Notwithstanding that user satisfaction studies are widely used to evaluate summarization algorithms, we favored more objective and quantitative evaluation metrics for assessing the performance of the proposed summarization method. Consequently, the chapter also presents a method to quantitatively measuring the quality of an image collection summary by estimating its ability to reconstruct the complete collection and its diverseness ability.

This chapter is organized as follows: Section 2 describes related work; Section 3 presents the proposed method; Section 4 presents experimental evaluation of the proposed strategy; and finally, Section 5 concludes the chapter.

3.2. Image collection summarization

Image collection summarization is the process that aims to select a small set of the most representative images (summary) to highlight a larger set. This process has became very important to enable interactive navigation and exploration of large-scale image collections [58]. Many applications can benefit from results of image collection summarization: (1) Image search engines such as Google Similar Images, Flickr and Yahoo, which use the conventional page-based navigation paradigm; (2) On-line shopping web sites, which visualize representative images for each product category; (3) Personal photo collection systems, which automatically organize image categories based on metadata information. Such applications have motivated researchers to investigate new summarization methods that allow users to explore large image collections in a efficient and intuitive way [59].

We propose an approach in which the summarization process fuses visual and textual content of the images. That is, an image can be modeled taking into account its visual (color, edges, textures) and textual (tags, caption, descriptions, etc) properties, which are called 'modalities' in this thesis. This fusion process allows to combine visual and textual content in the same space to produce a more semantic representation.

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ $\mathbb{R}^{l\times n}$ be an image collection. Let $S = \{s_1, s_2, \dots, s_m\}$ $\mathbb{R}^{l\times m}$ be a subset of X, S X. When m is small with respect to n, m n, S is said to be a summary of X. A good summary is expected to include images which are representative of the collection content. The representativeness must be given in terms of two main aspects, the visual content and the semantic content. For instance, a good summary of an image collection produced as the result of submitting the query 'apple' to Flicker, as depicted in Figure 3.1, must contain images corresponding to different senses of the word 'apple' (fruit, brand, city, etc), but also must contain images illustrating the visual variability in the collection (green apples, red apple, white logo, etc). In order to objectively evaluate the goodness of a summary we use two performance measures, diversity score and reconstruction error. Both measures are computed over a multimodal representation of the summary, i.e., taking into account both the visual and semantic richness of the collection.

Diversity score We define diversity of a summary as a score that measures how different are the images of a summary in terms of textual and visual content. A high diversity value means that a summary contains dissimilar images, which is a good indicator of topic collection coverage. On the contrary, a low diversity value means that a summary contains redundant images, which causes a low topic coverage.

Formally, diversity of a summary S is defined as the summation of distances among all the elements of the summary as follows:

$$Diversity(S) = \sum_{i=1}^{r}\sum_{j=1}^{r} d(s_i, s_j), \tag{3.1}$$

where,

$$d(S_i, S_j) = d(S_i^v, S_j^v) + d(S_i^t, S_j^t), \tag{3.2}$$

where $d(S_i^v, S_j^v)$ is the distance between the i-th and j-th elements of the summary taking into account its visual representation, and $d(S_i^t, S_j^t)$ is the distance between the i-th and j-th elements of the summary taking into account its textual representation. This metric is measuring the multimodal distance between the i-th element and j-th element of the summary. There are different distance measures that can be used in Equation 3.2. However, we model each modality using a bag-of-words model, which produces a representation interpreted as a probability distribution function (p.d.f). The details of this representation process is presented in Section 3.4.1. As a result of this process each modality (visual and textual) is represented as a p.d.f. One of the most commonly distance function used to compare p.d.f is the Kullback-Leibler Divergence (KLD), so we selected it in the experimental evaluation conducted in this chapter.

Reconstruction error A good summary is expected to include the representative visual patterns in the collection, i.e., the main building blocks of the collection visual content. An approach to measure if these building blocks are present in the summary is to attempt to represent the full collection only using the visual content of the summary. In this context, we can think of the summary as a basis where the images in the collection are represented. Formally, we can attempt to represent an image x_i as a linear combination of the images in the summary:

$$x_i = \sum_{j=1}^{m} h_{ji} s_j,$$

where h_{ji} is a set of coefficients.

The reconstruction error measures how good this representation is.Therefore, we minimize the error produced when x_i is reconstructed as follows:

$$\mathit{Reconstruction}(S) = \min_{H \geq 0} \sum_{i=1}^{l} \sum_{s_j \in S} (x_i \quad h_{ji} s_j)^2, \tag{3.3}$$

where S is the image subset (summary) that represents the image collection and H contains the weights to linearly combine elements of the summary to reconstruct the visual content of the collection. In principle, this minimization problem could be solved using some of the algorithms proposed in [60], which solve the more general non-negative matrix factorization problem:

$$\min_{H \geq 0} \; X \quad HS \;{}^2_F \quad , \tag{3.4}$$

where S is the matrix representation of a summary and X is the matrix representation of the complete image collection.

3.3. Related work

Different methods have been proposed to build image collection summaries including methods based on clustering [23, 24, 25], similarity pyramids methods [26], graph methods [28, 29], neural networks methods [61], formal concept analysis [30], and kernel methods [32], among others. In most of the cases, the summarization problem is approached as a non-supervised learning problem. Typically, image clusters are identified in the collection and representative images from each cluster are chosen to compose the summary.

Conventionally, the quality of a summary is calculated by using quality measures such as entropy, separation, cohesion, purity, among others [44]. Other measures commonly used on information retrieval such as precision, recall, f-measure and MAP are adapted to evaluate the impact of the summarization process in image retrieval tasks [62, 63].

Diversity is a characteristic of the results obtained in a retrieval system, which is related to the variety of documents representing different semantic concepts of the collection. This concept comes from the information retrieval area [45], where one of the main concerns is to balance diversity and relevance in search results [64, 65]. Whilst relevance is focused on measure whether images are relevant to the user, diversity is associated to the capacity of a system to show results that reflect the user's complete spectrum of interests [46].

Image result diversification aims to diversify image results in an image retrieval system and is used as mechanism in image collection summarization. In this context, some works such as described in [44] and in [66, 46] propose summarization models where a diverse set of images is evaluated.

Hybrid summarization methods such as the proposed in [67, 68, 32] compute image summaries by combining visual and semantic content. The main characteristic of these summarization methods is that the summary is obtained by clustering in a separately way both modalities and then combining them to obtain a final set of clusters.

Most of the existent research works in image collection summarization have two problems: (1) They use of only visual content to represent images, which does not take into account other information sources that can enrich image representation; and (2) The evaluation of the summarization is performed subjectively by using user satisfaction questionnaires, which does not allow to measure quantitatively how good is the summarization results.

In this chapter, we propose a strategy based on latent topics in which visual and text modalities are jointly used to better model image semantics in the summarization process. We also propose to assess the performance of the summarization process with objective evaluation metrics:

The *K-Medoids algorithm* [69] is a typical clustering-based method in which k clusters are obtained from which its corresponding medoids are selected as summary. This algorithm is commonly used in the literature as a baseline for comparing purposes.

The *Affinity propagation algorithm (AP)* [70] takes as input the similarity between pair of data points and iteratively exchange messages between such data points until a good set of exemplars are selected. The AP algorithm is being used as a state-of-the-art method for image collection summarization [59, 68, 71] and result diversification [72, 73].

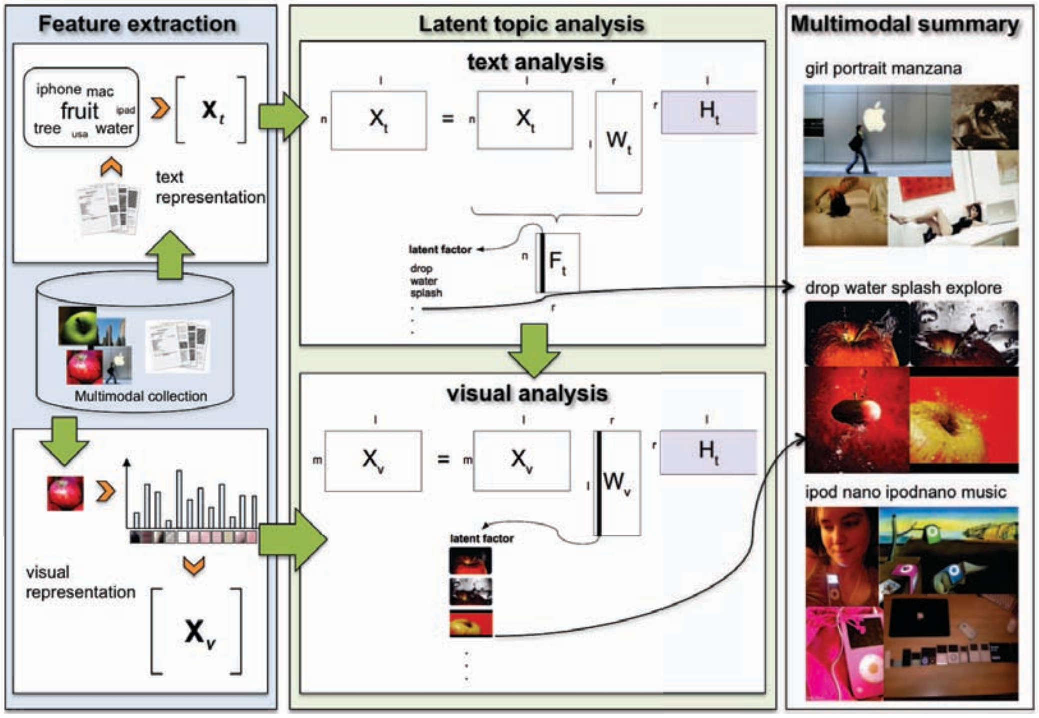


Figure 3.2.: Overview of the proposed method, which is composed of three main stages: feature extraction, latent topic analysis, and summary construction.

3.4. MICS: Multimodal image collection summarization

This chapter proposes a method for automatically building a summary that combines visual and textual content. The main goal of the proposed method is to obtain a multimodal summary that represents the complete collection. This process is composed of three stages: feature extraction, latent topic analysis, and summary construction. The overall process of building an image collection summary is illustrated in Figure 3.2.

In the f*eature extraction stage*, each image and its associated tags is processed to build a visual and text representation. This representation stage produces two matrices, X_t and X_v, where all images in the collection are indexed taking into account textual and visual content respectively. In the *latent topic analysis stage*, the obtained matrices are processed to obtain a latent representation in which both modalities are fused in the same space. In this new space each image is represented as a linear combination of latent factors coefficients and the membership level of each image to the corresponding latent factor. In the *summary construction stage*, a set of images and terms are selected to produce a multimodal summary from the latent factors found in the previous stare.

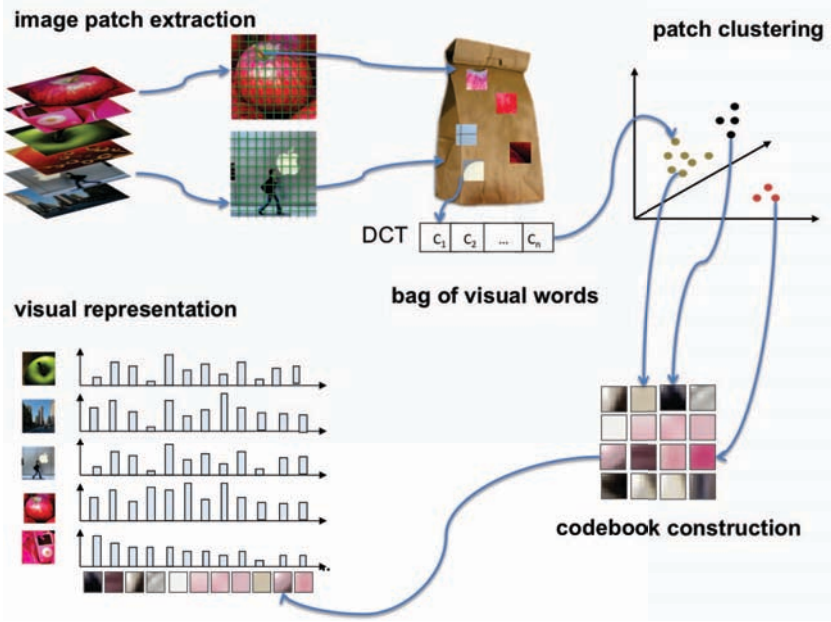


Figure 3.3.: Visual image representation following a bag of features model.

3.4.1. Image representation

In this work, we follow a bag-of-words (BoW) approach to represent image content, which is one of the most common representations used in text-mining and information retrieval (IR). We use a common scheme for representing both text and visual content since we are dealing with multimodal objects.

3.4.1.1. Visual representation

The visual content is represented using a bag-of-features [74] approach in which an orderless distribution of image features is constructed based on a predefined dictionary of visual patterns. As illustrated in Figure 3.3, this dictionary is built from the image collection and accounts for the occurrence of each visual pattern in the images. The bag of features is constructed using a training set of images that are split in sub-blocks of n×n pixels. A visual feature is extracted for each sub-block, to represent the associated visual patterns using rotation invariant properties. Then, all the extracted blocks are clustered to obtain k centroids, which are used as a reference dictionary of visual patterns. Finally, we build a histogram with the occurrences of visual patterns found in the image. This scheme is widely used in computer vision tasks, including image categorization and object recognition [75].

3.4.1.2. Text representation

To represent the tags associated to each image, we follow the Vector Space Model (VSM) [76]. This model is based on a vector representation, where each component of a vector indicates the frequency of a word (term) in the document. Formally, a document is expressed as $x_i = (x_{1,i}, x_{2,i}, x_{3,i}, \ldots, x_{m,i})$, where $w_{t,i} = \mathrm{tf}_t \cdot log\frac{|D|}{|\{t \in i\}|}$, tf_t is the term frequency of the term t in the document i, $|D|$ is the number of documents in the collection, and $log\frac{|D|}{|\{t \in i\}|}$ is the inverse frequency of the documents that contain t.

3.4.2. Latent topic analysis and summary construction

Let $X_t \quad \mathbb{R}^{m\times\ell}$ be the matrix that contains all the vectors representing the textual content of an image collection with l images, and let $X_v \quad \mathbb{R}^{n\times\ell}$ be the matrix that contains the vectors representing the visual content. The general summarization problem consists in finding a subset of images that is representative of both the visual content of the collection and the semantic (textual) content associated to them. This problem is addressed through two main strategies: (1) a strategy to find a small subset of images that is a representative of the visual content of the collection (X_v), and (2) a strategy to associate these images with the semantic classes found in the textual content (X_t). The two strategies are discussed in the following subsections.

3.4.2.1. Latent topic analysis

The goal of latent topic analysis is to discover latent "topics" that occur in a collection of documents. Intuitively, given that an image is about a particular topic, one would expect particular textual and visual terms to appear in the image. An image typically concerns multiple topics in different proportions. A topic model captures this intuition in a mathematical framework, which allows examining a set of images and discovering, based on the statistics of visual and textual terms in each, what the topics might be and what each image's balance of topic is.

A popular approach to latent topic analysis is to apply singular value decomposition (SVD). The problem of applying SVD to find the latent space, is that the codifying vectors could have, in general, negative values. This means that some documents are represented, not only by the presence of latent factors, but also by their absence.

To solve this problem, an additional restriction may be imposed on the basis vectors and codifying vectors to force all their entries to be positive. This is called Non-negative Matrix Factorization (NMF). Notice that in the case of SVD, the restriction is that the basis vectors must be orthogonal. This is accomplished by the Eigendecomposition. In NMF this restriction does not apply, and this in turn can have tremendous benefits in the semantic representation since, intuitively, different concepts or clusters do "not" have to be orthogonal or non-overlapping with each other. There are different ways to find an NMF [60].

Let $X = (x_1, \ldots, x_\ell)$ be the input data matrix that contains a collection of n data vectors as columns. We consider factorizations of the form:

$$X_{\pm} \quad F_{\pm}H_{+},$$

where $X \quad \mathbb{R}^{n\times\ell}$,$F \quad \mathbb{R}^{n\times r}$, and $H \quad \mathbb{R}^{r\times\ell}$, with $H \quad 0$. For reasons of interpretability it is useful to impose the constraint that the vectors defining F lie within the column space of X: $f_r = w_{1r}x_1 + \cdots + w_{\ell r}x_\ell = Xw_r$, or $F = XW$.

3.4.2.2. Summary construction

The image database is composed of two data modalities, herein denoted by X_v and X_t. The former is a matrix whose rows are indexed by n visual features and whose columns are indexed by l images. The latter has m rows to represent text terms and l columns for images as well. The construction of a latent semantic space may be done by decomposing the matrix of images that can have only visual features or only text annotations. However, to generate a semantic space for image indexing, we are interested in exploiting multimodal relationships between images and text. Figure 3.2 presents an overview of proposed summarization process. The image collection summarization method is presented in Algorithm 3.1 and described in the following paragraphs.

In a first step (line 04) , the textual matrix X_t is decomposed using convex NMF as follows,

$$\min_{W_t,H_t\geq 0} \quad X_t \qquad X_tW_tH_t \ {}^2_F \quad . \tag{3.5}$$

This has an interesting byproduct, the columns of $F_t := X_tW_t$ could be interpreted as clusters centroids that represent clusters among image tags which could be eventually associated to high-level semantic concepts. This in fact generates a new representation of the textual data, which is in fact a latent semantic representation. This process is applied to both visual and textual data to obtain a multimodal latent representation of the image collection.

In a second step, we fix H_t and apply CNMF to factorize the visual matrix:

$$\min_{W_v,H_t\geq 0} \quad X_t \qquad X_tW_tH_t \ {}^2_F \tag{3.6}$$

where $X_v \quad \mathbb{R}^{m\times l}$, $W_v \quad \mathbb{R}^{l\times r}$, and $F_v \quad \mathbb{R}^{m\times r}$. This is expressed in line 05 of the MICS algorithm. It is important to note that in this factorization we fix H_t to find W_v, therefore both information sources are combined in this step obtaining a new factorization that depends on text and visual information at the same time. In each column of W_v we find the importance degree of each image in the i-th cluster. As result of the optimization process we obtain W_v (line 05), which is used jointly with W_t to create the i-th multimodal cluster φ_i. Note that cell scores of each topic W_v^i are independently sorted to select the n most important (line 08). Each cell score represents the relative weight of each image to the i-th topic, that is to say, each cell score indicates the level of importance of each image in the i-th latent topic. Once the top cell scores are obtained for each topic, the corresponding images are added to the summary φ.

It is worth noting also that in the algorithm, H_t is the basis of the latent space in which an image is represented as a linear combination of the r columns of $F_v = W_v \cdot X_v$. The corresponding coefficients of the combination are codified in the columns of H_t. As it is shown in [63], each column of F_v corresponds to a cluster of the original objects, and each column of H_t corresponds to an object represented in the latent space. Thus, in each column of H_t we find the membership degree of each text term in the i-th cluster.

Algorithm 3.1 Multimodal image collection summarization (MICS) algorithm.

```
Input: Text matrix $X_t$, Visual matrix $X_v$, number of topics $k$, number of images per topic $n$
Output: Summary $\varphi$

begin
    $W_t, H_t = \arg\min_{W_t,H_t} ||X_t \quad X_t W_t H_t||_F^2$
    $W_v = \arg\min_{W_v} ||X_v \quad X_v W_v H_t||_F^2$
    $W \quad \{W_v^1, W_v^2, W_v^3, ..., W_v^k\}$
    for $i := 1$ to $k$, do
        $\varphi_i \quad n$ top images in $W_{:,i}^{(v)}$
    end for
    $\varphi = \varphi_1 \quad \varphi_2 \quad \varphi_3 \cdots \quad \varphi_n$
end
return $\varphi$
```

3.4.2.3. Topic textual description

The proposed summarization method described in Algorithm 3.1 allows to select textual terms in each latent topic. That is to say, $F_t = X_t W_t$ has in each cell the level of importance of each textual term in the i-th topic. A cell score in F_t indicates the relative weight of each term in each latent topic, so if they are sorted independently, they can be used to select the most important terms in each cluster (the highest cell scores). Figure 3.5 shows an example of terms obtained in each latent topic.

3.5. Experimental evaluation

The main goal of the experimentation conducted in this chapter was to objectively evaluate the proposed method and compare it to other state-of-the-art automatic summarization methods. To reach this goal, we use two image data sets crawled from Flickr, and two performance measures to evaluate the performance of the proposed method.

3.5.1. Datasets

In our experiments we use the MIRFlickr [77] and the Flickr4Concepts datasets, which are image collections crawled from Flickr.

MIRFlickr The MirFlickr data set that contains 25.000 images collected by downloading images from Flickr over a period of 15 months. The collection contains images that were ranked by the Flickr's interestingness rating. These images were manually annotated using 38 categories. Each image has 8 tags on average, which is used as the text modality. The complete collection has 1386 tags, which are in

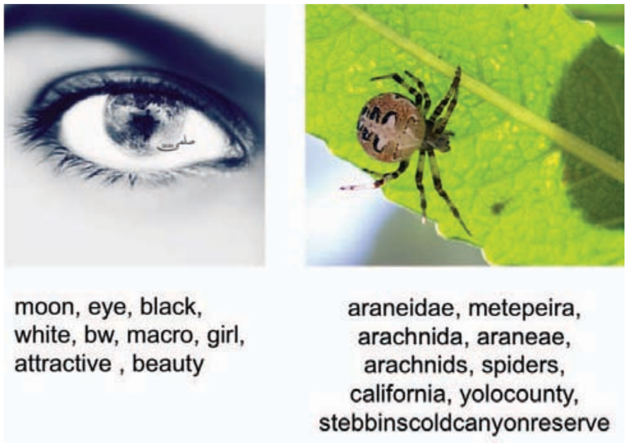


Figure 3.4.: Example of images and its associated tags of the Flickr4Concepts (left) and MirFlickr (right) data sets.

Table 3.1.: The Flickr4Concepts datasets. Four different one-term queries were used and each dataset corresponds to images returned from Flickr when queried with the corresponding term. The last column specifies the number of different text terms found in the dataset after pre-processing

Dataset	Number of images	Terms
apple	1263	837
love	724	666
closeup	1405	995
beauty	1490	1195

English and other languages. Figure 3.4 shows an example of a *beauty* image and its associated tags.

Flickr4Concepts We crawled 4882 images from Flickr and their associated tags for *apple*, *love*, *closeup*, and *beauty* query terms, as described in Table 3.1. These terms were chosen because they have different meanings depending on the context, which makes challenging the summarization process. For instance, the term *beauty* can be found in images of women, nature, cats, flowers, etc. Figure 3.4 (left) shows one of image and its associated tags of this data set. The associated terms are useful to see whether a summary can discriminate different semantic sub-groups of images that belong to the same concept. We select 4 concepts for this work, but the proposed method can be applied to more concepts and larger datasets. Table 3.1 describes the number of terms indexed from the Flickr4concepts dataset after a pre-processing step.

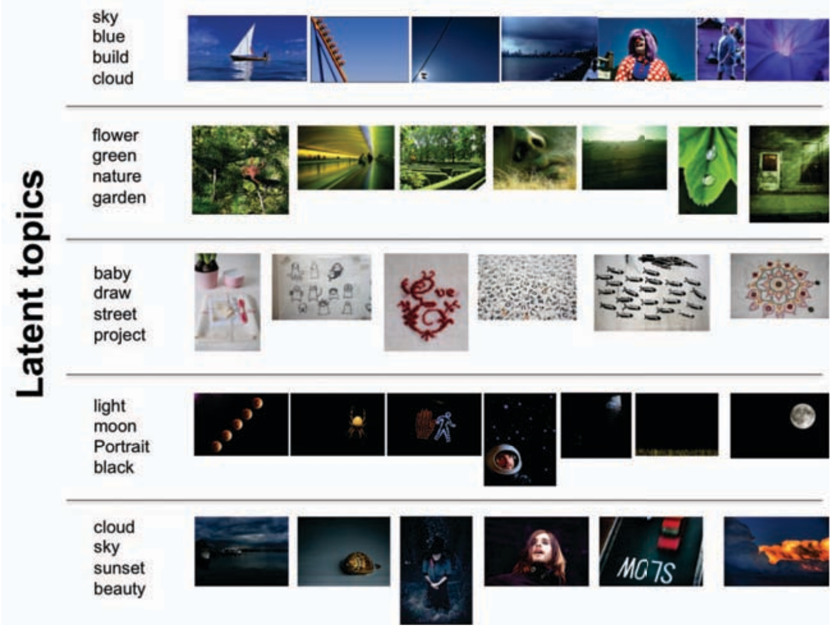


Figure 3.5.: Latent topics obtained from MIRFlickr. The proposed method automatically founds images and terms that are representative of the collection.

3.5.2. Image features

Visual content was indexed using the bag of visual words method presented in Section 3.4.1.1. Each image was split into patches of 8x8 pixels. The DCT (Discrete Cosine Transform) descriptor [78] was used to index each patch. The parameter of the patch clustering process k was set to 1000, thus we obtained a dictionary of 1000 patches. Finally, each image was indexed using a histogram of 1000 bins.

Text content was indexed using the vector space model presented in Section 3.4.1.2. We applied stop-words removal and stemming.

3.5.3. Summarization results

Figure 3.5 shows an example of the latent topics obtained from the MirFlickr data set. In this illustration, each latent topic contains the most representative images and terms of the corresponding latent topic. Note how each latent topic groups visual and text content in the same semantic space. For instance, sky concept is related to blue color, and black is related to night and moon concepts. This result is particularly interesting because each latent topic is expressing a type of image category automatically captured from the collection.

As it was described in Section 3.4.2, a multimodal summary is built by selecting the most important text terms and images of each latent topic. We selected 4 images and 4 terms to illustrate the method, but this is a parameter can be set according to the visualization metaphor used to display results.

Figure 3.6 shows some multimodal latent factors for the *apple* concept of the Flickr4Concepts data set. The number in each cluster represents the rank assigned

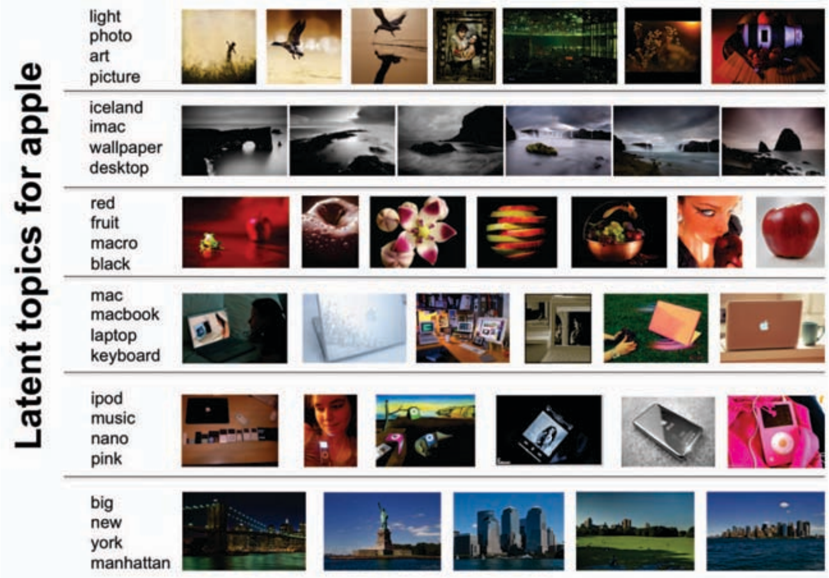


Figure 3.6.: Example of multimodal latent factors for the *apple* concept.

according to its importance (the weight of the cluster). Note that each cluster illustrated in Figure 3.6 captures representative images and text terms of the concept *apple* for different contexts. For instance, there are clusters for fruits, computer elements, buildings, and apple trees. This result is useful since users can see different semantic concepts that match with the query, so users have the opportunity of explore the subset that is more related to their needs.

Figure 3.7 shows the obtained multimodal latent topics for the *beauty* concept. In this case the clusters represent different high-level concepts of beauty. This concept is very subjective, so users annotate with this term images of flowers, animals, women, and art pictures.

Figure 3.8 shows the results for the multimodal latent topic obtained for the *close-up* concept. In the obtained summary there are images of people, flowers, insects, and animals. It is worth noting that this summary provides a good set of representative images of different topics.

Finally, Figure 3.9 shows the obtained latent topics for the *love* concept. Images of nature, babies, marriages, couples, and animals were automatically grouped.

It is worth to note that the proposed method is robust to junk images. In the matrix factorization process, each latent topic (columns of W_v and F_t) is representing a cluster in the latent space. In each cluster we select the most important images and concepts according to their importance in the respective cluster. This mechanism allows to select images that are representative of the collection and penalize junk images.

3.5.4. Performance evaluation of the summary

We are interested in objectively evaluating the quality of the summarization process. To do this, we use the performance measures defined in Section 3.2 and show the

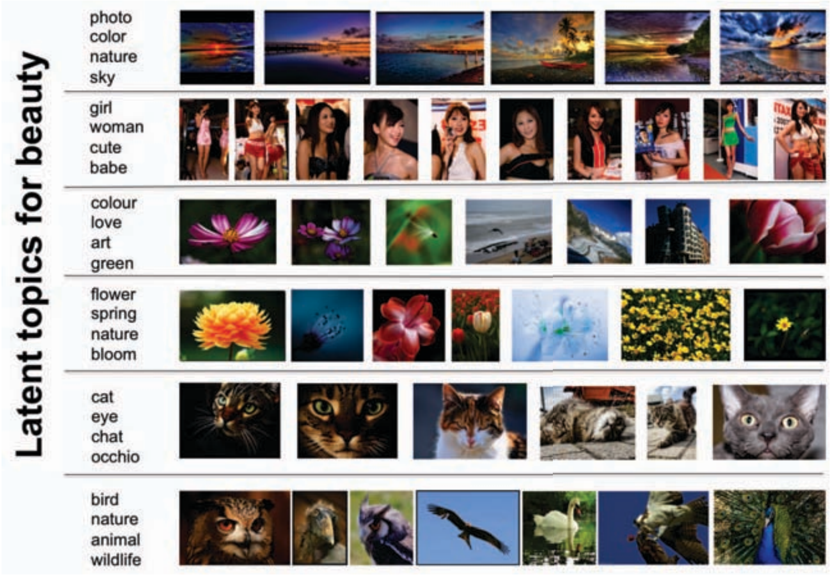


Figure 3.7.: Example of multimodal latent topics for the *beauty* concept.

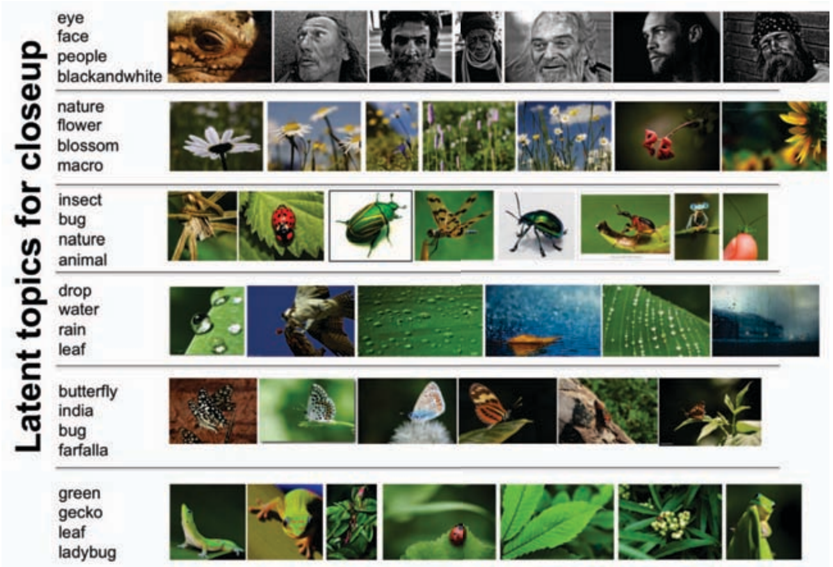


Figure 3.8.: Example of multimodal latent topics for the *closeup* concept.

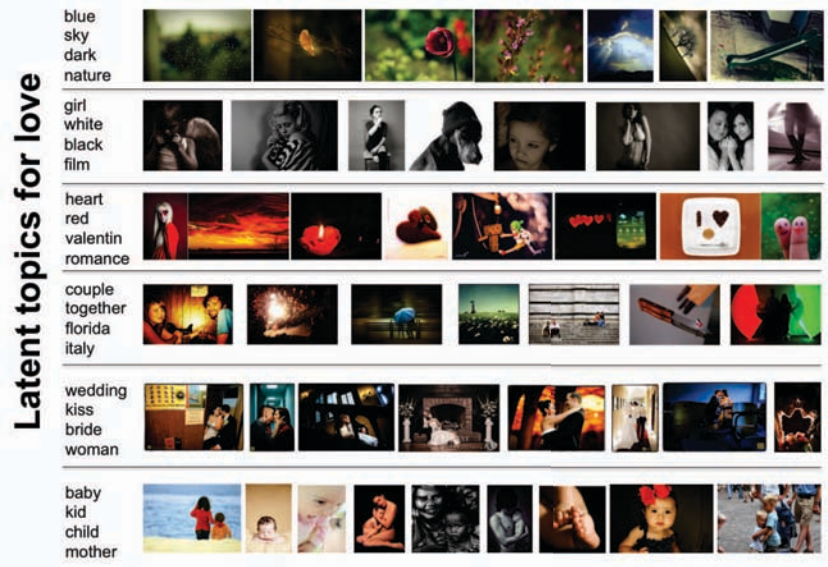


Figure 3.9.: Example of multimodal latent topics for the *love* concept.

obtained results in the following subsections.

3.5.4.1. Results of reconstruction error

In this analysis we evaluate the reconstruction error for different summary sizes of each dataset. A low error value indicates that the obtained summary is able to reconstruct the whole collection in a more precise way. Figure 3.10 shows the obtained plots for this performance measure. Table 3.2(a) shows the error average for each concept of the Flickr4Concepts. The Affinity Propagation algorithm obtains the lowest error when the summary size is lower than 50 latent topics. However, the MICS algorithm reaches the lowest error when the summary size is higher than 50 latent topics. In the case of the MirFlickr data set, the lowest error is obtained for our method in all the summary sizes.

3.5.4.2. Results of summary diversification

In this analysis we evaluate the ability of each method to build diverse summaries. A high value indicates that the summary is more diverse, that is to say, images that compose the summary cover in a better way the input image space. Figure 3.10 shows the obtained results for the Flickr4Concepts data set. The MICS algorithm outperforms the other three baseline methods for both the MirFlickr and Flickr4Concepts data sets. This fact can be attributed to the ability of the proposed method to select images from latent factors, which covers the complete image collection space. Table 3.3 shows the diversity average of the MirFickr and Flickr4Concepts data sets.

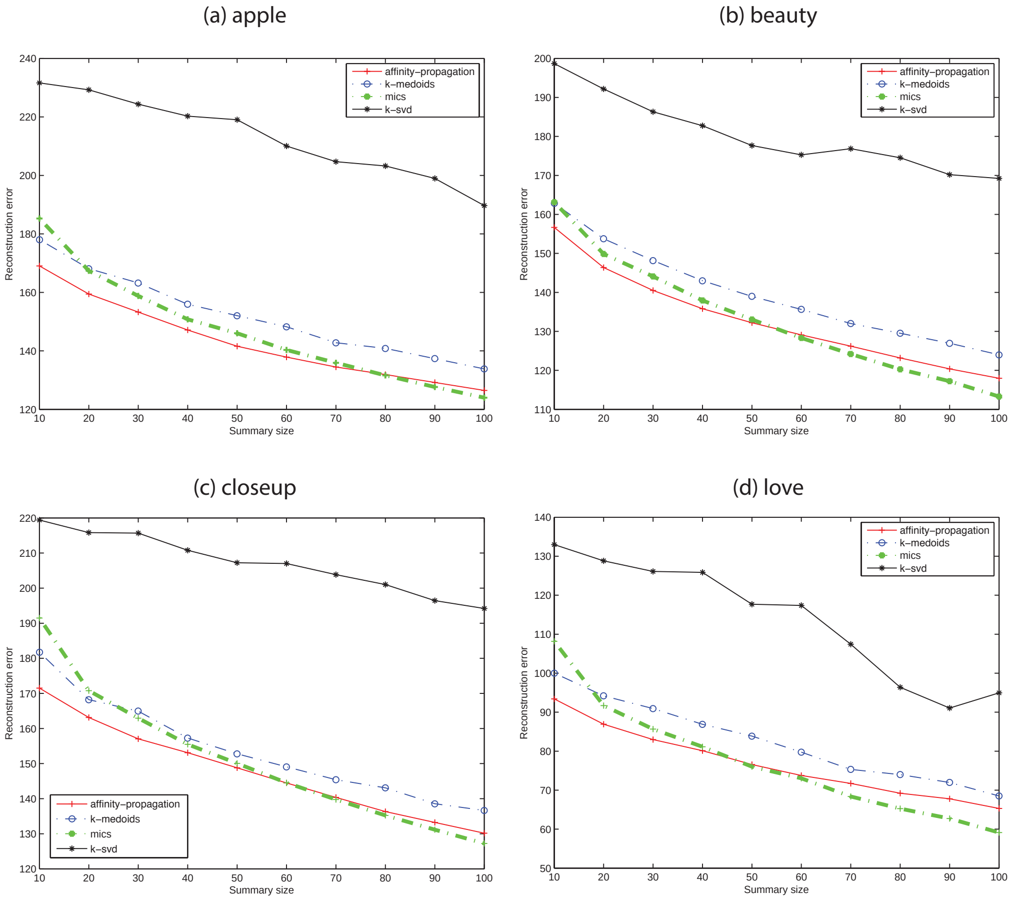


Figure 3.10.: Reconstruction error for (a) apple, (b) beauty, (c) closeup, and (d) love concepts of the Flickr4Concepts data set.

Table 3.2.: Reconstruction error for Affinity Propagation (AP), K-Medoids, and the proposed method (MICS). Columns correspond to different summary sizes. Columns corresponds to different summary sizes. The best performance values are shown in bold.

(a) Flickr4Concepts (average for the 4 concepts)

	10	20	30	40	50	60	70	80	90	100
AP	**147.6**	**138.9**	**133.4**	**129.0**	**124.7**	121.3	118.2	115.2	112.5	109.9
K-Medoids	156.2	148.3	138.6	137.3	132.5	127.3	124.0	121.3	118.7	115.4
MICS	156.2	146.9	136.7	131.2	125.8	**121.2**	**118.0**	**112.8**	**109.1**	**106.1**

(b) MIRFlickr

	10	20	30	40	50	60	70	80	90	100
AP	168.9	161.6	151.4	146.2	138.6	131.9	128.2	122.4	116.9	110.6
K-Medoids	173.2	165.0	159.9	155.2	146.7	142.5	138.4	133.2	123.9	114.4
MICS	**156.2**	**146.9**	**136.7**	**131.2**	**125.8**	**121.2**	**118.0**	**112.8**	**109.1**	**106.1**

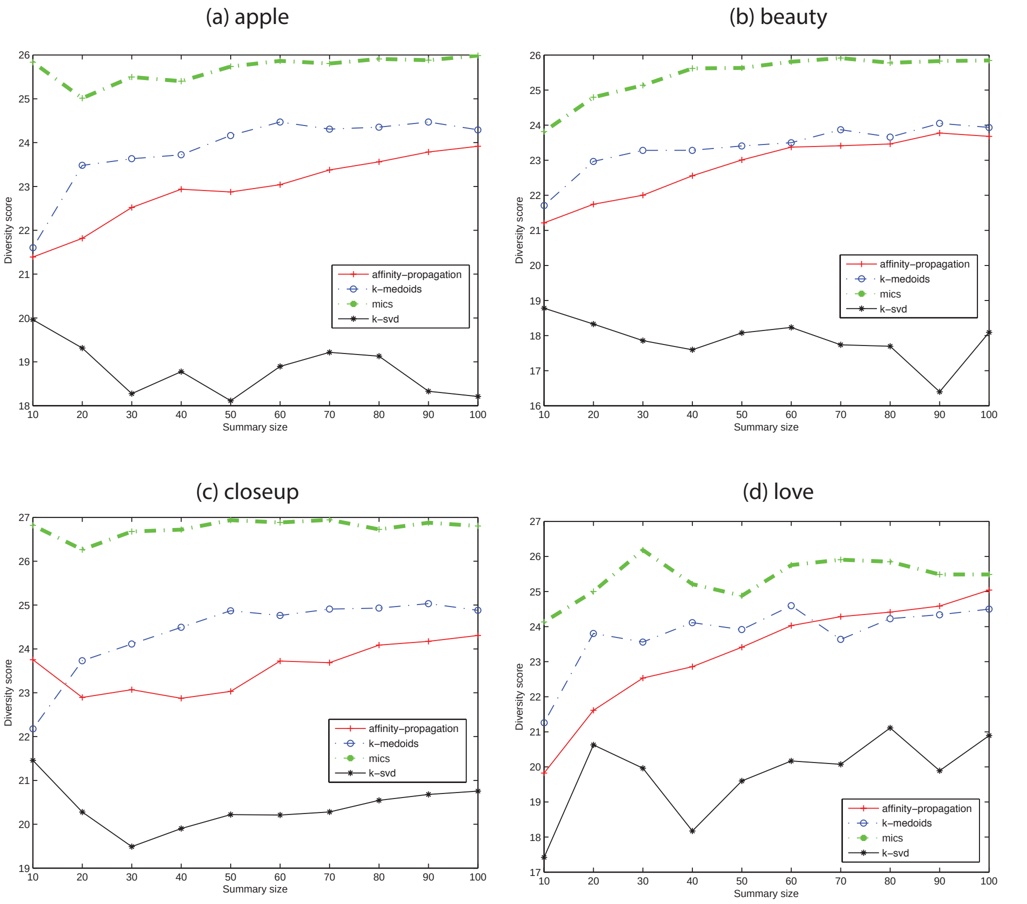


Figure 3.11.: Diversity score of the summary for (a) apple, (b) beauty, (c) closeup, and (d) love concepts of the Flicr4Concepts data set.

Table 3.3.: Diversity scores obtained for Affinity Propagation (AP), K-Medoids, and the proposed method (MICS). Columns correspond to different summary sizes. The best performance values are shown in bold.

(a) Flickr4Concepts (average of the 4 concepts)										
	10	20	30	40	50	60	70	80	90	100
AP	21.54	22.01	22.53	22.80	23.08	23.54	23.67	23.87	24.08	24.25
K-Medoids	23.46	23.39	23.77	23.77	23.99	24.36	24.36	24.12	24.31	24.52
MICS	**23.99**	**25.02**	**25.12**	**26.05**	**26.05**	**26.19**	**25.96**	**26.19**	**25.96**	**25.95**

(a) MIRFlickr										
	10	20	30	40	50	60	70	80	90	100
AP	20.95	20.32	20.22	20.36	20.42	20.67	20.73	20.78	20.87	20.84
K-Medoids	**21.62**	21.01	20.98	21.09	21.00	20.69	20.88	20.96	21.50	21.14
MICS	21.36	**22.03**	**22.10**	**21.97**	**21.47**	**21.89**	**21.60**	**21.66**	**21.78**	**21.52**

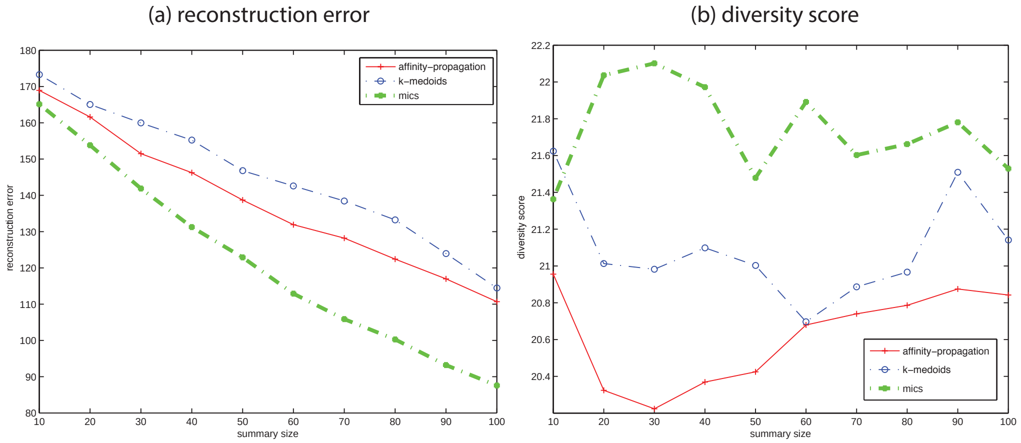


Figure 3.12.: Diversity score (a) and reconstruction error (b) of the summary for the MirFlickr data set.

Figure 3.12 shows the reconstruction error and diversity when the summary size is increased in the MirFlickr data set. The MICS algorithm outperforms the baseline algorithms. This result can be observed in Table 3.2 (b) and Table 3.3 (b) respectively.

3.6. Discussion

The main difference between the proposed method and the baseline algorithms is the way in which the summary is obtained. That is, the baseline methods use a clustering-based approach where summary elements are generated from closest images to the obtained centroid of each cluster. Instead, we perform a latent topic analysis in which summary elements are selected from the latent factors obtained in the matrix factorization process. This latent representation has two very important properties: (1) the latent space induces a sparse representation, which directly influences the diversity of the summary; (2) the summary is built from the original images, which avoid the selection of elements based on linear combinations of the input image dataset or cluster centroids.

Results show how the proposed method in general outperforms the baseline methods in terms of reconstruction ability and result diversification. In the Flickr4Concepts data set we observed that when the summary size is lower that 50, the AP algorithm performed slightly better than our proposed method in terms of reconstruction error. However, the AP algorithm obtains the worst performance in terms of diversity, which indicates that although the obtained summary has a good ability to reconstruct the image collection, it contains redundant images. In contrast, our algorithm maintains the best performance when the diversity is evaluated in all the experimented summary sizes.

3.7. Conclusion

We have presented a new method to build multimodal collection summaries based on latent topic analysis. We proposed the MICS algorithm, which uses a mechanism to fuse text and visual information in the same latent semantic space to better model image semantics. This method allows to build semantic summaries that involve text and visual content in the construction of the summary. The proposed method was applied to four image collections extracted from Flickr. We also used objective measures to validate the performance of the proposed method in the construction of multimodal image collection summaries. Results are encouraging and show the feasibility of using this method to offer the user more diverse and semantic results when interacting with image collection exploration systems.

Part II.

Visualization

4. Multimodal Image Collection Visualization Based on Non-negative Matrix Factorization

The work presented in this chapter was published at the 14th European Conference on Research and Advanced Technology for Digital Libraries [79].

Image collection visualization is an important component of exploration-based image retrieval systems. In this chapter we address the problem of generating an image collection visualization in which images and text can be projected together. Given a collection of images with attached text annotations, we aim to find a common representation for both information sources to model latent correlations among the collection. Using the proposed latent representation, an image collection visualization is built, in which both data modalities (images and text) can be projected simultaneously. The resulting image visualization allows to identify the relationships between images and text terms, allowing to understand the distribution of the collection in a more semantic way. The resulting visualization scheme can be used as the core metaphor in an interactive image exploration system.

4.1. Introduction

Multimedia content production, including documents with text, images, videos and audio, has experienced an exponential growth thanks to the development of computer systems and communications. The process of finding and accessing this vast volume of information requires effective and efficient computational methods. In the case of text, a lot of work has been done by the information retrieval community, and thanks to its development, today we have suitable text document search engines that allow us to easily find information. However, in the case of other type of multimedia content, such as images, the results are not as satisfactory yet. New mechanisms to explore large image collections are necessary to offer the user different alternatives for accessing and finding information.

Image collection exploration has been shown to be a promising strategy for image retrieval [48]. In this strategy, the user interacts with the system while it learns from the user's feedback to delivery more precise results. Image collection visualization plays an important role in this process. To construct the visualization, an image representation is projected into a two dimensional space, in which images with similar features are mapped to neighboring positions. This mapping allows the user to see image inter-relationships and to easily identify how they interact. In this way, it is expected that users can access images with similar properties in the same region of the screen.

The visualization may be built using the visual content of the image to organize the image collection according to some visual similarity properties, for instance using colors and textures. However, low-level features are usually not enough to establish a meaningful criteria for image search due to the semantic gap, i.e. the discrepancy between computed features and human interpretation of images [80]. Thus, images with similar low-level visual features may appear in the same region of the screen even though they represent different semantic concepts. To provide a more semantic and organized visualization, some learning approaches have been proposed to adapt the position of images in the screen according to the user's preferences [81] or some predefined semantic categories in the collection [82]. However, these schemes have two limitations that prevent them from being used for analyzing massive image collections. First, they do not allow to clearly identify the underlying semantic structure of the collection, since these algorithms allow to project only visual information. Second, these learning algorithms rely on user's feedback or structured metadata to learn the semantic organization of images, requiring expensive efforts to collect user profiles or reliable annotations.

Most of the systems for image collection exploration use mainly visual features to generate the visualization of the collection, ignoring other possible information sources that may complement the image representation from a semantic viewpoint, such as unstructured text. The problem addressed in this chapter is the design of a strategy to construct an image collection visualization using both, visual features and available text data.

In this chapter we propose a model to construct an image collection visualization using the two mentioned information sources (modalities) in the same representation space. We propose the construction of a latent multimodal space in which visual features and text terms can be represented together. This multimodal representation is built using Non-negative Matrix Factorization (NMF) algorithms to compute a latent space for images that can be spanned using either text terms or visual features. The location of text terms and visual features can be identified in the latent space, bringing a unified way to analyze the relationships between both modalities. This latent multimodal representation is then used to project both, text terms and images in the same 2D plane to construct an image collection visualization with a semantic organization given by text data and at the same time marks the regions of the screen in which semantic concepts can be found.

The organization of this chapter is as follows: Section 2 presents a discussion of some related work. Section 3 presents the NMF algorithms and visualization methods. The experimental evaluation and results are presented in Section 4. Finally, Section 5 presents the conclusions and future work.

4.2. Related work

Image collection visualization is an active area of research whose main aim is to offer intuitive and efficient mechanisms to visualize image collections. Most of the visualization approaches reported in the literature are *unimodal*, that is, only one modality (visual content) is shown in the visualization metaphor. One of the most used approaches is 2D similarity-based visualization, where images are placed within 2D canvas based on their mutual visual similarities. A comprehensive review of state-of-the-art methods for image collection visualization and browsing models can be found in [83] and [48].

Image collection visualization is typically associated with image collection exploration and retrieval. Google Image Swirl [84] and Ostensive Image Retrieval [49] are examples of image collection exploration methods that use some sort of visualization mechanism. In both systems, the user is shown a set of piles of images organized according to clusters of similar images. Once the user selects one pile, it goes to the front (expanded in a circle) whilst the rest goes to the background. While the user interacts with the collection, the navigation path is visualized to the user in a spiral shape allowing the user to get back to previous navigation points. In both cases, even though the collection has additional data modalities, the visualization only involves visual features.

The use of text data for image retrieval has been investigated by Jeon et al. [85], who modeled the relationships between image features and text terms using cross media relevance models. Later, Hare et al. [86] and Rasiwasia et al. [87], have extended the idea of including the semantic information of text data in the image representation using latent semantic indexing methods. They used a common indexing technique for visual features and text terms, so the system provides a mechanism to find images through keywords even though they do not necessarily have attached text. These approaches suggest that the correlation between images and unstructured text can be a powerful mechanism to construct new image search solutions.

Little work has been done in multimodal image collection visualization. In this chapter we propose to build a *multimodal* visualization that includes both visual and textual content in the same visualization space, such that the visualization is more semantic and allows the user to explore the image collection in an intuitive and efficient way.

4.3. Multimodal image collection visualization

The proposed framework consists of two main operations: to build a multimodal image representation and to project images and terms into a 2D canvas. The proposed multimodal image representation is built using a NMF algorithm to decompose the original data matrix in several latent factors that may be understood as image clusters. This operation is done using visual features as well as available text data. The image collection visualization is then built using a projection algorithm to reveal the relationships between images and text.

4.3.1. Non-negative matrix factorization

The general problem of matrix factorization is to decompose a matrix X into two matrix factors A and B:

$$X_{n \times l} = A_{n \times r} B_{r \times l} \tag{4.1}$$

This could be accomplished by different methods, including singular value decomposition (SVD). In that case $X = U\#V^T$, so $A = U$ and $B = \#V^T$.

This type of factorization may be used to do latent semantic analysis (LSA). In this case, X is a term-document matrix. The columns of A can be interpreted as the elements of a basis for the latent space, which is r-dimensional. The columns of B are the codifying vectors, i.e., they correspond to the representation of the documents in the latent space. The problem of applying SVD to find the latent space, is that the codifying vectors could have, in general, negative values. This means that some documents are represented, not only by the presence of latent factors, but also by their absence.

To solve this problem, an additional restriction may be imposed on the basis vectors and codifying vectors to force all their entries to be positive. This is called Non-negative Matrix Factorization (NMF). Notice that in the case of SVD, the restriction is that the basis vectors must be orthogonal. This is accomplished by the Eigen-decomposition. In NMF this restriction does not apply, and this in turn can have tremendous benefits in the semantic representation, since intuitively, different concepts or clusters do "not" have to be orthogonal or non-overlapping with each other.

There are different ways to find a NMF [60], the most obvious one is to minimize:

$$||X \quad AB||^2 \tag{4.2}$$

An alternative objective function is:

$$D(X|AB) = \sum_{ij} \left(X_{ij} \log \frac{X_{ij}}{(AB)_{ij}} \quad X_{ij} + (AB)_{ij} \right) \tag{4.3}$$

In both cases, the constraint is $A, B \quad 0$.

Both objective functions are non convex, so there is no straightforward algorithm that guarantees finding the global optimum. A gradient-descent technique could still be applied however. For instance, Lee and Seung [88, 60] proposed an iterative algorithm using recursive updating rules for A and B that achieve a good compromise between speed and ease of implementation. In this work we used an algorithm to optimize the objective function presented in Equation 4.3.

NMF has been used to address different problems in computer vision [88, 89], machine learning [90, 91] and text-mining [63], among others.

4.3.2. NMF-based multimodal image representation

The image database is composed of two data modalities, herein denoted by X_v and X_t. The former is a matrix whose rows are indexed by n visual features and whose columns are indexed by l images. The latter has m rows to represent text terms and l columns for images as well. The construction of a latent semantic space may be done by decomposing the matrix of images that can have only visual features or only text annotations. However, to generate a semantic space for image indexing, we are interested in exploiting multimodal relationships between images and text.

The proposed strategy consists in the construction of a multimodal matrix $X = [X_v^T \, X_t^T]^T$. Then, the matrix is decomposed using NMF as follows:

$$X_{(n+m)\times l} = W_{(n+m)\times r} H_{r\times l} \tag{4.4}$$

where W is the basis of the latent space in which each multimodal object is represented by a linear combination of the r columns of W. The corresponding coefficients of the combination are codified in the columns of H. As was shown in [63], each column of W corresponds to a cluster of the original objects, and each column of H corresponds to an object represented in the latent space.

The next required step for image collection exploration is to manage objects that have partial information, i.e. images without associated text content. An object with only visual information, represented by the vector $y \quad \mathbb{R}^n$ of visual features, is mapped to the latent space by finding $h > 0$ that satisfy the following equation:

$$y = W^v h, \tag{4.5}$$

where $W^v_{n\times r}$ is a trimmed version of W that includes only visual features and does not include the rows corresponding to text terms. This equation is solved by applying a modified version of the NMF algorithm that only updates h keeping W unchanged. This strategy allows us to map both multimodal and unimodal objects to the same latent space, enabling the system to visualize those images without text annotations.

One important aspect in the proposed multimodal scheme is that we can represent text terms in the same space as images are being represented. The latent semantic space is indexed by r latent factors that can be understood as the membership degree of each image to each of r clusters. In the same way, each of the m text terms have a representation in the latent semantic space given by the rows of the matrix $W^t_{m\times r}$. Thus, since the position of text terms is known, we can analyze their neighborhood to understand image semantics.

4.3.3. Multimodal visualization

Given a set of objects, the visualization process consists in organizing both images and concepts in a two dimensional space. This can be accomplished by dimensionality reduction techniques. The generic problem of dimensionality reduction is the following. Given a set of k points, $\{x_1, \ldots, x_k\} \quad \mathbb{R}^n$, find a set of points $\{y_1, \ldots, y_k\} \quad \mathbb{R}^l (l \quad n)$ such that y_i represents x_i. Given a set of objects, the visualization process consist in organizing both images and concepts in a two dimensional space.

We use Principal Component Analysis (PCA) algorithm to reduce the dimensionality of text data and images taking their representation in the latent space. As input, PCA receives a transformation matrix T obtained as follows,

$$T = \left[W^{\mathrm{T}}_{rxm} \, H_{rxl}\right],$$

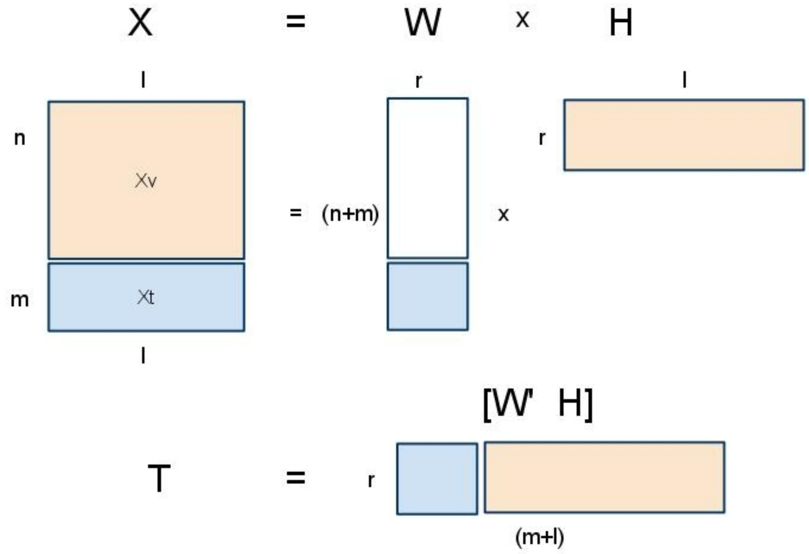


Figure 4.1.: Illustration of the latent factor analysis in which visual and textual content are combined. In this analysis m elements of W (blue) and l elements of H (brown) are concatenated to build T, which is the multimodal latent representation of each image in the data set.

where W^{T}_{rxm} is the representation of concepts in the latent space and H_{rxl} is the representation of images in the latent space. Since we aim to visualize concepts and images in the latent space, we use only the m vectors of W corresponding to concepts and all vectors of H corresponding to all images. Once we obtain the two principal components of T, we project each object in a 2-dimensional coordinates system. Figure 4.1 shows an illustration of the latent factor analysis to build the proposed multimodal visualization. In this analysis m elements of W (Blue) and l elements of H (brown) are concatenated to build T, which is the multimodal latent representation of each image in the data set.

4.3.4. Impact of text and visual content in visualization

A natural question that arises is what is the impact of text modality over visual modality in the resulting visualization. In order to perform this analysis, we define a trade-off between X_v and X_t in a convex combination. Formally, the following equation expresses this combination:

$$\begin{bmatrix} (1 \quad \alpha)X_v \\ \alpha X_t \end{bmatrix} = \begin{bmatrix} (1 \quad \alpha)W_v \\ \alpha W_t \end{bmatrix} H_v, \tag{4.6}$$

where α range from 0 to 1.

4.3.5. Measuring the quality of the visualization

In a 2D visualization each image is projected in a x and y coordinates system. Images that belong to a class should be arranged close each other in the visualization space. This supposition allows to think that images of two different classes should not overlap each other, that is to say, they do not share common properties (visual or textual) so they should be organized in different visualization zones. Based on this supposition, we propose to measure the overlapping among all classes to evaluate the quality of the complete visualization. To reach this goal, we propose to model the visualization layout of each class as a probability distribution function.

4.3.5.1. Visualization layout as a probability distribution function

Figure 4.2 illustrates the arrangement of a specific class in which each point represents the coordinates of an image in the visualization space. To model a class visualization as a probability distribution function we propose the following steps:

- Divide the visualization space into a grid of $10x10$ cells
- Count the amount of images in each cell
- Generate a histogram with the probability of occurrence of images in each cell of grid.

According to this representation, each class is defined as a probability distribution function, which allows to perform a comparison between classes to measure its overlapping.

4.3.5.2. Overlapping

The overlapping between two classes can be calculated as the intersection between each pair of histograms thus:

$$Int(h_i, h_j) = \sum_{k=1}^{n} min\left(h_i(k), h_j(k)\right),$$

where h_i and h_j are the histograms of the class i and class j respectively, and $h(k)$ is the k-th bin of histogram h. A high score of the intersection measure indicates that a pair of classes are highly overlapped, so both classes are distributed in the same zone of the visualization space.

4.3.5.3. Overlapping graph construction

To calculated the complete overlapping of all classes, we build a class overlapping matrix K in which each cell contains the overlapping (histogram intersection) between each pair of classes of the collection. This matrix can be also visualized as a graph in which classes and its overlapping are represented. We propose the following steps to determine the impact of visual and textual content in the quality of the visualization as follows:

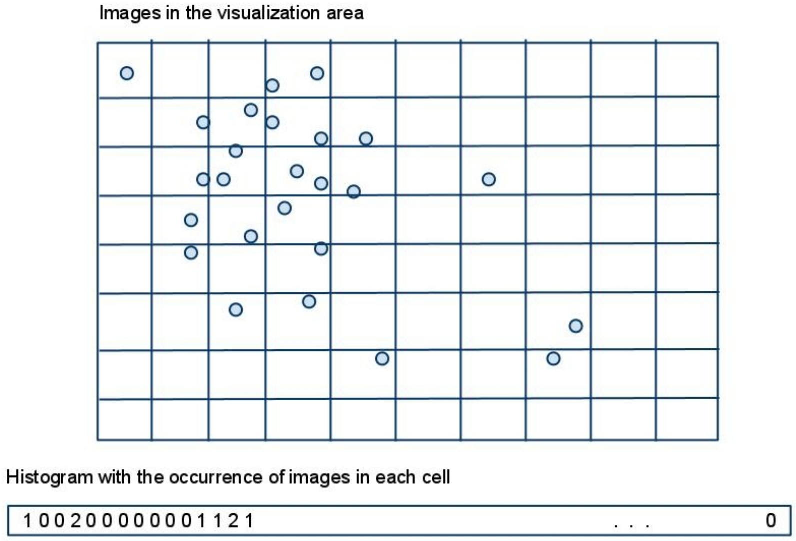

Images in the visualization area
Histogram with the occurrence of images in each cell
1 0 0 2 0 0 0 0 0 0 0 1 1 2 1
. . .
0

Figure 4.2.: Grid representation of the the visualization of a specific class. Each point represents the coordinates of an image in the visualization space. A histogram is built by counting the points in each cell of the grid. According to this model, each class is defined as a probability distribution function, which allows to perform comparison between classes.

- The graph $G(V, E)$ represents the relationship among all classes, where V represents classes and E represents the overlapping
- Edges are drawn when the intersection is higher than 0.5 (empirically set)
- We range α from 0 to 1 in Equation 4.6 looking for a graph with the maximum number of independent nodes

A graph with a high number of independent nodes indicates that the corresponding classes are visualized separately, that is to say, images of different classes are arranged away each other.

4.4. Experimental evaluation

Some experiments were performed in a well-known image data set. The goal was to make a qualitative evaluation of the visualization produced by the proposed method. These results helped us to assess the potential of the approach.

4.4.1. Experimental setup

Data set In this evaluation, we used a subset of the Corel image database which is composed of 2,500 images in 25 categories. The name of each category was used as concept and one image per concept was created. This dataset has become a *de facto* benchmark in automatic image annotation and retrieval research and has allowed the comparison of different strategies during the last years. It has been criticized of being easy and unrealistic [92], but other authors agree that it reflects important characteristics for image retrieval evaluation [78]. We believe that the most important attribute of this data set is its standard experimental protocol and clear ground-truth, that guarantee fair conditions to the academic community for reproducing experiments and present comparable evaluations.

Data representation Visual image content is represented using a bag of features approach, following a standard configuration. Each image is split in non-overlapping blocks of 8×8 pixels, and for each block the SIFT descriptor [93] is computed. Then, using an image training set, a codebook of SIFT descriptors is built using the k-means algorithm. We set the number of elements in the codebook to be 1,000. Then, a histogram is built for each image, counting the occurrence of the elements of the codebook in the image, looking for the most similar one for each block.

The matrix X_v^T, is then constructed using a vector in $\mathbb{R}^n$ for each image, in which $n = 1000$ is the number of visual features. To build X_t^T, a binary vector in $\mathbb{R}^{25}$ for each image is built using the category information, in which the i-th position is 1 if the image belongs to i-th class and 0 otherwise. This information is used to simulate keywords associated to image contents following a bag of words approach for text data. Notice that this representation can be easily extended to any vector space model for text data, in which unstructured annotations or multiple tags are available for each image. Categorical information has been used in this work to simplify the experiments and the analysis of the obtained results.

Finally, we computed the NMF factorization as $X_{(1000+25)\times 2500} = W_{(1000+25)\times 30}H_{30\times 2500}$. We set the value of the r parameter empirically to 30, which was determined by maximizing a standard measure in an image retrieval task. The concatenation of both feature vectors (visual and text) for each image was normalized to have norm $\ell_2 = 1$ to avoid scale problems in the 2D visualization.

Performance evaluation In order to objectively evaluate the proposed multimodal visualization method, we define the following measure based on the *stress* function defined in Equation 2.1:

$$Stress = \frac{\sum_{i,j}(d^T_{i,j} \quad D_{i,j})^2}{\sum_{i,j}(D^2{}_{i,j})}, \tag{4.7}$$

where $d^T_{i,j}$ is the distance between the i-th and j-th images using its textual representation, and $D_{i,j}$ is the distance between the i-th and j-th images in the projected space. The projected space is obtained by projecting the latent representation of images in a 2-dimensional space using PCA.

We evaluate the impact of the visualization when one modality or two modalities are used in the visualization process as follows:

- **Visual**: The visual information is used to be directly projected to the 2D space

- **Latent-visual**: A factorization of the visual matrix X_v is performed. The matrix H_v is used as input to the PCA algorithm obtain a (x,y) set of coordinates, one for each image.
- **Latent-multimodal**: A factorization of the a multimodal matrix X is performed. The matrix H is used as input to the PCA algorithm obtain a (x,y) set of coordinates, one for each image.

For these three scenarios the semantic stress of Equation 4.7 is calculated. A higher value of this measure indicates a poor visualization, and a high value indicates a better visualization. An intuitive explanation of this hypothesis is that images that belong to the same semantic class should be projected closer each other in the visualization space.

4.4.2. Results

4.4.2.1. Multimodal visualization

We built some multimodal visualizations where concepts and images are distributed in the 2D coordinates system generated by PCA. Figure 4.3 shows the complete image collection together with the text terms, according to the representation in the latent multimodal space. Even though some images are occluded and the layout has not been optimized, the user can get oriented in the metaphor thanks to the presence of the text terms in the visualization. It is especially remarkable that some similar text terms, from the semantic viewpoint, are represented closely in the latent space, and therefore in the visualization as well. For instance, notice the close position of the terms *plants* and *forest* as well as *beach*, *boats* and *isles*. Individually, all of them are identified as completely different categories, but they share many similarities in terms of visual properties, as well as from the semantic perspective. The coherence of their positions supports the idea that the NMF algorithm is providing a consistent representation for images since a semantic perspective, and also shows how the visual patterns are revealing meaningful connections between text terms. Other examples can be found by observing the positions between *volcano-mountain* and *flags-cards*.

Also, the quadrants of the visualization can be conceptualized with some other high-level interpretations of the image categories. For instance, the concepts at the southern part of the visualization may be associated to open landscapes, such as those for *boats*, *beach*, *aviation* and *mountains*, among others, while the northern part, may be associated to closed landscapes for *roses*, *fruits*, *cats* and *dogs*. The western may be associated to more artificial scenes such as those related to *cats*, *flags*, *cards* and *drinks*, while the eastern may be more associated to natural landscapes. In that way, we can notice that images are being organized in a high-level semantic way among the latent semantic space.

To further consider the correspondence between images and text terms, we select some categories to evaluate the distribution of the images around the corresponding text term.

For the selected terms, we project all the images belonging to that categories, that is, 100 images per class, and the selected terms are highlighted. Figure 4.4 shows the images corresponding to the categories *drinks* and *roses*. Observe that

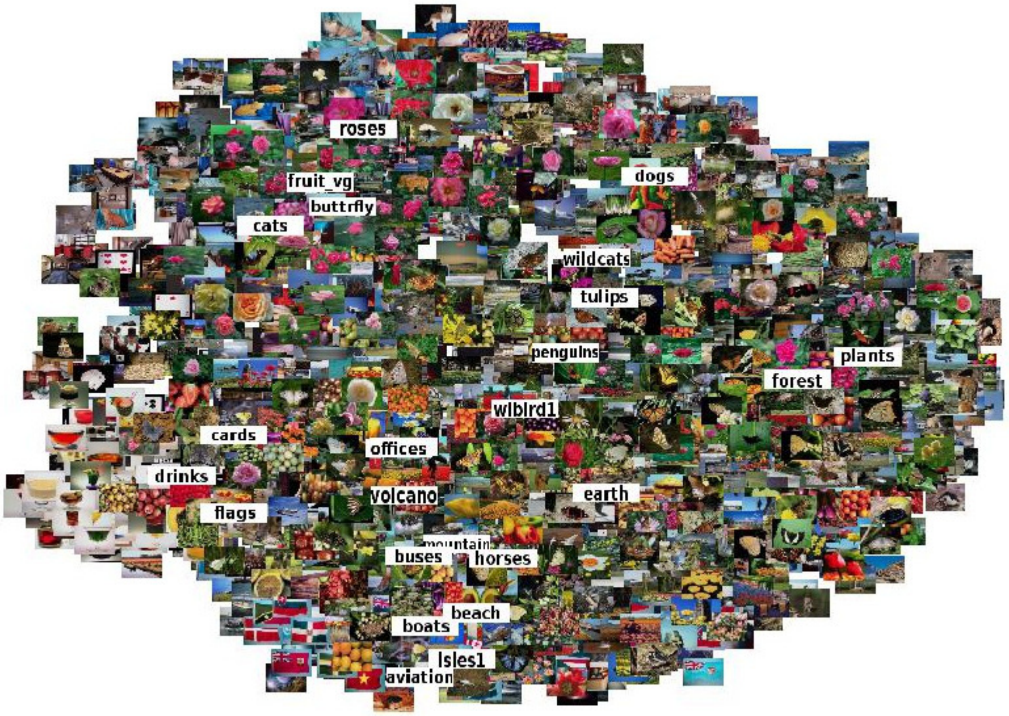


Figure 4.3.: Multimodal visualization of the complete image dataset and concepts

almost all images in the category *drinks* are compacted around the corresponding term, which means that a potential user exploring the collection will be able to find relevant images associated to this concept just in the region around the term. In the case of the category *roses*, images are mildly spread around the screen, but they are still surrounding the corresponding text term. Interestingly, the category *roses* shows a particular subset of pictures that are relatively far from the term, basically in an opposite side of the projection. However, this subset has been placed close to some semantically related text terms: *plants* and *forest*, in a region in which a potential user might find these roses relevant as well.

Figure 4.5 illustrates another case using the categories *flags* and *penguins*. In this case, the images are even more spread around the screen, suggesting that may be difficult to find images of these categories just by exploring the region surrounding the corresponding text term. This problem may be related to the particular choice of visual features, since it has been suggested that color information might improve classification and retrieval performance for the dataset used in this work. Nevertheless, although the images are more spread in this case, they preserve certain level of organization in terms of the region of the screen that is being occupied.

4.4.2.2. Visual and text impact

Figure 4.6 presents an example of a graph for $\alpha = 0.1$ in Equation 4.6. Note how some classes are not connected (penguins, drinks, etc) because they have a low semantic similarity, and other classes such earth and volcano are connected because

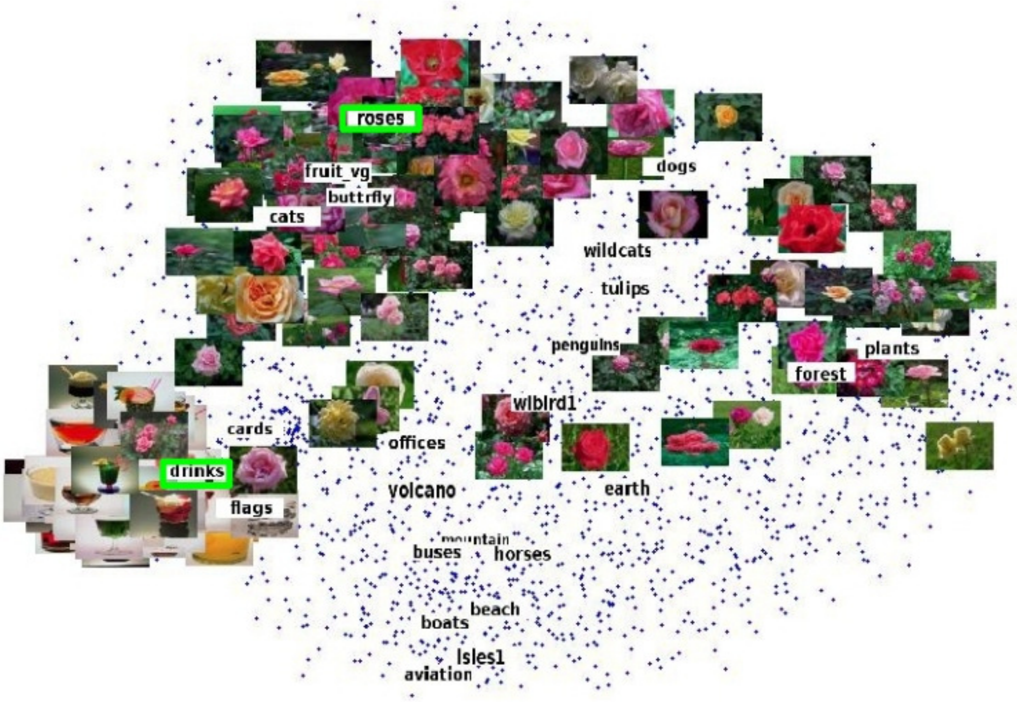


Figure 4.4.: Multimodal visualization highlighting drinks and roses

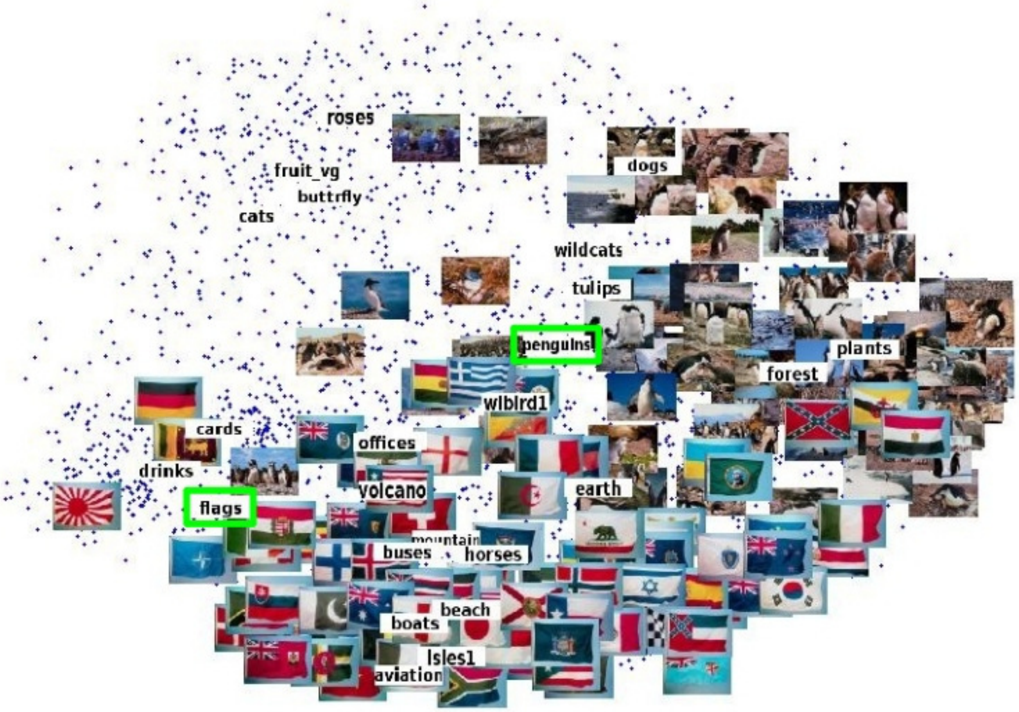


Figure 4.5.: Multimodal visualization highlighting flags and penguins

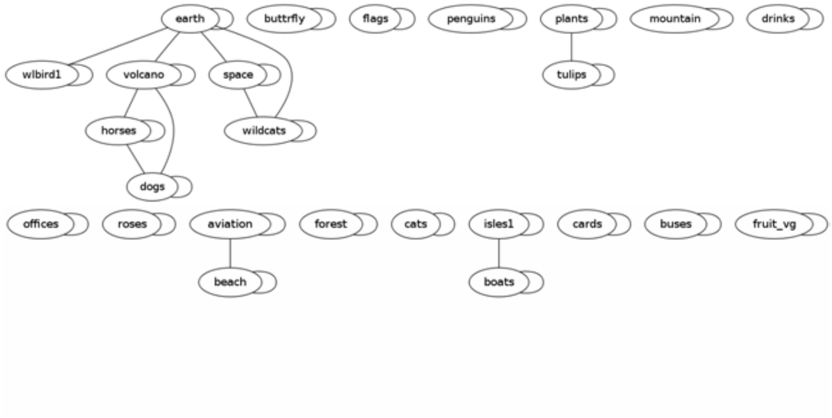


Figure 4.6.: Graph representation of a multimodal visualization. Edges indicate that classes are closer each other.

of its semantic similarity. Figure 4.7 shows the visualization of the corresponding obtained graph.

4.4.2.3. Semantic stress

Figure 4.8 shows the semantic stress obtained for three image representations when the number of latent factors is increased. The visual curve corresponds to the semantic stress obtained when the original visual representation is used in the projection process. Note that it remains constant. The visual latent representation obtains a better performance with respect to the original visual representation, which indicates that the latent representation improves semantic relationships between images. The best performance is obtained for the latent multimodal representation, which is obtained by fusing text and visual modalities in the same latent space.

4.5. Conclusions

This chapter presented a method that brings a semantic organization of the image collection. This is reached by performing a joint analysis of visual features and text terms to construct a meaningful visualization. We used a Non-negative Matrix Factorization algorithm to built a latent space for multimodal data, in which images and text terms can be represented together. We showed the potential of the proposed strategy, following a qualitative and quantitative evaluation of a multimodal image collection visualization. The first clear advantage of the proposed approach is the ability to locate text terms in the 2D canvas to guide the user in a hypothetical exploration process. This result makes an important difference among the state-of-the-art methods for image collection exploration, which are mainly based on visual features.

The proposed approach also showed a consistent distribution of text terms and images along the collection visualization. Interestingly, the text terms were clustered in high-level semantic regions as the result of the joint analysis with visual

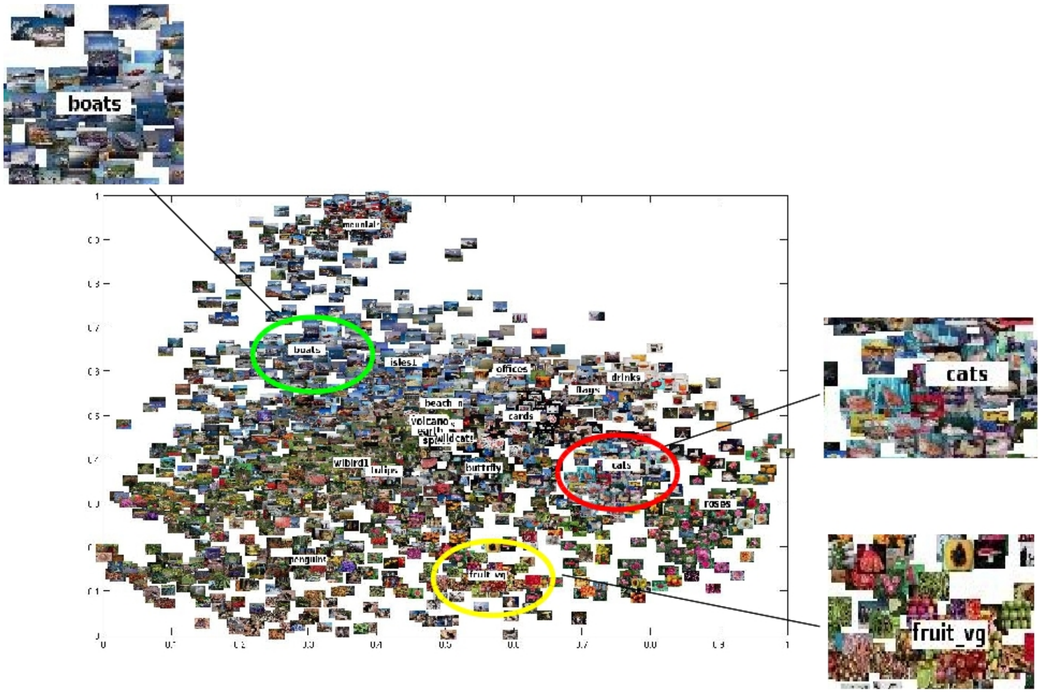


Figure 4.7.: Multimodal visualization. Highlighted zones indicate multimodal groups that were automatically organized close to a descriptive concept such as boats, fruits and cats.

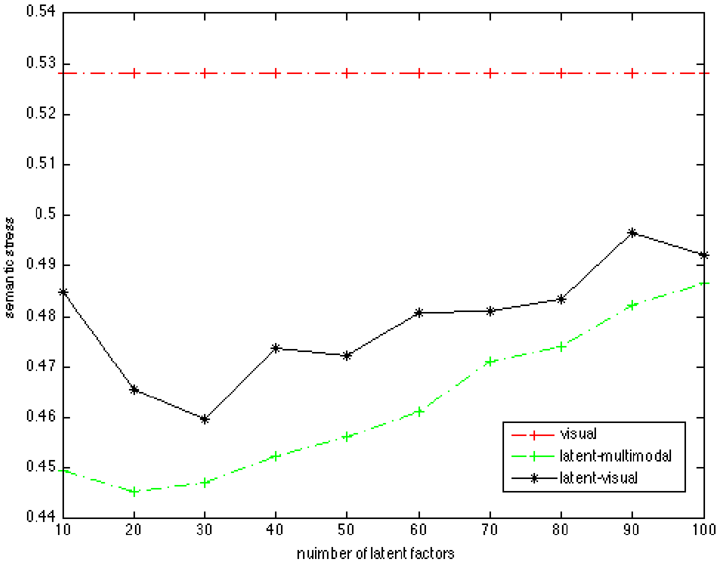


Figure 4.8.: Semantic stress obtained for different image representations: visual representation, visual latent representation and multimodal latent representation.

features. These high-level concepts, such as natural or artificial scenes, closed or open landscapes, etc., were observed among the organization of the text terms on the canvas, and are mainly related to visual characteristics in the image collection.

Part III.

Interaction

5. A Kernel-based Framework for Image Collection Exploration

This work has been published in the Journal of Visual Languages & Computing [94].

Image search systems still face challenges. The keyword-based query paradigm used to search in image collection systems, which has been successful in text retrieval, may not be useful in scenarios where the user does not have the precise way to express a visual query. Image collection exploration is a new paradigm where users interact with the image collection to discover useful and relevant pictures. This chapter proposes a framework for the construction of an image collection exploration system based on kernel methods, which offers a mathematically strong basis to address each stage of an image collection exploration system: image representation, summarization, visualization and interaction. In particular, our approach emphasizes a semantic representation of images using kernel functions, which can be seamlessly harnessed across all system components. Experiments were conducted with real users to verify the effectiveness and efficiency of the proposed strategy.

5.1. Introduction

Image collections are growing very fast due to the easy access to low-cost digital cameras and other image capturing devices. For instance, Flickr, a photo sharing system on the web, receives about 5,000 new photos per minute. Finding useful information in these large image collections is a very difficult task. Current approaches for image search are based on keywords or image examples provided by users [95], requiring a clear goal during the search process. These systems are useless when all what is required is to understand the structure of an image collection, i.e., to understand what kind of images are available in the collection and what are the relationships between them. For instance, in a personal photo collection that has been built for years, someone would like to explore the contents of the entire collection to identify meaningful patterns, special events, remember occasions and re-discover images that were archived a long time ago. This task becomes challenging when the collection comprises thousands of images that are shared between multiple users, such as in a scientific research community, training in medical imaging, and even in social networks and web communities. A system to explore and understand an image collection could rise the awareness of all the useful material archived in these repositories.

Building an exploration system involves several technical problems to process and organize visual information, following a strategy for *image collection exploration.*

We define an image collection exploration system as the composition of four main components: content representation, summarization, visualization and interaction. *Content Representation* is the ability of the system to encode and recognize visual configurations in images. This is usually done using features that discriminate different images and match similar ones. *Summarization* consists in selecting an iconic portion of the repository that represents a set of images to build an overview that allows the user to see example images of potential contents. These example images should be representative pictures of semantic subsets in the collection, assuming the existence of subsets that share meaning and/or visual relationships. *Visualization* consists in organizing a group of images into the screen following a metaphor that represents relationships between them. *Interaction* is the stage in which the user interacts with the system. At each step of the exploration process, the user performs some actions to feedback the system and the system learns from this feedback to refine search results.

Previous works that address the image collection exploration problem focus on parts of the whole problem. The retrieval community makes special efforts on image content representation to build systems that correctly match similar images [48]. However, these systems are oriented to support specific query paradigms as opposed to offer exploration capabilities. Other works that concentrate on summarizing and visualizing image collections [31, 96, 28, 97, 98], usually build their systems on top of very basic image representations and may not evaluate interaction functionalities. Research efforts in interactive search systems may be close to integrate all the pipeline [8], however, delegate the major responsibility to the interaction component, and do not explore dependencies between different subsystems. A holistic view of the problem may provide insights of which parts of the whole problem contribute to bring a better experience for users.

In this chapter, we propose the design of an image collection exploration system that brings an alternative way to access image repositories. The proposed system allows to explore image collections in an intuitive way where the system visualizes image results using a 2D metaphor exploiting image similarity, summarizes image results, and learns from the user interaction in order to refine the user search needs. Besides, we propose a framework, completely based on kernel methods, to model the computational methods behind each stage of the system. In this framework, it is possible to use any valid image kernel, which enables the system to use discriminative and expressive content representations. In order to harness the power of kernel functions to build meaningful representations, we also propose the construction of a semantic kernel that combines different visual features according to information of categories to better discriminate between different classes of images. This representation is exploited in each component of the system to generate relevant summarizations, useful visualizations and appropriate interaction responses.

The contributions of this work are: 1) the design and implementation of a complete image exploration system which includes all four components: a common content representation, summarization, visualization and interaction. Up to our knowledge, this is the first work that presents a fully integrated system for image exploration. 2) A semantic image representation based on kernel functions, that integrates category information of images together with visual features. This representation is the result of an optimization procedure that selects the best combination of visual features with respect to a category discrimination criterion. 3) To take advantage of the semantic image representation provided by kernel functions, the proposed computational framework is entirely based on kernel methods. Then, the framework exploits non-linear patterns in the collection and separates the representation from the algorithms.

This chapter is organized as follows: Section 2 reviews some related work; Section 3 presents material and methods; Section 4 presents experimental evaluation; and finally, Section 5 presents the concluding remarks.

5.2. Related work

5.2.1. Image representation

In computer vision applications, visual features are usually organized in a feature vector, in such a way that images are represented as points in $\mathbb{R}^n$ (where n is the number of features). Kernel methods are also popular in computer vision to analyze and recognize images and objects. Kernel representations are useful to find nonlinear patterns in data using the kernel trick, which is a way of mapping observations into an inner product space, without explicitly computing the mapping, and expecting that non-linear patterns in the original representation space map to linear patterns in the kernel-induced space [20]. Intuitively, kernel functions provide a similarity relationship between objects being processed. Kernel functions have been successfully used in a wide range of problems in pattern analysis since they provide a general framework to decouple data representation and learning algorithms.

Kernels on images have been widely proposed: histogram intersection kernel [99], graph kernels [100] and pyramidal kernel [101]. In this work, we use *histogram intersection kernel* (HIK) [102], however, the proposed modular strategy can be applied to any kind of valid image kernel. In our work, we have chosen the HIK since it has exhibited excellent performance when image features are represented by histograms. HIK has been successful in different domains such as pedestrian detection [103, 104], scene recognition [105], pattern recognition [101, 106], and image recognition [107, 108], among others. Recently, the HIK has attracted a lot of attention in the computer vision community not only because of its excellent performance as similarity measure but because of its fast evaluation speed [109].

Besides the use of a powerful kernel for image analysis, our work extends the representation of images to include semantic information extracted from image categories. Usually, kernel functions for images are used for classification tasks, i.e., supervised learning settings. However, the problem herein considered is not a classification problem, but an image exploration system. In that respect, we need to incorporate semantic information directly in the representation, which is achieved using optimization strategies on kernel functions.

5.2.2. Image collection summarization

In general, keyword-based search systems for images may be a restrictive query strategy for users, specially when specific visual compositions are required. In such cases, additional query tools are useful. Interactive search systems for images have been studied during the last years to allow users to introduce additional hints to the system about the images that should be retrieved. The use of example images as queries may provide a good head start for finding related images, so, this is one of these potential hints [95]. However, finding a good query image is a serious problem by itself.

To overcome this problem, the system may help the user by showing a set of images from the collection, to allow the selection of interesting images and conduct an exploration. The first problem of this approach is the selection of a proper set of images to show to the user. Simple strategies may be used such as just picking a fixed number of random pictures from the database. But that can be very frustrating for users if the set of images does not contain something representative, and asking for more random pictures might be hopeless. Summarization is a more promising strategy, which builds on top of the extraction of meaningful patterns from the image collection to select representative images in an automatic and informed way.

In most of the cases, the summarization problem is approached as a non-supervised learning problem. Typically, image clusters are identified in the collection and representative images from each cluster are chosen to compose the summary. Previous works that have addressed the construction of image collection summaries are based on different methodologies, from clustering methods [23, 24] to similarity pyramids [26], including graph methods [28, 29], neural networks [61], formal concept analysis [30] and kernel methods [32]. Most of these works use low-level visual features for the construction of summaries, and it is well known that the quality of the resulting clusters depend on the underlying content representation. The more semantic the representation, the higher the quality of the summary. In our work, we make a special effort to build an image representation that incorporates semantic knowledge for the construction of summaries.

5.2.3. Image collection visualization

After a summary has been built, the next challenge is to show the summary to the users. Again, simple strategies can be employed, such as sorting the set of representative images in a sequential list. However, to take advantage of the ability of human beings to interpret visual information, a more powerful visualization can be constructed, in such a way that relationships between images may be easily identified. The organization of images in a 2D map is a useful metaphor for achieving that. Thus, projecting the high dimensional representation of images into the screen actually helps to reveal groups of images that share similar contents.

Methods like Multidimensional Scaling MDS [9], Principal component analysis (PCA) [110], Isometric Mapping (Isomap) [12], Local Linear Embedding (LLE) [11], and combinations of them, have been used to generate the navigation map. Some works focus on nonlinear dimensionality reduction, also known as manifold learning. Weinberger et al. [111] for example, address the problem of learning a kernel matrix for nonlinear dimensionality reduction, trying to answer the question of how to choose a kernel that best preserves the original structure. In [21], kernel Isomap is proposed, which addresses generalization and topological stability problems of the original Isomap.

Nguyen et al. [8] use some of the mentioned methods and focus on how to organize the obtained projection taking into account a trade-off among image collection overview (summary), visibility (occlusion) and structure preservation (dimensionality reduction). Janjusevic et al. [40] address the problem of how to optimize the limited space to visualize an image collection. In [39], the authors propose a modification of MDS that solves the overlapping and occluding problems, using a regular grid structure to relocate images. Chen et al. [112] propose a pathfinder-network-scaling technique that uses a similarity measure based on color, layout and texture.

Liu et al. [113] use a browsing strategy based on a one-page overview and a task-driven attention model in order to optimize the visualization space. Users can interact with the overview with a slider bar that allows to adjust the image overlapping. Porta [4] proposes different non-conventional methods for visualizing and exploring large collection of images, using metaphors such as cubes, snow, snakes, volcanos and funnels, among others.

Most of these algorithms and strategies try to preserve image relationships from a low-level feature space into a 2D map. However, once again, the original representation in an exploration system requires to represent semantic relationships among images to help the user visualize meaningful patterns. In our work, we develop a common framework to exploit the same image representation used to build summaries to also organize meaningful visualizations.

5.2.4. User interaction and feedback

When designing image search systems, the problem of the semantic gap has to be addressed, i.e., managing the difference between low-level image representations and the semantic high-level human perception. One successful strategy is modeling human interaction, i.e., processing the user's feedback. In an interactive framework for image search, the user is an active element of the exploration process. The most popular model to involve the user in the process is the well known Relevance Feedback (RF) mechanism, which is widely used in text retrieval.

RF was introduced in the mid-1960's by the text retrieval community to improve retrieval effectiveness. RF is a process to automatically adjusting an existing query by using user's feedback. In the context of images, this mechanism is known as Interactive Content-Based Image Retrieval (ICBIR), where the image query is refined in an iterative process to model the user's needs. Basically, this model consists of a process where images are given to the users, they select which are relevant to the query and the system returns a new result based on this feedback. During this iterative process, the ICBIR system learns the correspondence between low-level features and high-level concepts.

The most representative works in this area can be organized as follows: (1) *Query point movement* [114, 115, 116], where the query is moved in the search space to relevant images according to the user feedback; (2) *Active learning* [117, 118, 119], where the system trains a Support Vector Machine (SVM) to select the closest images to the SVM boundary in order to show to the user a new set of images; (3) *The ostensive model* [120, 48], where users do not have to rank or label images, instead, the user only selects an image from the retrieved set and a new set is returned in response. In this model, the query is built from the user's interaction with the collection through time, and is calculated as a weighted sum of the features of the current image and the previously selected ones during the navigation. (4) *Long-term learning* [121], where all the interaction with the system in different sessions is recorded and used to train a learning model.

In our work, we formulate a RF mechanism that is incorporated in the proposed general framework. Our strategy takes advantage of the interaction process on top of a common kernel representation, which has been semantically adapted from the very beginning, providing an additional layer of abstraction to tackle the semantic gap.

5.2.5. Integrated exploration systems

We consider that an integrated system for image collection exploration should involve four important components: 1) An appropriate image content representation; 2) a summarization component to prevent the user from reviewing many pages; 3) a visualization strategy to organize exploration results in a meaningful way; and 4) an interactive mechanism that allows users to lead the system towards the desired results. Most of the works reviewed in this Section focus on one or two of these components, making contributions to the understanding of important, but independent problems.

In this work, we propose a system that integrates all of these components in a seamless framework. This allows to study the exploration process as a whole. Our approach exploits the ability of kernel methods to build expressive content representations for images and, at the same time, allowing methods to take advantage of nonlinear patterns in the representation. This basically allows the system and the user to manipulate a common content representation of images.

5.2.6. Data sets and evaluation

In the literature we find different works that address particular stages of the image collection exploration process, which use the Corel dataset in summarization [122, 123], visualization [8, 124, 40], and interaction [125, 126]. Other image data sets such as Flickr-based datasets have been recently used, but also for evaluating standalone exploration tasks: visualization [32, 67], summarization[127, 43], and interaction [128, 129]. Other important characteristic of these works is that they do not use a standard evaluation framework, as is the case of the TREC evaluation benchmark in content-based image retrieval tasks, where a complete process is defined to evaluate the performance of a new algorithm. This phenomena can be attributed to the fact that in image collection exploration there are many aspects that make difficult and challenging the evaluation process. In a recent survey of interactive image

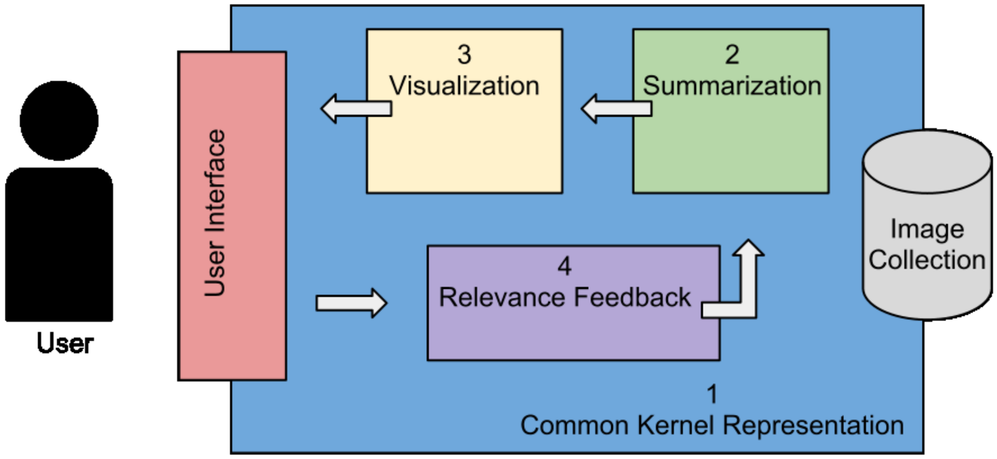


Figure 5.1.: Image collection exploration framework.

search reported in [130] authors describe in detail the difficulty and challenging of evaluating image collection exploration systems. Authors argue that evaluation of image collection exploration systems is in their ”infancy” and ”some works provide no evaluation, because they present a novel idea and only show a proof of concept.

5.3. Material and methods

The proposed framework is illustrated in Figure 5.1, which shows the integrated system working on top of a common kernel representation of images. The first component is the one that supports the whole framework providing a powerful representation of images using kernel functions. The following components, i.e., summarization, visualization and relevance feedback are formulated as kernel algorithms to exploit nonlinear patterns in the representation.

In this Section, we start describing the process of building a discriminative image representation that includes visual features and semantic information from image categories. Then, the algorithms for summarization and visualization are described. Finally, the relevance feedback mechanism is formulated, to complete the kernel-based framework for image collection exploration.

5.3.1. Image representation

The aim of the feature extraction process is to identify and extract relevant information from the image that allows discrimination of different image classes. Feature extraction approaches are based on the calculation of objective content measures related to visual patterns such as colors, textures, edges, among others. In general, images can be represented as a set of histograms for different visual features. In this work we extract the gray histogram, RGB histogram (color) [131], Sobel histogram (edges) [132], and Local Binary Patterns (LBP) histogram (texture) [133].

Although histograms can be seen as feature vectors, they have particular properties that can be exploited by a similarity function. Consider h as a histogram with n bins, associated to one of different visual features. The histogram intersection kernel between two histograms is defined as $k_\cap(h_i, h_j) = \sum_{l=1}^{n} min\,(h_i(l), h_j(l))$, where h_i, h_j are histograms and $min(\cdot,\cdot)$ is the minimum value between two histogram bins. Intuitively, this kernel function is capturing the notion of common area between both histograms. Note that, if the input histograms have n bins, the histogram intersection kernel is actually embedding these objects into a feature space of a dimensionality many times greater than n.

In this work, four different histograms were calculated for each image. Using $k_\cap$ and these four visual features, we obtain four different kernel functions that will be used for building a new kernel. A kernel function using just one low-level feature provides a similarity notion based on particular aspect of the visual perception. For instance, the RGB histogram feature is able to indicate whether two images have similar color distributions. However, we aim to design a kernel function that provides a better notion of image similarity according to prior information. We construct the new kernel function using a linear combination of kernels associated to individual features as follows,

$$K_\alpha = \sum_{i=1}^{N} \alpha_i K_i \tag{5.1}$$

The simplest combination is obtained by assigning equal weights to all basis kernel functions, so the new kernel induces a representation space with all visual features. However, depending on the particular class, some features may have more or less importance. For instance, in a class where images share the same color distribution the RGB feature will not be a good discriminant, so in this case other features such as texture and edges will be more suitable to discriminate it. Kandola et al. [134] proposed the kernel alignment strategy in the context of supervised learning to combine different visual features in an optimal way with respect to a domain knowledge target (*ideal kernel*). The empirical kernel alignment, is a similarity measure between two kernel functions, calculated over a data sample. If K_α and K_t are the kernel matrices associated to kernel functions k_α and k_t in a data sample S, the kernel alignment measure is defined as:

$$A_S(K_\alpha, K_t) = \frac{K_\alpha, K_t\ _F}{\sqrt{\ K_\alpha, K_\alpha\ _F}\sqrt{\ K_t, K_t\ _F}}, \tag{5.2}$$

where $\ \cdot,\cdot\ _F$ is the Frobenius inner product defined as $\ A, B\ _F = \sum_i \sum_j A_{ij} B_{ij}$, K_α is the linear combination of basis kernels, that is, the combination of all visual features given by $k_\alpha(x, y) = \sum_f \alpha_f k_\cap\,(h_f(x), h_f(y))$, where $h_f(x)$ is the f-th feature histogram of image x, and α is a weighting vector. The definition of a target kernel function K_t, i.e. an ideal kernel with explicit domain knowledge, is done using labels associated to each image that are extracted from previous information (class labels). It is given by the explicit classification of images for a particular class using y_n as the labels vector associated to the n-th class, in which $y_n(x) = 1$ if the image x is an example of the n-th class and $y_n(x) = \ \ 1$ otherwise. So, $K_t = yy'$ is the

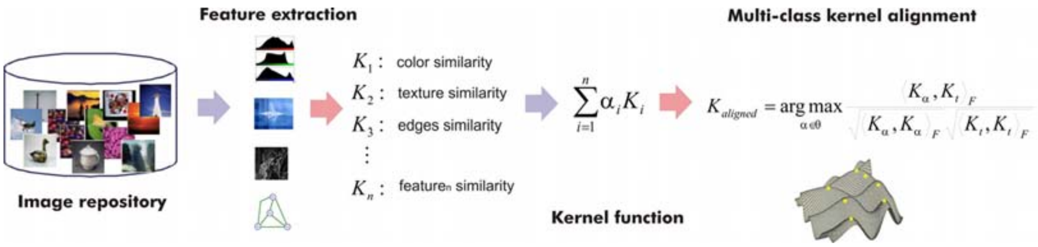

Feature extraction
Multi-class kernel alignment
Image repository
K_1 : color similarity
K_2 : texture similarity
K_3 : edges similarity
K_n : featuren similarity
$\sum_{i=1}^{n} \alpha_i K_i$
Kernel function
$K_{aligned} = \arg\max_{\alpha \in \Theta} \frac{\langle K_\alpha, K_t \rangle_F}{\sqrt{\langle K_\alpha, K_\alpha \rangle_F} \sqrt{\langle K_t, K_t \rangle_F}}$

Figure 5.2.: Image feature extraction process.

kernel matrix associated to the target for a particular class. This configuration only considers a two-class case. We need to build a new kernel function that takes into account the information of all classes simultaneously (*multi-class* case).

Vert [135] proposes a strategy that addresses the multi-class problem in the context of multi-class classification. Therefore, we adapted that strategy in the context of image collection summarization. The author proposes to build the ideal kernel matrix as follows:

$$K_t(x, x') = \begin{cases} 1 & \textit{if } \text{y=y'} \\ \quad 1/(Q \quad 1) & \text{otherwise} \end{cases}, \tag{5.3}$$

where Q is the number of classes. K_t is, by construction, a valid kernel and we will call it the *ideal kernel*. Under some regularity assumptions on K_α, the alignment function is differentiable with respect to α. Upon this assumption we can use a *gradient ascent* algorithm in order to maximize the alignment between the combined kernel and the ideal kernel as follows:

$$\alpha^* = \underset{\alpha \in \Theta}{argmax}\, A_S(K_\alpha, K_t) \tag{5.4}$$

Kernel alignment strategy has been used in the context of supervised learning and in classification problems. We use it for both, supervised and non-supervised learning in the context of image collection exploration. Figure 5.2 shows the process of proposed kernel alignment process.

5.3.2. Summarization

In large collections of images it is not possible to show all images at once to the user due to the limitations of screen devices. Therefore, it is necessary to provide a mechanism that *summarizes* the image collection. This summary represents an overview of the data set and allows the user to start the navigation process. A good summary corresponds to a representative set of samples from the collection, i.e., a set that includes prototypical images from the different categories present in the collection. The most common strategies to build image collection summaries are based on clustering algorithms [25]. Figure 5.3 shows a picture of the process used for summarizing an image collection. In the proposed framework, *kernel k-means* [20] was used to summarize the collection. As input, the clustering algorithm receives the aligned kernel matrix previously discussed. The parameter k of the algorithm

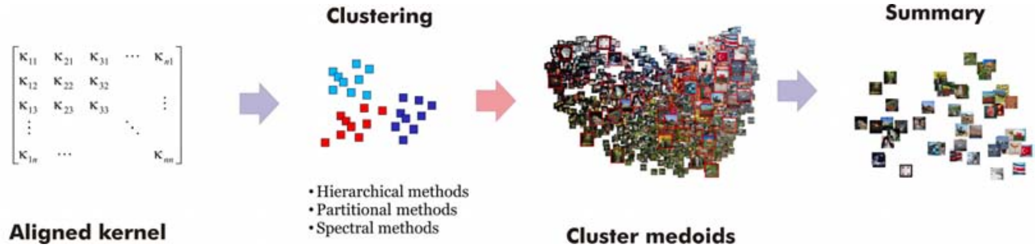


Figure 5.3.: Summarization process based on the kernel matrix.

can be set according to the number of semantic categories of the image collection. In this chapter we propose to analyze the entropy of the summary to determine a value that yields a good summary. Our hypothesis is that the larger the number of image examples, the more descriptive the image summary is.

5.3.3. Visualization

Given a set of images, the visualization consists in organizing those images in a 2D coordinates system. Hence, the goal is to find a 2D coordinate to each image given the kernel function. In this work, we use *kernel principal component analysis* (KPCA) [136] to find a representation of the image collection in a low-dimensional space. Images are represented based on their projections on the two principal components. This produces a 2-dimensional representation that is expected to preserve, to some extent, the similarity relationship, i.e., two similar images are expected to be projected to the same region. Figure 5.4 shows a 2D visualization of 2,500 images.

The aligned kernel matrix K_α obtained in Section5.3.1 is given as input to the algorithm, which generates a set of 2D coordinates, one for each image. Figure 5.5 shows the general process to build a 2D visualization from the kernel matrix.

5.3.4. Interaction

Exploratory image collection search involves several tasks performed by the user and other by the system. The complete interaction loop is shown in Figure 5.6. First, the system selects a representative image subset of the collection. Then, the user determines which images are relevant. Next, the system captures this feedback and reformulates the query using a relevance feedback (RF) model. Then, the system ranks the image set to select the r most relevant in order to reduce the search space. Thus, the system computes a new summary taking as input those images. Finally, the user determines whether the task has been completed, otherwise, the process is repeated until the user interests have been meet.

In our strategy, the search space is reduced at each feedback iteration. In the first stage, the system shows an image collection summary that represents the complete image collection. When the user selects a set of relevant and non-relevant images the first time, the system estimates the position of the new adjusted query and discards the $(l \quad r)$ farthest images from that position, with l being the size of the image collection and r the number of relevant images considered in this first

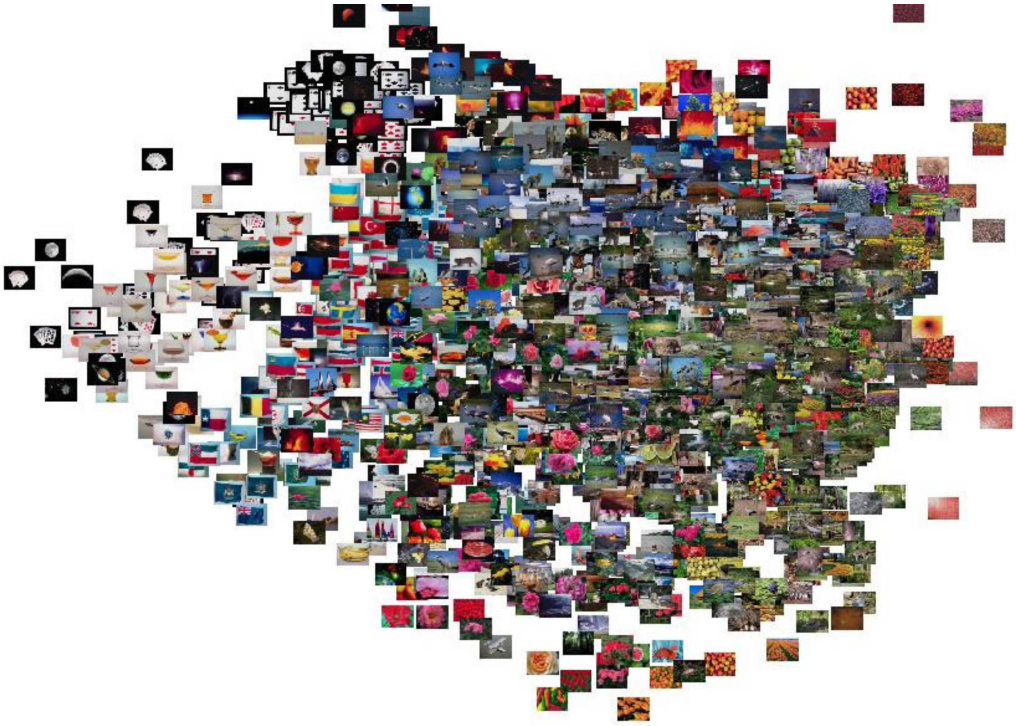


Figure 5.4.: 2-dimensional visualization. Similar images are projected close to each other in the visualization space.

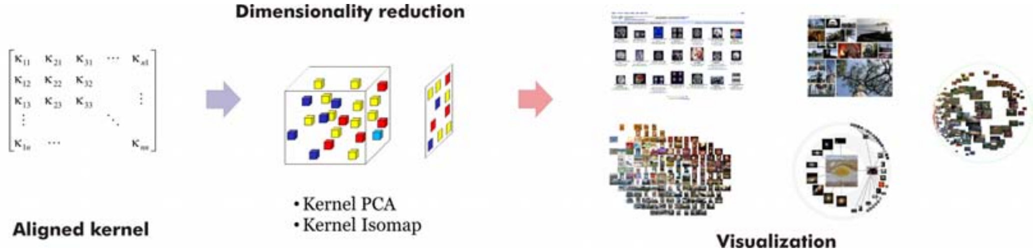

Dimensionality reduction
κ_{11} κ_{21} κ_{31} $\cdots$ κ_{n1}
κ_{12} κ_{22} κ_{32}
κ_{13} κ_{23} κ_{33}
κ_{1n} $\cdots$ κ_{nn}
Aligned kernel
• Kernel PCA
• Kernel Isomap
Visualization

Figure 5.5.: Visualization process based on the kernel matrix.

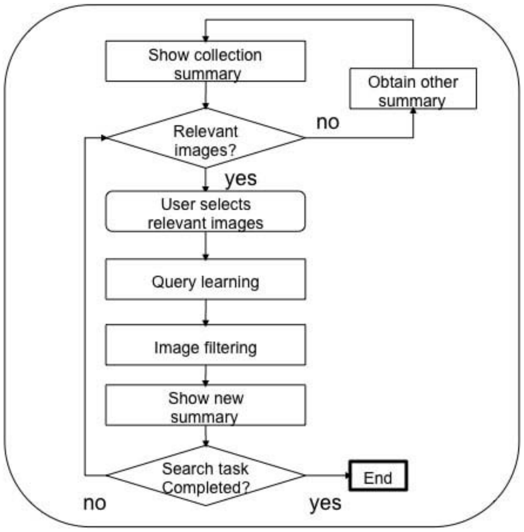


Figure 5.6.: Interaction process. In each iteration the query is refined with the user's feedback.

iteration. Then, using only the r most relevant images, a new summary is built to help the user iteratively concentrate on the kind of images that she/he is interested in. Note that the r parameter must be dynamic as the user interacts with the system, each iteration being smaller. This is equivalent to reduce the search space by taking a hypersphere centered in the query position, with a decreasing radius each time. At the beginning, the radius covers the complete image collection, since the user is exploring its contents. At the end, it is expected that the user may have access to a reduced set of highly relevant images, since the provided feedback relocated the center of the hypersphere to an interesting region. We explored two different functions to reduce the search space at each iteration: (1) A constant percentage of the current image set: $r_{t+1} = \sigma r_t$, and (2) An exponential decreasing size: $r_{t+1} = exp(\ \ \sigma r_t)$. The first function models a more exploratory task with slow convergence, compared to the second, which leads the user to a small image set faster. In both cases the user may control the σ parameter which defines the actual convergence speed.

5.3.4.1. Kernel-based relevance feedback model

Primal model Rocchio [137] proposed a relevance feedback model to learn from the user's feedback. In this model, results are exposed to the user who selects which documents are considered relevant and which are not. In this way, this task is performed iteratively such that in each iteration the query is closer to the positive feedback and away from the negative feedback. The model assumes a vector-space representation, i.e., both documents and queries belong to a real-valued vector space. The Rocchio's formula is expressed a follows,

$$q_{t+1} = \alpha q_t + \frac{\beta}{R}\sum_{j=1}^{R} d_j \quad \frac{\gamma}{NR}\sum_{j=1}^{NR} d_j, \tag{5.5}$$

where t denotes the t-th iteration, q_t is the query vector in the previous iteration, d_j is the j-th document vector from the relevant and non-relevant sets, R is the number of relevant documents, NR is the number of non-relevant documents and the parameters α, β, γ are used to weight the importance of each component in the formula. Once the new query has been reformulated using this model, the search algorithm must be triggered to identify the most relevant documents. This model has been widely used by the information retrieval community due to its simplicity, speed and easy implementation. It is important to note, that this model depends on the vector representation. In the proposed framework, each image may be represented as a complex object with structured data that can be compared to other objects using a similarity measure or kernel function. So that, in the proposed model, the original Rocchio's formula cannot be applied.

Dual model Formally, a kernel function is a function κ that for all x, z in a certain set X (the problem space) satisfies,

$$\kappa(x, z) = \ \phi(x), \phi(z) \ , \tag{5.6}$$

where ϕ is a mapping from X to an inner product space F (the feature space), $\phi : x \qquad \phi(x) \quad F$.

The function κ is a kernel function with F its corresponding feature space. This means that we can compute the inner product between the projection of two points into the feature space without explicitly evaluating their coordinates [20]. In the context of information retrieval, the problem space, X, corresponds to the original data and the feature space, F, corresponds to its representation. This means that the similarity between documents and queries must be calculated in the feature space, and this is exactly what the kernel function does. Following this idea, we formulate the Equation 5.5 in terms of points into the feature space and define its computation in terms of kernel functions.

First of all, we need to compute the center of mass of the relevant and the non-relevant images. The center of mass in the feature space can be computed as the vector $\phi_S = \frac{1}{\ell}\sum_{j=1}^{\ell}\phi(x_j)$. So we can compute the dot product between relevant vectors and their center of mass as follows,

$$\left\langle \phi(x), \phi_s^{rel} \right\rangle = \left\langle \phi(x), \frac{1}{R}\sum_{j=1}^{R}\phi(y_j) \right\rangle = \frac{1}{R}\sum_{j=1}^{R}\kappa(x, y_j), \tag{5.7}$$

where $\{y_1, \ldots, y_R\}$ is the set of relevant images and ϕ_s^{rel} is its center of mass. Using the same analysis for the non-relevant images,

$$\left\langle \phi(x), \phi_s^{nonrel} \right\rangle = \left\langle \phi(x), \frac{1}{NR}\sum_{j=1}^{NR}\phi(z_j) \right\rangle = \frac{1}{NR}\sum_{j=1}^{NR}\kappa(x, z_j), \tag{5.8}$$

where $\{z_1, \ldots, z_{NR}\}$ is the set of non-relevant images and ϕ_s^{nonrel} is its center of mass. Therefore, the similarity between an image x and the center of mass of relevant and non-relevant items can be expressed in terms of the kernel function calculated between relevant/non-relevant images and all the images involved in the current iteration. The Rocchio's model transforms the user's query according to the user's feedback, resulting in a new query vector q_{i+1}. Since we do not have an explicit vector representation for the query, let's introduce the notation $\phi(q)$ as the representation of the user's query in the feature space. Now, the Rocchio's model can be expressed in terms of the dot products in the feature space:

$$\phi(x_i), \phi(q_{t+1}) \quad = \quad \alpha \;\; \phi(x_i), \phi(q_t) \; + \left\langle \phi(x_i), \left(\frac{\beta}{R} \sum_{j=1}^{R} \phi(x_j) \right) \right\rangle \quad \left\langle \phi(x_i), \left(\frac{\gamma}{NR} \sum_{j=1}^{NR} \phi(x_j) \right) \right\rangle,$$

where x_i is a given image, and t is the t-th iteration. According to the Equation 5.6, we can rewrite last Equation in terms of the kernel function between the query and each i-th document in the collection,

$$\kappa(x_i, q_{t+1}) = \alpha\kappa(x_i, q_t) + \beta\kappa(x_i, r_s) \quad \gamma\kappa(x_i, nr_s), \tag{5.9}$$

where $\kappa(x_i, r_s) = \frac{1}{R}\sum\limits_{j=1}^{R}\kappa(x_i, x_j)$ and $\kappa(x_i, nr_s) = \frac{1}{NR}\sum\limits_{j=1}^{NR}\kappa(x_i, x_j)$, according to Equations 5.7 and 5.8 respectively.

Note that, even though we do not have an explicit representation of the reformulated user's query, we can calculate the similarity measure between any image in the collection and the new adjusted query. The similarity measure will be used as a score function to select the most relevant images.

5.4. Experimental evaluation

5.4.1. Dataset

The experiments were conducted on a subset of 25 classes from the Corel image database, each class has 100 items, leading to a collection of 2,500 images. This data set has been extensively used to evaluate performance in conventional image retrieval tasks, but not in a complete image collection exploration system, therefore we propose a general experimental evaluation setup suitable for image collection exploration systems that uses this dataset. Our results are not directly comparable with previous works reported in the literature since previous works use experiment setups that are heterogeneous and measure the performance of specific standalone image collection exploration tasks. In this chapter we propose an evaluation setup that can be used in future works to perform a complete evaluation of an image collection exploration system as an overall process.

5.4.2. Aligned kernel matrix

For the experimental evaluation, the following image features are used: Gray histogram, RGB color histogram, Sobel histogram (borders) and local binary patterns (texture). As described in Section 5.3.1, an optimal linear combination of feature kernels is found by optimizing the function described in Equation 5.4. The optimization was performed with a gradient ascent algorithm that started with random α values, step size $\eta = 0.1$, $\quad \alpha_i = 0.1$ and 100 iterations.

The optimal kernel matrix can be computed with the α obtained after the optimization process,

$$K_\alpha = 0.1537 \cdot K_{gray} + 0.0507 \cdot K_{lbp} + 0.1023 \cdot K_{sobel} + 0.1537 \cdot K_{rgb}. \tag{5.10}$$

Note that the color kernel has the highest weight in the linear combination. It means that colors have different distributions among different classes in the Corel datasets, and so allow to discriminate contents. In contrast, the texture kernel (LBP) has the lowest weight, which indicates that texture is not a good class discriminant in this data set.

5.4.3. Summarization and visualization

A good summary is a representative set of samples from the collection, i.e., a set that includes prototypical images from the different categories present in the collection. Based on this idea, we define a supervised summarization quality measure that makes use of the image labels. This measure corresponds to the entropy of the summarization and is calculated as follows:

$$H_{summary} = \quad \sum_{i=1}^{C} (\frac{\#C_i}{k}) log_2(\frac{\#C_i}{k}), \tag{5.11}$$

where C is the number of classes, $M = \{m_1, \ldots, m_k\}$ is the set of k medoids obtained in the clustering process, and $\#C_i = \quad \{m_j \quad M | m_j \quad C_i\}$ is the number of medoids in M that belong to class C_i. The quantity $\frac{\#C_i}{k}$ represents the proportion of samples in the summary that belongs to class C_i. The maximum entropy is obtained when this value is the same for all classes, e.i., $\quad i, \frac{\#C_i}{k} = \frac{1}{C}$. In this case, all the classes are equally represented in the summary. The maximum entropy depends on the number of classes, $H_{summary} = log_2(C)$. In this experimental setup $log_2(C) = log_2(25) \quad 4.64$. With this measure defined, we aim to assess the quality of the summaries generated for the following kernel functions: an *ideal kernel function* using the Equation 5.3, which will have the maximum entropy since it has the a priori class labels information; a *base-line kernel function* as a combination of the base kernel functions (RGB, Sobel, LBP and Gray) with equal weights; and the *aligned kernel* built as was suggested in Section 3.

Figure 5.7 shows the quality of the three summaries: *ideal kernel, basis kernel,* and *aligned kernel*. Results show that the *aligned kernel* outperforms the baseline, which indicates that the proposed method improves the quality of the summary. All three kernels increase the summary entropy when the number of medoids is increased; with *k=50* medoids the summary entropy is close to the maximum. A minimum of this parameter is related to the number of representative images in the summary. In our experiments we know a priori 25 classes, but results show that we reach a meaningful discrimination when k=50, that is, with two times the number of classes. Entropy is a measure that can be useful to automatically found k in a collection where a priori class information is unknown. The obtained entropy for different values of k shows that the minimum value of this parameter can be found through entropy analysis of the summarization process.

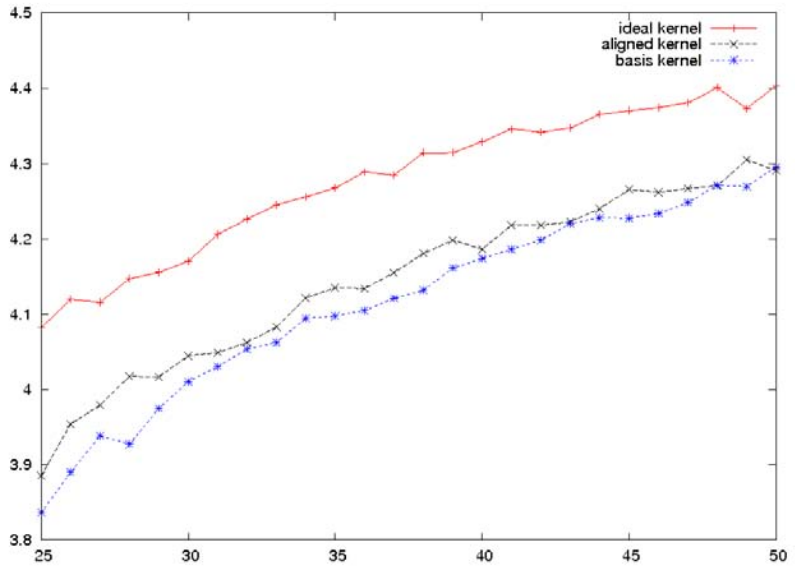


Figure 5.7.: Entropy vs number of centroids (average for 100 runs). The kernel that involves domain knowledge (*aligned kernel*) outperforms the base-line kernel.

Figure 5.8 shows the visualization of the entire data set using KPCA and a visualization of the automatically generated summary is shown in Figure 5.9.

5.4.4. Interaction

We implemented a prototype system using a web-enabled user interface. The system was deployed on a computer with a 3.0 GHz Pentium D processor and 4 GB of RAM memory. A screenshot is shown in Figure 5.10.

The summarization algorithm was empirically set to find 100 representative images from the results in each iteration. The system allows users to select relevant images in each iteration and provides a search button to request an update of the image collection visualization. We aim to identify the way in which the relevance feedback parameters influence the response of the system in a *category search* task. The purpose of this task is to find as many images as possible of a particular category.

The system evaluation involved 10 master students aged 25-30 years old whom had experience with search engines. We met each participant separately and followed the procedure outlined as follows: (1) each user was given a training session on the system; (2) each user was asked to search images related to one of the 25 possible categories according to a one random image generated as starting point; (3) each

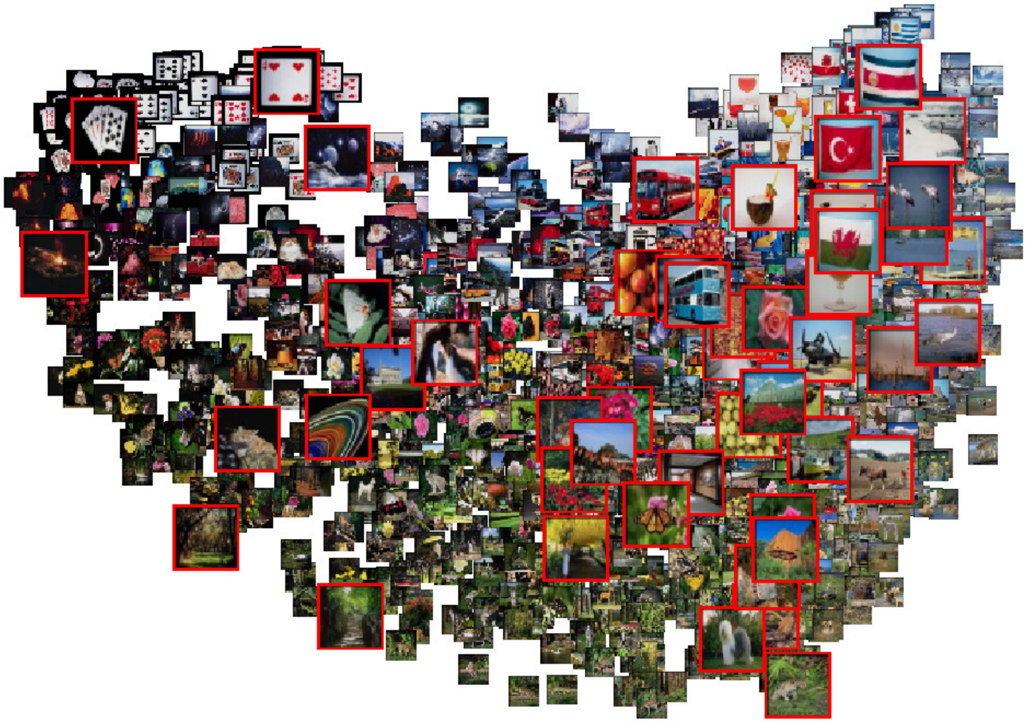


Figure 5.8.: 2D visualization of the complete dataset highlighting the most representative images

user had to complete at least 10 iterations in the relevance feedback loop; (4) in each iteration, the number of images of the target category was recorded in the server log. Experiments were grouped in three particular evaluations: weighting impact, filtering behavior and global performance.

5.4.4.1. Weighting impact

This evaluation aims to analyze the system response using different values of the parameters α, β and γ. These parameters weight the importance of the selections made in previous iterations, the relevant examples selected by the user and the non-relevant images, respectively. Figures 5.11 and 5.12, show the visualization of one of the experiments at iterations 3 and 4 respectively. The larger the number of iterations, the higher the amount of images of the target category. For this experiment, we calculated the recall of the first 10 iterations as is shown in Figure 5.13. Using the parameters $\beta = 0.4$ and $\gamma = 0.6$, indicates that non-relevant images have more importance in the definition of the user's query. The plot associated to these parameters showed the lowest recall since the query is forced to the direction of the negative examples and the user obtains more non-relevant images. The zig-zag shape indicates that the interaction with the user has a contrary effect, since relevant images selected by the user have less importance than the other amount of non-relevant ones. With $\beta = 0.8$ and $\gamma = 0.2$, we give more importance to the

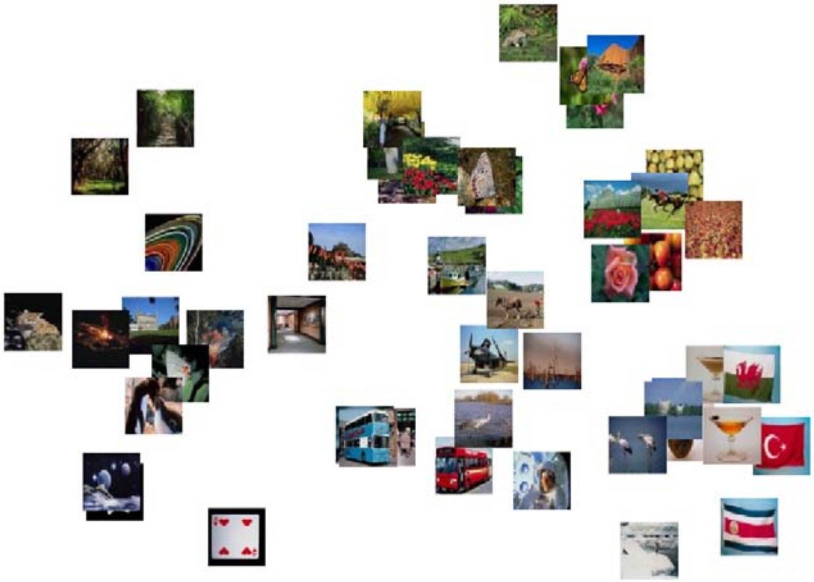


Figure 5.9.: 2D visualization of the most representative images in the collection

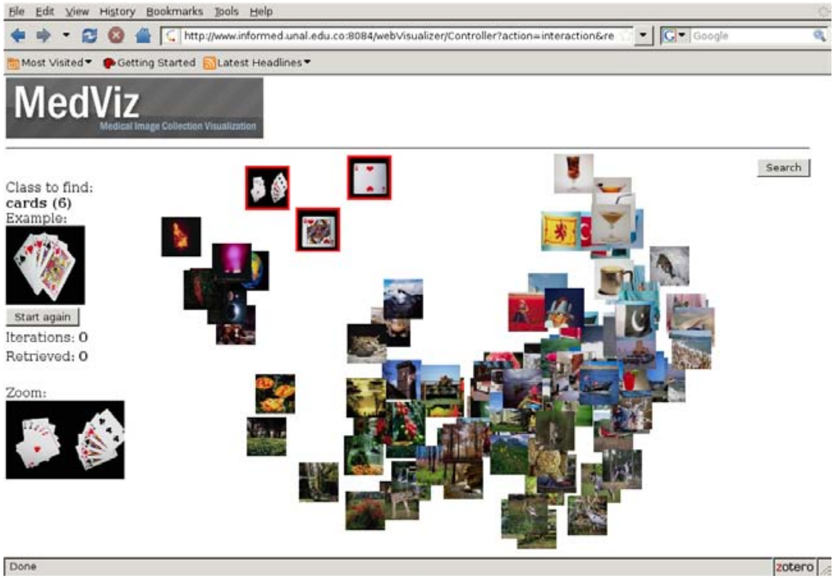
positive examples selected by the user. Surprisingly, this configuration does not lead the best results, because of the lack of discrimination ability between positive and negative examples. When β,γ were set to the same weight, the performance reached a maximum value.

5.4.4.2. Filtering behavior

We also investigate the filtering behavior to identify the influence of the r parameter in the exploration process, which is related to the number of similar images considered for selecting the nearest images to the query at each iteration. Instead of using a decreasing function for this parameter, we fixed it to some constant values to analyze the impact of this parameter in the convergence of the categorical search task. We fixed the category called *buses* for the weighting impact and filtering behavior tasks, and evaluated its performance at each iteration. The lowest performance is obtained with large values of the parameter r, this is because of the search space remains very large among the different iterations and the summary includes images of non-relevant categories. Another extreme is to reduce the search space to a number of images less than the total number of relevant items, that is in our experiments $r = 50$. That means that good results are obtained by the user at the first iterations since the search space is smaller than the appropriate. It is interesting to highlight that the parameter should be set near to the size of the categories, in our case, when $r = 100$, see Figure 5.14.

5.4.4.3. Interaction performance

Finally, the global performance evaluation was conducted with a fixed configuration of the parameters analyzed on the previous two evaluations, selecting those that


File Edit View History Bookmarks Tools Help
http://www.informed.unal.edu.co:8084/webVisualizer/Controller?action=interaction&re
Google
Most Visited
Getting Started
Latest Headlines
MedViz
Medical Image Collection Visualization
Search
Start again
Done
zotero

Class to find:
cards (6)
Example:

Iterations: 0
Retrieved: 0

Zoom:

Figure 5.10.: Screenshot of the system. The user selects relevant images by double clicking on them

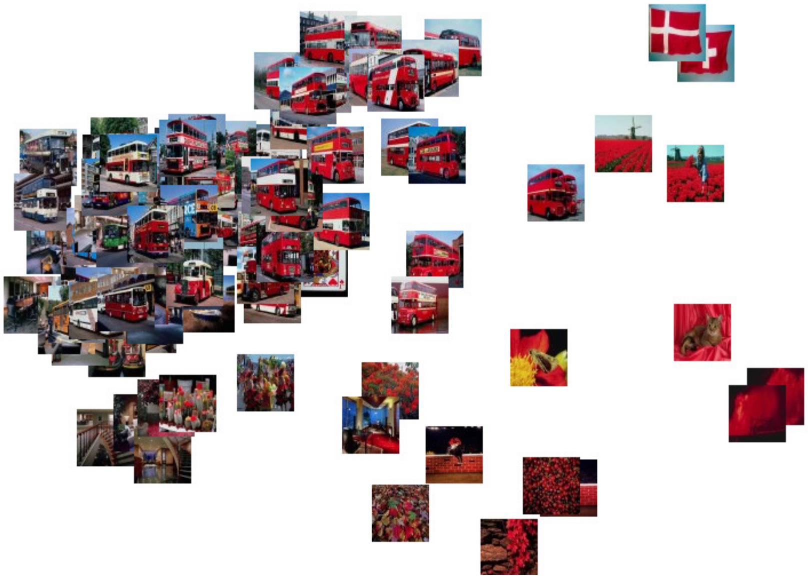


Figure 5.11.: Visualization of the class *buses* at iteration 3. Some images are visually similar one another but they do not belong to the *buses* category.

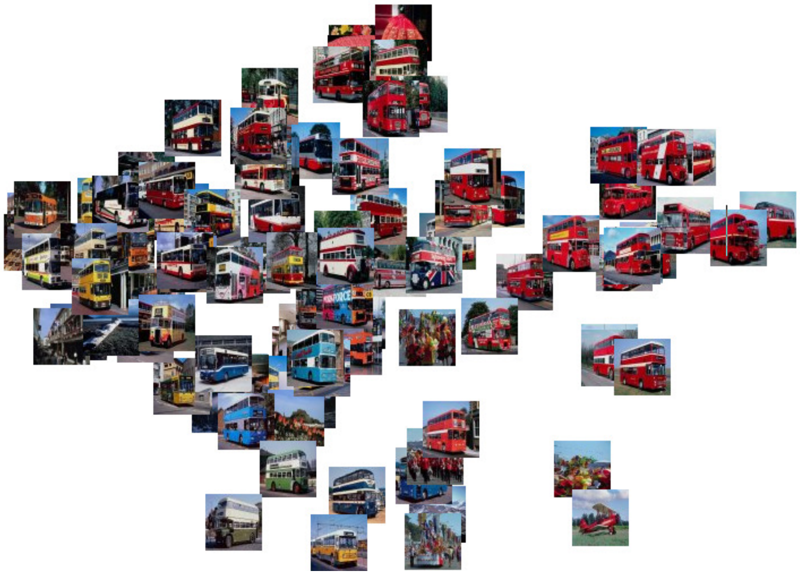


Figure 5.12.: Visualization of the class *buses* at iteration 4. Some images are visually similar one another but they do not belong to the *buses* category.

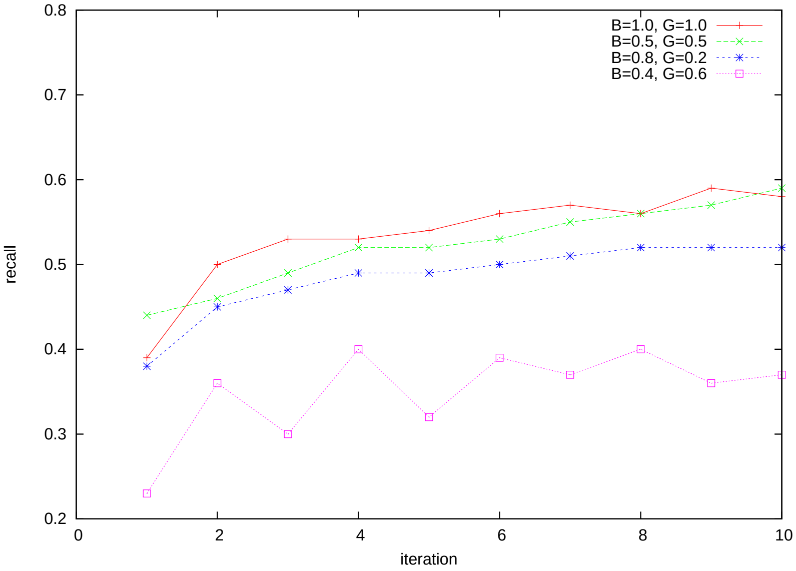


Figure 5.13.: Recall vs. iterations for the *buses* category. Each curve corresponds to different combinations of values for β and γ parameters.

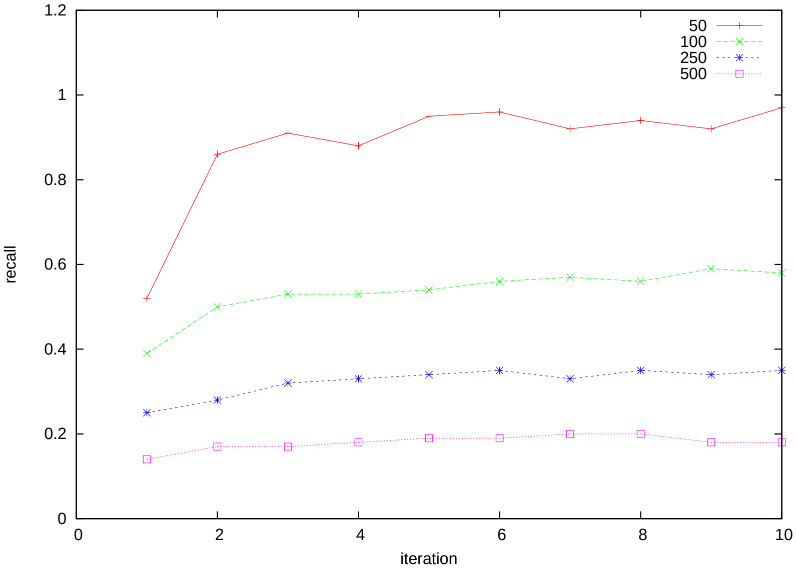


Figure 5.14.: Recall vs. iterations for the *buses* category. Each curve corresponds to a different value of r, the number of shown images.

leads to the best configuration. In the global performance experiments, the user was asked to search images from a category randomly assigned. Figure 5.15 shows different recall plots for several categories. Note that finding images may be less or more difficult according to the underlying target category. For instance, *buses* is easier than the *wildcats*, since in the former only 2 iterations are needed to obtain 80% of recall, while in the later, only the 15% is obtained. Figure 5.16 shows the average recall for all categories in the collection, that shows an increasing recall value when the number of interactions increases.

5.5. Conclusions

This chapter presented a fully functional model for image collection exploration. The main contribution of this chapter is the proposal of a modular framework completely based on kernel methods including, image representation, summarization, and interaction. Kernel methods have shown to be a powerful tool to model image semantics and to build better similarity measures. Traditionally, image collection exploration approaches use feature vector representation to model image content following the standard practice in text retrieval. This feature vector is then used to calculate a standard similarity measure such as cosine similarity. However, in the case of images this is not necessarily the best way to proceed, image similarity calculation could involve more complex procedures that combine the different visual features. The proposed framework exclusively uses image similarity values which are modeled as kernels. So, the proposed approach can be applied to any kind of valid image kernel. This characteristic allowed us to formulate an image similarity learning process that optimally combine different visual kernel to generate a kernel

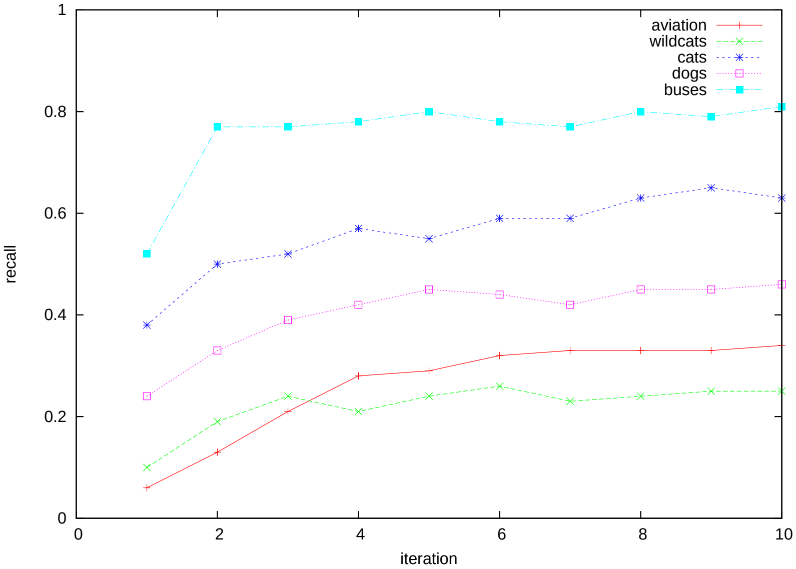


Figure 5.15.: Recall at first 10 iterations. The class *buses* reaches the highest recall, whilst *wildcats* has the lowest.

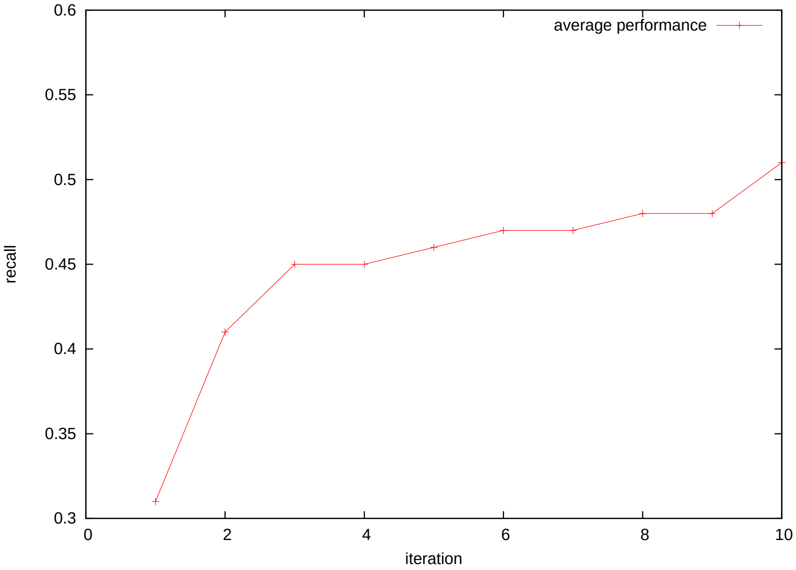


Figure 5.16.: Average recall for all classes at the first 10 iterations.

that better reflects the semantics of the collection. The chapter also contributed the formulation of a kernelized form of the Rocchio's model, which effectively fits in the kernel-based strategy. This kernelized relevance feedback model can be used as baseline in the comparison with future kernel relevance feedback models. We conducted experiments using an image collection with the order of thousands of images, and the results demonstrate the potential of the proposed strategy to provide effective exploratory access to it.

6. Monitoring of Illicit Pill Distribution Networks Using an Image Collection Exploration Framework

This work has been published in the Forensic Science International Journal [138].

This chapter proposes a novel approach for the analysis of illicit tablets based on their visual characteristics. In particular, the chapter concentrates on the problem of ecstasy pill seizure profiling and monitoring. The presented method extracts the visual information from pill images and builds a representation of it, i.e. it builds a pill profile based on the pill visual appearance. Different visual features are used to build different image similarity measures, which are the basis for a pill monitoring strategy based on both discriminative and clustering models. The discriminative model permit to infer whether two pills come from a same seizure, while the clustering model groups of pills that share similar visual characteristics. The resulting clustering structure allows to perform a visual identification of the relationships between different seizures. The proposed approach was evaluated using a data set of 621 Ecstasy pill pictures. The results demonstrate that this is a feasible and cost effective method for performing pill profiling and monitoring.

6.1. Introduction

According to the World Drug Report 2011 published by the United Nations Office on Drugs and Crime [139] amphetamine type stimulants (ATS) represents one of the most significant drug problems worldwide with an annual prevalence ranging between 0.3% and 1.3% of the worldwide population aged 15-64 (13.7 and 56.4 million people). Many efforts have been undertaken during the last decade to propose new methods for highlighting the links between seizures and thereby allowing the deciphering of traffic mechanisms of criminal organizations. Illicit pill monitoring attempts to determine the origin of the pills, the routes used to traffic them and the chemicals used for their production. This process implies establishing links between pills belonging to different seizures. A common assumption made by different illicit pill profiling and monitoring methods is that pills having similar physical and chemical profiles are likely connected. Indeed, large scale projects regrouping different forensic laboratories have investigated the feasibility of deploying harmonized methods for the profiling of amphetamine [140, 141, 142, 143, 144, 145] or methamphetamine [146, 147, 148] by building physical and chemical profiles of pills. New analytical development like Solid Phase MicroExtraction (SPME) [149, 150, 151, 152] or Isotope-Ratio Mass Spectrometry [148, 153, 154] (IRMS) have also been tested for bringing additional knowledge regarding the composition of the illicit drugs seizures. However, ensuring that data obtained by methods such as GC-MS or GC-IRMS are comparable is expensive and time consuming, since it requires the adoption of highly standardized processes. It is therefore important to offer alternative methods that are less expensive, easier to deploy and to maintain in the long term. For instance, the actual philosophy consists in developing databases that can be shared between several laboratories. In turn, physical properties of ATS such as diameter, weight and thickness are readily determined in a first step and were shown useful [155]. The visual features were also showed promising at this stage [156, 157, 158, 159, 160]. Processes to measure these properties or to extract features may easily be performed automatically and do not require tedious and expensive standardization procedures. Therefore, this chapter proposes an agile method for illicit pill profiling and monitoring that relies on the assumption that two illicit pills, which are visually similar, are likely related. In particular, the chapter proposes a pill profiling method based on pill visual appearance, and a pill monitoring strategy that combines a discriminative and clustering model based on pill visual similarity functions. In general, visual similarity depends on different visual features such as color, texture and shape. This chapter explores different image representations and evaluates them according to their performance for discriminative and clustering tasks that support the proposed pill monitoring strategy. The proposed approach offers two main tools for performing pill monitoring: first, a discriminative model that is able to determine whether two different pills belong to the same batch and visual exploration tool, based on a pill clustering, which shows the relationships between different pill and production batches in the pill database. Thus, we provide the users with a fast and reliable method to guide them during their inquiries, while more sophisticated and thereby more expensive methods such as Gas Chromatography coupled to Mass Spectrometry (GC-MS) or IRMS may be used only when a court requires evidence.

6.2. Methodology

6.2.1. Image acquisition

Images of illicit pills were acquired using a Nikon D90 camera with a 5000° K illuminant inside a digital imaging lightbox. The pills were laid on a black photo paper containing a white square of 2.5mm side [21]. Each original image was binarized to discriminate all the objects from the background, while the background of the resulting image was corrected to black; and the largest object, i.e. the pill itself, was extracted. This process is illustrated in Figure 6.1.

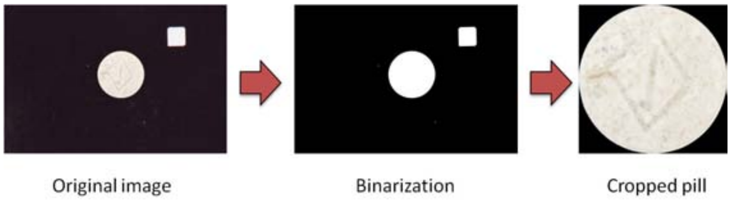


Figure 6.1.: Diagram that illustrates the pre-processing applied to the original pill pictures.

6.2.2. Feature extraction

Two main types of visual features were used to characterize the visual content of images. One based on standard visual features that globally characterize color, texture and shape. These processes were applied to the images acquired as described in the previous Section; and the other one adapted to the particularities of this collection through the construction of a visual codebook. Both strategies are described in the following Subsections (Figure 6.2).

6.2.2.1. Standard visual features

The aim of feature extraction is to identify and extract the relevant information. For this task we follow a Content-based Image Retrieval (CBIR) strategy, where visual patterns such as texture, color and edges were extracted to represent the image content. These features can be represented as histograms, which indicate the frequency associated with visual patterns present in the image. The following standard visual features were extracted:

- *Gray.* A histogram is built by counting the number of occurrences of each one of the 256 intensity values in a gray version of the pill image.
- *Color.* A RGB color space was used to represent pill color that is divided into 512 elements using 8 bins per color channel [161]. A 512 bins histogram (8x8x8=512) is built by counting the number of pixels in the image that correspond to each one of the 512 subspaces.
- *Local binary patterns.* Local binary patterns [162] are a measure of image texture. The goal is to define a set of binary patterns (black and white) that can be found around each pixel. This pattern is identified by analyzing the intensity of the pixel and their 8 nearest neighbors. The intensity of each pixel is compared with the intensity of its 8 neighbors. A value of 1 is assigned to the neighbors with higher intensities, while 0 otherwise. Thus, a word of 8 bits is assigned to each pixel corresponding to a value between 0 and 255.
- *Sobel edges.* The Sobel operator [163] detects image borders by calculating the derivative in one local region of the image by identifying the magnitude and change vector direction. The information codified in the histogram is related to the change in magnitude, which indicates the presence of salient or

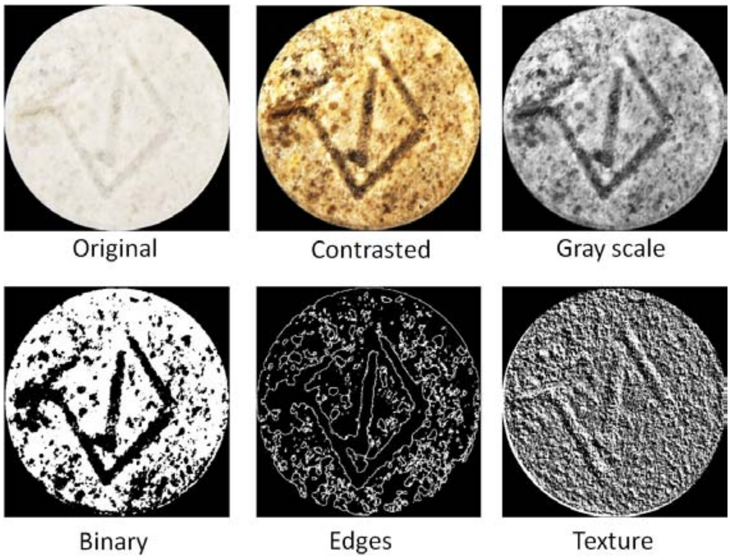


Figure 6.2.: Visual characteristics extracted of a pill.

smooth edges. It is common to use a 3 × 3 operator to determine the changes in intensity of the neighbors and to represent these changes using a 512 bins histogram.

- *Tamura texture.* This visual feature captures texture information such as coarseness, contrast, directionality, linelikeness, regularity and roughness [164]. A 512 bins histogram is built with the space generated by the first three characteristics, as described for the color histogram.
- *Invariant feature.* This type of feature extracts texture characteristics that are invariant to different conditions such as rotation, translation and scale [161].

6.2.2.2. Non-standard features: bag of visual features

The bag of features (BoF) [165] approach is another manner to represent the content of an image. The BoF representation is inspired by the human visual system, since it perceives an object by integrating its constituent components, referred to as patches [166]. An image is thus represented by a histogram that accounts for the amount of small patches present in it. The BoF method involves three steps: (1) Feature detection and description, which consists in dividing the image into small patches, typically of 8x8 pixels. A local descriptor is computed to each patch, commonly descriptors such as the scale-invariant feature transform (SIFT) or the discrete cosine transform (DCT) coefficients. Here we chose to build a vector representation with 192 DCT coefficients for each patch by concatenating the first 64 DCT coefficients obtained for each RGB color channel; (2) Codebook construction, which consists in clustering the resulting representation vectors in order to select a set of k=1000 cluster centroids that are representative. Here, we used a k-means clustering algorithm

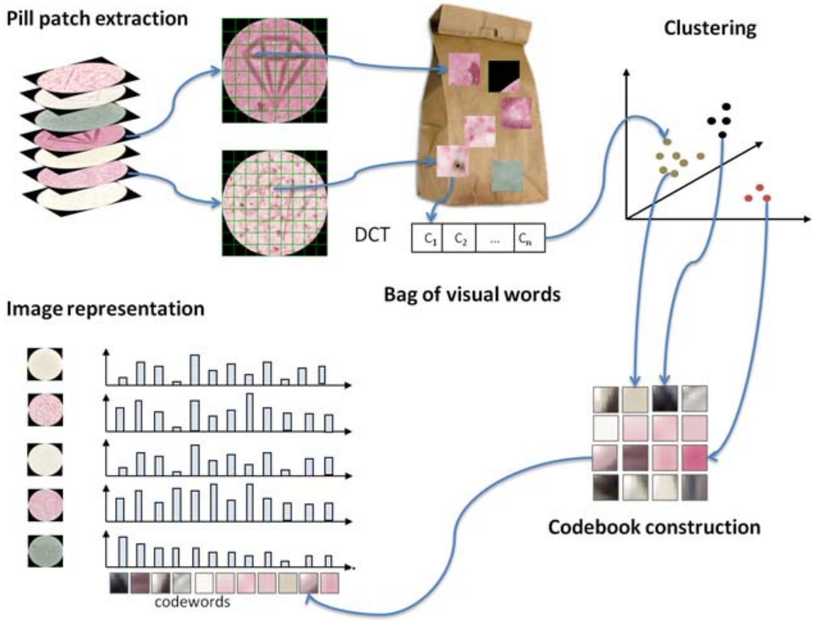


Figure 6.3.: Bag of features process to represent pill Ecstasy images.

and we obtained a 1000 words codebook; and finally (3) BoF representation, which consists in representing each image by a k-bin histogram by finding and counting for each patch the closest word in the codebook. An overview of the BoF process is illustrated in Figure 6.3.

6.2.3. Pill visual similarity

6.2.3.1. Visual feature similarity

Given a particular visual feature, the similarity between two pill images can be obtained by comparing the corresponding histograms extracted from each image for the particular visual feature. There are different alternatives to compare histograms. In this work, we use a methods referred to as histogram intersection [167], which has been shown one of the best metrics to compare histograms in CBIR systems. The histogram intersection between two histograms is defined as,

$$k_{\cap}(h_a, h_b) = \sum_{i=1}^{n} min\left(h_a(i), h_b(i)\right),$$

where h_a and h_b are for histograms a and b and the min function retrieves the minimum value between the i-th bin of both histograms. In other words, the function measures the amount of overlap of the two histograms, if both histograms are identical the overlap is maximum.

The resulting similarities between pills can be represented using a distance matrix, as illustrated in Figure 6.4 for a set of 237 ecstasy pills. Each cell (pixel) represents

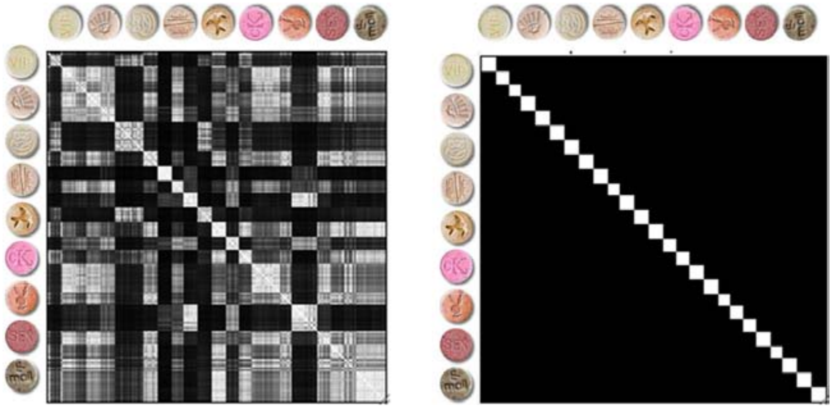


Figure 6.4.: Similarity matrix for a collection of 237 pills obtained using the histogram intersection of Gray visual feature (left). As an example, the ideal distance matrix obtained using a priori knowledge (right).

the similarity between the i-th pill and the j-th pill. The whiter the intensity of the pixel, the higher the similarity between the corresponding pills.

6.2.3.2. Evaluation of the similarity measures

Evaluating different similarity measure is a difficult task. Some methods might discriminate between groups of pills while others not and vice versa. Therefore, statistical tools are required that enable a rigorous evaluation. The simplest analysis consists in representing graphically the distributions of similarity values obtained for both black (negative distribution) and white pixel (positive distribution) of the target matrix shown in Figure 6.4 (right). A perfect method will lead to non overlapping distribution, that is, a threshold value can be chosen that unambiguously classify the intra-batch from the inter-batch similarity values. A more sophisticated approach consists in reporting the rate at which true positive events occurs versus the false positive rate. In that case, the perfect method is the one with a true positive rate of 1 and a false positive rate of 0. Again, this condition only occurs if a threshold value allows to discriminate unambiguously inter- from intra-batch values. In real conditions, this analysis, referred to as ROC (Receiver Operating Characteristics) curves allows, indeed, to determine the threshold value that meet users' expectations, that is the best compromise between true and false positive rate. Another possibility is to evaluate the impact of similarity in other tasks such as clustering. This impact can be analyzed measuring the entropy of the resulting groups of pills after the clustering process. A perfect clustering is that in which images of the same batch are clustered in the same group, that is to say, entropy measures the degree to which each cluster consists of pills of a single batch.

6.2.3.3. Area overlap

The goal of the similarity learning task is to find a similarity measure of visual characteristics that assigns a high value to pills from the same production batch (in our case pills coming from a same seizure), while returning a small value otherwise (pills coming from seizures knowing being unrelated by law enforcement information). Figure 6.4 (right) shows the ideal similarity matrix or target matrix. This latter allows us to measure the discrimination power obtained with different visual features. The power of discrimination (POD) for a given similarity function is evaluated as explained below:

1. The different intra-batch similarities are recorded in a 100-bins histogram, h_{intra}.
2. Another histogram, h_{inter}, is built in a similar way as the first one, but this time sampling pairs of pills from different batches, i.e., inter-batch similarities.
3. Both histograms were normalized with L_1-norm.
4. Both histograms are compared calculating the amount of overlap using histogram intersection: $K_\cap(h_{inter}, h_{intra})$. Figure 6.5 shows an illustration of h_{inter} and h_{intra} for a particular visual pill similarity function.

6.2.3.4. ROC analysis

To evaluate the ability of the a particular visual similarity function to discriminate whether a pair of pills belong to the same batch or not, we used the area under the ROC (Receiver Operating Characteristics) curve [168]. In this context, a visual similarity function along with a given threshold is evaluated as a classifier as follows:

- *True positive* (tp): the pills belong to the same batch and their similarity is higher or equal than the threshold,
- *True negative* (tn): the pills belong to different batches and their similarity is lower or equal than the threshold,
- *False positive* (fp): the pills belong to the different batches and their similarity is higher or equal than the threshold,
- *False negative* (fn): the pills belong to the same batch and their similarity is higher than the threshold.

To construct the ROC curve, the threshold value is varied within a given interval, calculating the corresponding values for the sensitivity (true positive rate, $TPR = TP/(TP + FN)$) and specificity (false positive rate, $FPR = FN/(FP + TN)$), which are then plotted to form the ROC curve.

6.2.3.5. Entropy

We want to measure the impact of visual similarity functions in the construction of pill clusters. We used the entropy measure [169], which is a classification-oriented measure commonly used to validate cluster quality taking into account cluster labels, in our case, a priori batch information. Entropy measures the degree to which each cluster consists of pills of a single batch. A cluster with high entropy has pills from different batches, a cluster with low entropy is more homogeneous, i.e., most of the pills belong to the same batch, thus, a lower entropy value is preferred. The entropy of a particular clustering is calculated as follows. For cluster j we compute P_{ij}, the probability that a pill of cluster i belongs to batch j as $P_{ij} = m_{ij}/m_i$, where m_i is the number of objects in cluster i and m_{ij} is the number of objects of pills of batch j in cluster i. Using this batch distribution, the entropy of each cluster i is calculated using $e_i = \quad \sum_{i=1}^{K}(m_i/m)e_i$, where K is the number of clusters and m is the total number of pills.

6.3. Results and discussions

In order to evaluate our strategy, we build a collection of 621 pictures of ecstasy pills consisting of 215 batches i.e. 187 batches of 2 pills, 2 batches of 3 pills, 1 batch of 4 pills, 4 batches of 8 pills, 5 batches of 9 pills and 16 batches of 10 pills. A second collection of images was prepared by resizing the original pictures to 256x256 pixels in order to accelerate the extraction process.

6.3.1. Comparison of visual similarity functions

The POD for each visual similarity function was evaluated as described in Section 6.2.3.3. Pairs of pills from the training set belonging to the same batch were randomly selected and their similarity measured using a given similarity function. The results obtained for intra-batch distances (corresponding to the white pixels in Figure 6.4 (right)) were represented by a histogram h_intra, while the resulting inter-batch distances (black pixels in Figure 6.4 (right)) were represented by a histogram h_inter Superimposing both histogram permits to visualize the POD. Indeed, overlapped histograms means poor POD, while well separated histograms means high POD. Example of such histograms are presented in Figure 6.5, while Table 6.1 shows the obtained histogram intersection for each visual similarity function. Note that the BoF representation led to the lowest overlap, 10.59%, followed by RGB visual similarity function with 14.81%.

The POD for each visual similarity function was also evaluated using a ROC analysis, as discussed in Section 6.2.3.2. Table 6.1 shows both the area overlap and the area under the ROC curve (AUC).

Figure 6.6 shows the ROC curve computed for each visual feature, and the column "Area under ROC curve" of Table 1 shows the corresponding area under the ROC curves [168]. With the exception of linear binary pattern and color, the visual similarity functions were ranked in the same order by both area overlap and AUC. The BoF visual similarity function was best ranked by the three proposed evaluation methods, that is, area overlap, ROC analysis, and clustering entropy. This result can be attributed to the capacity of BoF to find patterns that are characteristic from an image collection, which then constitute the codebook. Indeed, this approach was shown successful in other domains such as natural scene analysis, medical image annotation and content-based image retrieval.

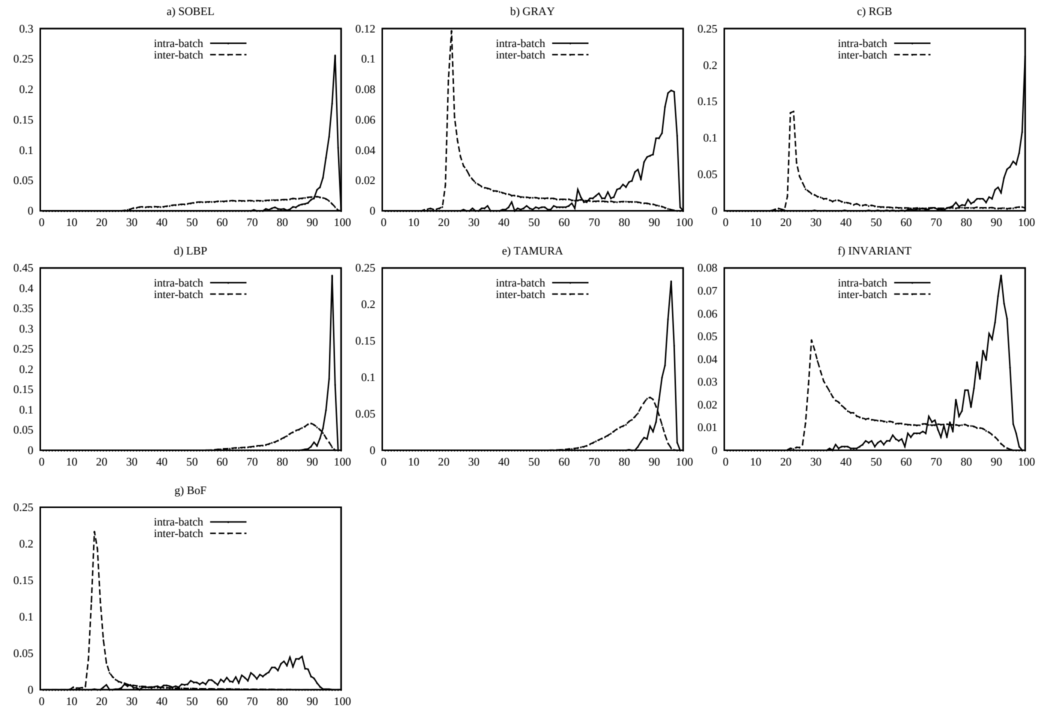


Figure 6.5.: Overlap between inter-batch and intra-batch distributions obtained using (a) sobel edge, (b) gray, (c) color, (d) invariant feature, (e) local binary patterns, (f) tamura texture and (g) bag of feature.

Visual feature	Area	Area under	Area under
	overlap	ROC curve	entropy curve
Bag of features	10.59%	0.9874	327.843
Local binary pattern	17.72%	0.9713	369.247
Tamura texture	26.51%	0.9421	430.746
Gray	23.13%	0.9514	433.599
Sobel edges	24.71%	0.9493	475.048
Invariant feature	35.00%	0.9076	499.384
Color (RGB)	14.81%	0.9689	533.384

Table 6.1.: Performance evaluation of each visual similarity function.

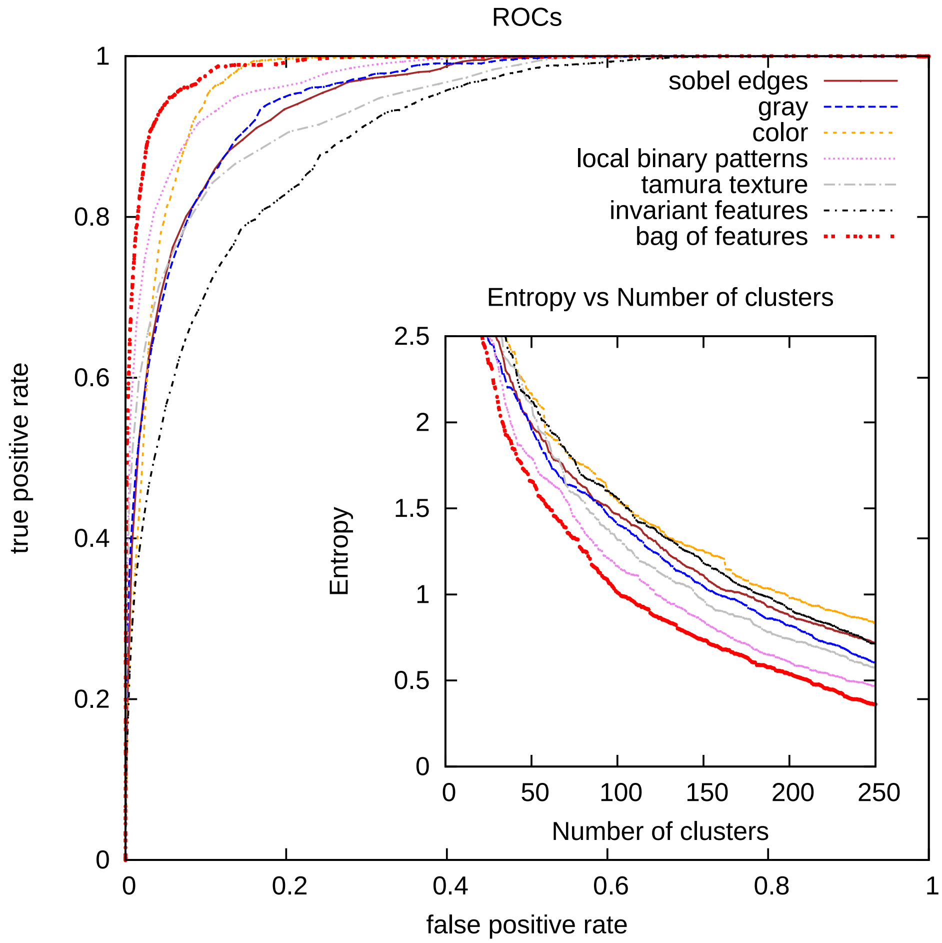


Figure 6.6.: ROC curves for each visual similarity functions. Insert: corresponding entropy curves.

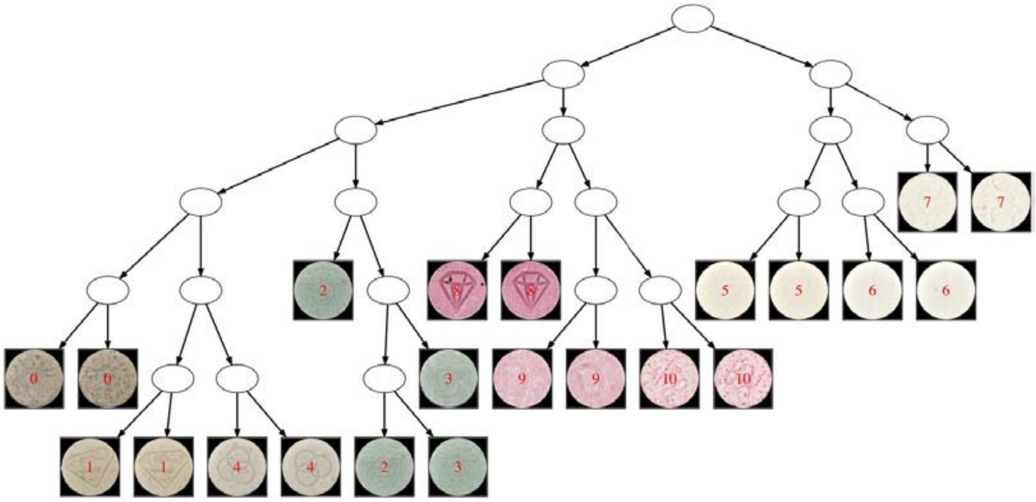


Figure 6.7.: Dendrogram generated by hierarchical clustering. Number inside indicates the corresponding batch membership.

6.3.2. Impact of visual similarity function in clustering

Cluster analysis is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other that to the ones in other clusters. Clustering is a common technique for statistical data analysis used for knowledge discovery. In this chapter we used a hierarchical clustering algorithm called single-linkage to analyze relationships among ecstasy pills. Hierarchical clustering creates a hierarchy of clusters that can be represented by a tree structure called dendrogram. The algorithm starts by merging individual elements into clusters and progressively merges the resulting new clusters according to the distance between them. In our case, the similarity matrix is used as input to the algorithm since it reflects the distance (similarity) between pill images. Figure 6.7 shows an example of a resulting dendrogram. For illustration purposes we superimposed the batch number inside 22 pills. Note that images of batches 2 and 3 are automatically organized in the same subtree, which allows revealing connections between tablets of different batches.

We also performed an objective evaluation of the impact of visual similarity functions in clustering using the entropy measure described in Section 6.2.3.5. We calculated the entropy varying the number of clusters, that is, we cut the dendrogram obtained for each visual similarity function and calculated its entropy. Figure 6.6b shows a plot of the entropy versus the number of cluster. This plot was generated for ranges X: [5,250] clusters and Y:[0,2.5] entropy. The column "Area under entropy curve" in Table 6.1 shows the area under entropy curves for each visual feature. Note that BoF obtains the lowest value, which means that this feature is the best one for clustering as was described in Section 6.2.3.2.

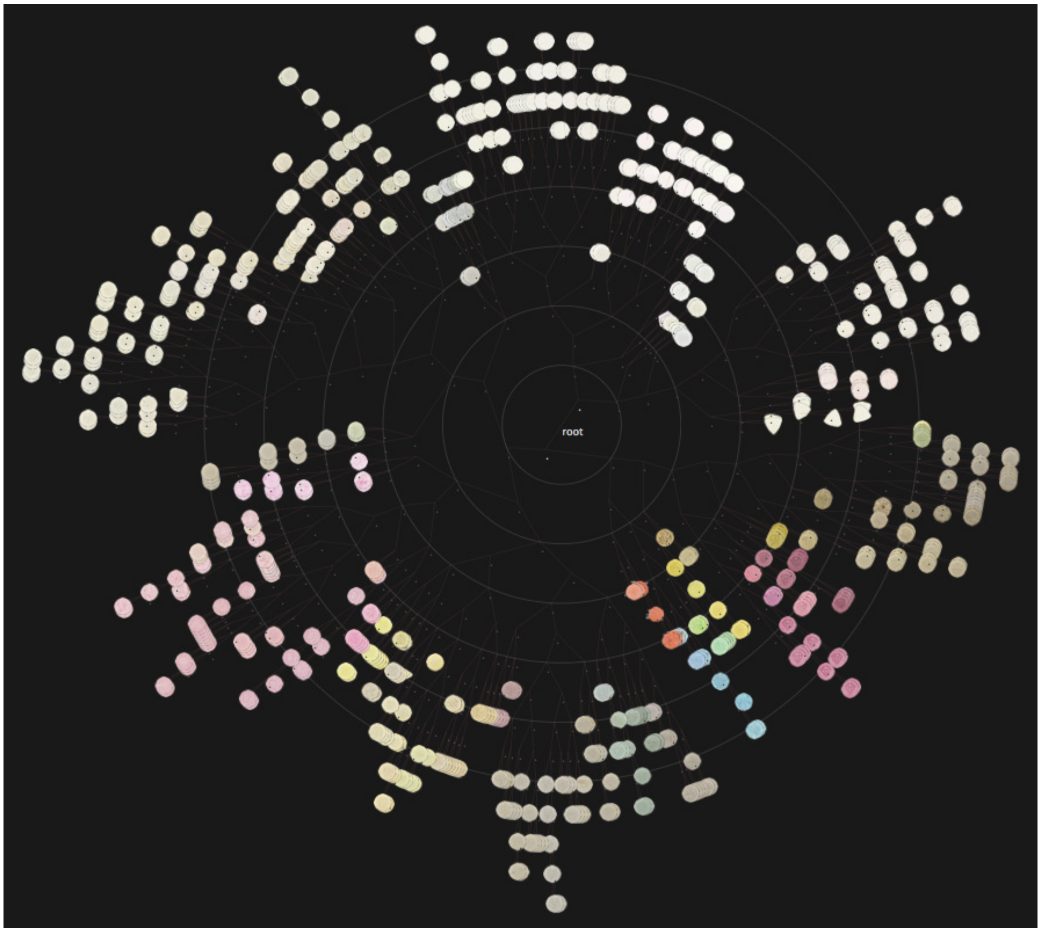


Figure 6.8.: Ecstasy pill explorer prototype. 621 pills can be accessed using this exploration tool.

6.3.3. Exploration of the image collection

The hierarchical structure of the obtained dendrogram allows to organize the image collection in a way that can be exploited to explore the relationships between tablets. We are interested in offering exploration tools that allow the user to visually navigate the image dataset. To reach this objective we developed a component based on the JavaScript InfoVis Toolkit[1] , which is an information visualization component that allows to visualize and explore a hierarchy of objects. In this prototype, see Figure 6.8, we deployed the complete image dataset. The component uses a radial tree layout [170], in which the view is determined by the selection of a focus node. The algorithm used by the component is based on the radial layout method [171] by linearly interpolating the polar coordinates of the nodes and constraining the layout during interactions to keep it as similar as possible to the previous layout.

However, when the number of images is large, it is necessary to display a summary of the collection, i.e., a subset of the collection that is representative and that can be

[1] http://www.thejit.org

used as starting point in the exploration process. In a future work we will address this issue by finding a representative image (centroid) to be displayed instead of all the members of the group.

Finally, an important aspect, of the profiling problem is the fact that images of different seizures can belong to the same production batch. Additional information (geo-location, infra-red spectra, etc.) can be used to build complementary visual similarity functions and combined with the existing information to achieve more accurate clustering, if necessary.

6.4. Conclusions

This chapter presented a new approach for the profiling of illicit tablets using pictures of ecstasy pills. This powerful tool might help finding links between illicit drugs seizures by exploring the collection of image to decipher the structure of criminal organizations behind this traffic. The costs for its implementation are low in comparison with more sophisticated methods such as GC-MS and GC-IRMS, since it only relies on the visual characteristics of the pills. Interestingly, this little and easy to get information permit to efficiently group pills according to their seizures. As expected, the similarity function that allows to extract features that are characteristic of this data set performed better than the similarity functions based on standard features. In addition, supplementary information, such as IR spectra, can readily be included in our procedure, thereby allowing to refine the results if necessary. We are currently working to include different kind of additional information such as seizure reports (date of seizure, place, active substances and cutting agent) and IR spectra to offer the users with the opportunity to select different levels of analysis according to their needs and to the availability of the data.

7. Conclusions

This thesis has presented a strategy to address the problem of involving multimodal information in the construction of each stage of an image collection exploration system: representation, summarization, visualization and interaction. This thesis focused on combining visual and textual modalities in the same representation space by using machine learning and image processing techniques. The proposed methods were evaluated with performance measures, which allowed to determinate the performance of such methods in an objective way. This thesis evaluated the proposed strategy in different contexts such as natural images, biomedical imaging, and drug intelligence. Results show that combining different information sources improve the quality of each stage of an image collection exploration system.

The main contribution of this research work is a family of algorithms and models based on kernel methods and matrix factorization, which allowed to fuse both modalities in the same representation space outperforming the use of only one modality. The following subsections discuss different aspects of the addressed problems and of the strategies used to tackle them.

7.1. Image collection representation

This dissertation has studied different mechanisms to represent images by combining visual and textual content. This stage is very important because it is the basis of the other stages. To combine both information sources, this work proposed a set of methods based on matrix factorization. Different visual representations were studied such low-level features (color, edges, textures, etc.) and the bag-of-visual-words model, in combination with the vector space model to index text content. Both modalities can be fused by combining the obtained feature vectors (visual and textual) or by factorizing one modality to obtain a basis to factorize the other modality as it was presented in Chapter 4 and Chapter 3. This fusion can be also obtained by calculating separately the kernel functions for each modality and summing them to obtain a combined kernel matrix as was described in Chapter 5.

7.2. Image collection summarization

In this work, multimodal image collection summarization is approached by selecting a set of prototypical images from a larger set of images. To perform this task, this thesis proposed the MICS algorithm presented in Chapter 3, which was compared to other state-of-the-art methods outperforming them. This comparison was made by means of objective measures that allow to evaluate the performance of the obtained summary in terms of its reconstruction ability and diversity of the summary.

A summarization experiment was also conducted to automatically find groups of Ecstasy pill images that share visual properties in the analysis of illicit pill distribution networks. The power of discrimination of the proposed method was evaluated using clustering-based measures and ROC analysis as it is described in Chapter 6. Results were promising and published in one of the main journals of the forensic science.

7.3. Image collection visualization

Visualizing image results is a very important task, which uses a visualization metaphor to represent image relationships. As was described in Chapter 2, there are different visualization metaphors to visualize image results. This thesis proposed a new way of visualizing images and text terms in the same visualization space as it was presented in Chapter 4. The proposed method is based on matrix factorization to build a latent space in which images and text terms can be represented together. The proposed strategy was evaluated in an objective way. The main advantage of the proposed approach is the ability to locate text terms in a 2D canvas that can guide the user in the exploration process. This result makes an important difference among the state-of-the-art methods for image collection exploration, that are mainly based on visual features without a clear identification of the regions in the screen.

7.4. Image collection interaction

The interaction stage of an image collection exploration system consists in a process in which the user interacts with the image collection. This interaction allows the system to capture user's feedback. Once the system visualizes image results, the user navigates the collection by selecting images that may be related with her/his needs as it was presented in Chapter 5. These interaction happens in an iterative way and can be used to refine image results. In this thesis, a system prototype was build which was named *MedViz*. The prototype allowed to evaluated with real users how the proposed interaction system performed while the user provided feedback. A kernelized version of the Roccio's model was proposed, which obtained good results in the first iterations. This functionality was also implemented in the *Informed* system described in Chapter 1, in which results of a query are visualized in a 2D visualization metaphor.

An exploration system prototype was also built in the context of drug intelligence, which was presented in Chapter 6. This prototype allowed to visualize a summary obtained from an image pill collection in order to automatically detect relationships between different production bathes of Ecstasy pills.

7.5. Future work

This thesis has addressed important challenges of image collection exploration from a multimodal perspective showing interesting results that motivate the continuation of the work. There are other important aspects that are also challenging and need to be addressed: (1) Large-scale of current image collections is one of the main concerns of image collection exploration. The size of nowadays collections is huge and growing, so existent methods need to be adapted to deal with this problem. As future work, it would be interesting to experiment the proposed methods with larger data sets to validate scalability and whether it is necessary to adapt such methods; (2) Information visualization is an active area that aims to create new visualization paradigms according to the user needs and information context. In particular, multimodal visualization metaphors in which more than one modality can be exploited; (3) With the increasing use of mobile devices, current visualization paradigms are not suitable to explore image collections. Users are now accessing search engines through mobile browsers, which use the same desktop-browser paradigm. However, there are device restrictions such as screen size, navigation controls, memory, among others, which arise new challenges in the construction of image collection exploration systems; and (4) User evaluation is another research area in which different issues have to be addressed. This thesis performed some user evaluations in controlled environments. However, massive user evaluations need to be conducted to validate the experimental results obtained in this thesis, which were addressed mainly from a quantitative perspective.

Bibliography

[1] J. Camargo and F. Gonzalez, "Visualization of large collections of medical images," in *VI Congreso Colombiano de Computacion, 4CCC*, 2009.

[2] H. Liu, X. Xie, X. Tang, Z.-W. Li, and W.-Y. Ma, "Effective browsing of web image search results," in *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval.* New York, NY, USA: ACM Press, 2004, pp. 84–90. [Online]. Available: http://dx.doi.org/10.1145/1026711.1026726

[3] T. Jankun-Kelly and K.-L. Ma., "Moiregraphs: Radial focus+context visualization and interaction for graphs with visual nodes," in *In Proceedings of 2003 Symposium on Information Visualization*, 2003, pp. 8–15.

[4] M. Porta, "Browsing large collections of images through unconventional visualization techniques," in *AVI '06: Proceedings of the working conference on Advanced visual interfaces.* New York, NY, USA: ACM Press, 2006, pp. 440–444. [Online]. Available: http://dx.doi.org/10.1145/1133265.1133354

[5] A. Del Bimbo, "A perspective view on visual information retrieval systems," *Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on*, pp. 108–109, Jun 1998.

[6] J. Zhang, *Visualization for Information Retrieval.* Springer, 2008.

[7] J. D. Stuart K. Card and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann Publishers, 1999.

[8] G. P. Nguyen and M. Worring, "Interactive access to large image collections using similarity-based visualization," *Journal of Visual Languages & Computing*, vol. 19, no. 2, pp. 203–224, April 2008.

[9] M. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17(4), pp. 401–419, 1958.

[10] I. Jolliffe, "Principal component analysis," *Springer-Verlag*, 1989.

[11] L. S. S. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. v.290 no.5500, pp. 2323–2326, 2000.

[12] J. B. de Silva V. Tenenbaum and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 260, pp. 2319–2323, 2000.

[13] T. Kohnen, “Self-organizing maps,” *Springer Series in Information Sciences*, vol. 30, 2001.

[14] P. Demartines and J. Herault, “Cca: Curvilinear component analysis,” in *15th Workshop GRETSI*, 1995.

[15] J. A. Lee, *Nonlinear Dimensionality Reduction*, ser. Information Science and Statistics. Springer New York, 2007.

[16] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15(6), pp. 1373–1396, 2003.

[17] C. A. J. Lee and M. Verleysen, “Locally linear embedding versus isotop,” *11th European Symposium on Artificial Neural Networks*, pp. 527–534, 2003.

[18] J. Sammon, “A nonlinear mapping algorithm for data structure analysis,” *IEEE Transactions on Computers*, vol. 18(5), pp. 401–409, 1969.

[19] G. Hinton and S. Roweis, “Stochastic neighbor embedding,” *Advances in Neural Information Processing Systems 15*, pp. 833–840, 2003.

[20] J. Shawe Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[21] H. Choi and S. Choi, “Robust kernel isomap,” *Pattern Recognition*, vol. 40, no. 3, pp. 853–862, March 2007. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2006.04.025

[22] S. Sekine and C. Nobata, “A survey for multi-document summarization,” in *Proceedings of the HLT-NAACL 03 on Text summarization workshop - Volume 5*, ser. HLT-NAACL-DUC ’03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 65–72.

[23] D. Stan and I. K. Sethi, “eid: a system for exploration of image databases,” *Inf. Process. Manage.*, vol. 39, no. 3, pp. 335–361, May 2003. [Online]. Available: http://dx.doi.org/10.1016/S0306-4573(02)00131-0

[24] I. Simon, N. Snavely, and S. M. Seitz, “Scene summarization for online image collections,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2007.4408863

[25] A. Gomi, R. Miyazaki, T. Itoh, and J. Li, “Cat: A hierarchical image browser using a rectangle packing technique,” in *Information Visualisation, 2008. IV ’08. 12th International Conference*, 2008, pp. 82–87. [Online]. Available: http://dx.doi.org/10.1109/IV.2008.8

[26] J.-Y. Chen, C. A. Bouman, and J. C. Dalton, “Hierarchical browsing and search of large image databases,” *Image Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 442–455, 2000. [Online]. Available: http://dx.doi.org/10.1109/83.826781

[27] M. Worring, O. de Rooij, and T. van Rijn, “Browsing visual collections using graphs,” in *Proceedings of the international workshop on Workshop on multimedia information retrieval.* New York, NY, USA: ACM, 2007, pp. 307–312. [Online]. Available: http://dx.doi.org/10.1145/1290082.1290125

[28] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, “Hierarchical clustering of www image search results using visual, textual and link information,” *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 952–959, 2004. [Online]. Available: http://dx.doi.org/10.1145/1027527.1027747

[29] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, “Web image clustering by consistent utilization of visual features and surrounding texts,” in *MULTIMEDIA ’05: Proceedings of the 13th annual ACM international conference on Multimedia.* New York, NY, USA: ACM, 2005, pp. 112–121. [Online]. Available: http://dx.doi.org/10.1145/1101149.1101167

[30] H. Nobuhara, “A lattice structure visualization by formal concept analysis and its application to huge image database,” in *Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on*, 2007, pp. 448–452. [Online]. Available: http://dx.doi.org/10.1109/ICCME.2007.4381774

[31] J. Li, J. H. Lim, and Q. Tian, “Automatic summarization for personal digital photos,” in *Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, vol. 3, 2003, pp. 1536–1540 vol.3. [Online]. Available: http://dx.doi.org/10.1109/ICICS.2003.1292724

[32] J. Fan, Y. Gao, H. Luo, D. A. Keim, and Z. Li, “A novel approach to enable semantic and visual image summarization for exploratory image search,” in *Proceeding of the 1st ACM international conference on Multimedia information retrieval.* New York, NY, USA: ACM, 2008, pp. 358–365. [Online]. Available: http://dx.doi.org/10.1145/1460096.1460155

[33] Y. G. J. Fan and H. Luooncept, “Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation,” *IEEE Trans. on Image Processing*, vol. 17(3), pp. 407–426, 2008.

[34] H. L. Y. G. J. Fan and R. Jain., “Incorporating concept ontology to boost hierarchical classifier training for automatic multilevel video annotation,” *IEEE Trans. on Multimedia*, vol. 9(5), pp. 939–957, 2007.

[35] B. Johnson and B. Shneiderman, “Treemaps: A space-filling approach to the visualization of hierarchical information structures,” in *IEEE Information Visualization*, 1991.

[36] B. B. S. B. . W. M. Bederson, “Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies,” *ACM Transactions on Graphics (TOG)*, vol. 21(4), pp. 833–854, 2002.

[37] K. Andrews, W. Putz, and A. Nussbaumer, “The hierarchical visualisation system (hvs),” in *Information Visualization, 2007. IV ’07. 11th International Conference*, 2007, pp. 257–262. [Online]. Available: http://dx.doi.org/10.1109/IV.2007.112

[38] J. Kruskal and M. Wish, “Multidimensional scaling,” *Sage Publications*, 1978.

[39] G. Schaefer and S. Ruszala, “Image database navigation on a hierarchical mds grid,” 2006.

[40] T. Janjusevic and E. Izquierdo, “Layout methods for intuitive partitioning of visualization space,” *Information Visualisation, 2008. IV ’08. 12th International Conference*, pp. 88–93, July 2008. [Online]. Available: http://dx.doi.org/10.1109/IV.2008.55

[41] S. Kullback and R. A. Leibler, “On information and sufficiency, annals of mathematical statistics,” vol. 22, pp. 79–86, 1951.

[42] R. C. Dubes, “How many clusters are best. an experiment,” *Pattern Recognition*, vol. 20(6), pp. 645–663, 1987.

[43] P.-A. Moëllic, J.-E. Haugeard, and G. Pitel, “Image clustering based on a shared nearest neighbors approach for tagged collections,” in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, ser. CIVR ’08. New York, NY, USA: ACM, 2008, pp. 269–278.

[44] R. V. Zwol, V. Murdock, L. G. Pueyo, and G. Ramirez, “G.: Diversifying image search with user generated content,” in *In: Proc. MIR 0́8: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 67–74.

[45] K. Song, Y. Tian, W. Gao, and T. Huang, “Diversifying the image retrieval results.” in *ACM Multimedia*, K. Nahrstedt, M. Turk, Y. Rui, W. Klas, and K. Mayer-Patel, Eds. ACM, 2006, pp. 707–710.

[46] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, “Improving recommendation lists through topic diversification,” in *Proceedings of the 14th international conference on World Wide Web*, ser. WWW ’05. New York, NY, USA: ACM, 2005, pp. 22–32.

[47] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, “Visual diversification of image search results,” in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 341–350.

[48] D. Heesch, “A survey of browsing models for content based image retrieval,” *Multimedia Tools and Applications*, vol. 40, no. 2, pp. 261–284, November 2008.

[49] J. Urban, J. M. Jose, and C. J. Rijsbergen, “An adaptive technique for content-based image retrieval,” *Multimedia Tools Appl.*, vol. 31, no. 1, pp. 1–28, 2006.

[50] L. Kontsevich and B. Calkins, "Xcavator," http://www.xcavator.net, [last visited 2009, April 20].

[51] J. Z. Wang, "Automatic photo tagging and visual image search." [Online]. Available: http://www.alipr.com[lastvisited2007,April21]

[52] M. Lux, "Caliph & emir," http://nixbit.com/cat/multimedia/graphics/caliph–emir/, [last visited 2009, April 21].

[53] D. F. Swayne, D. Temple Lang, A. Buja, and D. Cook, "GGobi: evolving from XGobi into an extensible framework for interactive data visualization," *Computational Statistics & Data Analysis*, vol. 43, pp. 423–444, 2003.

[54] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 985–1002, 2008. [Online]. Available: http://www.hpl.hp.com/personal/Yuli_Gao/google_demo/index.htm[lastvisited2009,April21]

[55] Y. Jing, H. Rowley, J. Wang, D. Tsai, C. Rosenberg, and M. Covell, "Google image swirl: A large-scale content-based image visualization system," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 539–540.

[56] H. Kang and B. Shneiderman, "Visualization methods for personal photo collections: Browsing and searching in the photofinder." in *IEEE International Conference on Multimedia and Expo (III)*, 2000, pp. 1539–1542.

[57] L. Shi, J. Wang, L. Xu, H. Lu, and C. Xu, "Context saliency based image summarization," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 270–273.

[58] Y. Jing, S. Baluja, and H. Rowley, "Canonical image selection from the web," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval.* New York, NY, USA: ACM, 2007, pp. 280–287.

[59] C. Yang, J. Shen, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," *Pattern Recognition*, vol. 46, no. 3, pp. 948 – 961, 2013.

[60] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[61] D. Deng, "Content-based image collection summarization and comparison using self-organizing maps," *Pattern Recognition*, vol. 40(2), pp. 718–727, 2007.

[62] C. Carpineto, M. DÁmico, and G. Romano, "Evaluating subtopic retrieval methods: Clustering versus diversification of search results," *Information Processing & Management*, vol. 48, no. 2, pp. 358 – 373, 2012.

[63] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM, 2003, pp. 267–273.

[64] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proceedings of the 14th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '06. New York, NY, USA: ACM, 2006, pp. 707–710.

[65] F. Sun, M. Wang, D. Wang, and X. Wang, "Optimizing social image search with multiple criteria: Relevance, diversity, and typicality," *Neurocomputing*, vol. 95, no. 0, pp. 40 – 47, 2012.

[66] J. Wang, L. Jia, and X.-S. Hua, "Interactive browsing via diversified visual summarization for image search results," *Multimedia Systems*, vol. 17, no. 5, pp. 379–391, 2011.

[67] R. Raguram and S. Lazebnik. Computing Iconic Summaries of General Visual Concepts.

[68] H. Xu, J. Wang, X.-S. Hua, and S. Li, "Hybrid image summarization," in *Proceedings of the 19th ACM international conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 1217–1220.

[69] Y. Hadi, F. Essannouni, and R. O. H. Thami, "Video summarization by k-medoid clustering," in *Proceedings of the 2006 ACM symposium on Applied computing*, ser. SAC '06. New York, NY, USA: ACM, 2006, pp. 1400–1401.

[70] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[71] Y. Jing, M. Covell, and H. A. Rowley, "Comparison of clustering approaches for summarizing large populations of images," *2013 IEEE International Conference on Multimedia and Expo (ICME)*, vol. 0, pp. 1523–1527, 2010.

[72] Z.-Q. Zhao and H. Glotin, "Diversifying image retrieval with affinity-propagation clustering on visual manifolds," *MultiMedia, IEEE*, vol. 16, no. 4, pp. 34–43, 2009.

[73] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma, "Improving web search results using affinity graph," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 504–511.

[74] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[75] A. Bosch, X. Muñoz, and R. Martí, "Which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, no. 6, pp. 778–791, June 2007.

[76] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[77] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval.* New York, NY, USA: ACM, 2008.

[78] J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon, "Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces," in *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval.* New York, NY, USA: ACM, 2008, pp. 359–368.

[79] J. E. Camargo, J. C. Caicedo, and F. A. Gonzalez, "Multimodal image collection visualization using non-negative matrix factorization," in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science, M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, Eds. Springer Berlin Heidelberg, 2010, vol. 6273, pp. 429–432.

[80] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, August 2000.

[81] I. Bartolini, P. Ciaccia, and M. Patella, "Pibe: Manage your images the way you want!" *Data Engineering, International Conference on*, vol. 0, pp. 1519–1520, 2007.

[82] H. Ding, J. Liu, and H. Lu, "Hierarchical clustering-based navigation of image search results," in *MM '08: Proceeding of the 16th ACM international conference on Multimedia.* New York, NY, USA: ACM, 2008, pp. 741–744.

[83] S. Marchand-Maillet, E. Bruno, and E. Unig, "State of the art image collection overviews and browsing."

[84] "Google image swirl," http://image-swirl.googlelabs.com.

[85] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* New York, NY, USA: ACM, 2003, pp. 119–126.

[86] J. Hare, P. Lewis, P. Enser, and C. Sandom, "A linear-algebraic technique with an application in semantic image retrieval," *Lecture Notes in Computer Science*, vol. 4071, p. 31, 2006.

[87] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *Multimedia, IEEE Transactions on*, vol. 9, no. 5, pp. 923–938, 2007.

[88] D. D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[89] J. Tang and P. H. Lewis, "Non-negative matrix factorisation for object class discovery and image auto-annotation," in *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval.* New York, NY, USA: ACM, 2008, pp. 105–112.

[90] A. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, p. 4, 2009.

[91] C. Ding, X. He, and H. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. SIAM Data Mining Conf.* Citeseer, 2005, pp. 606–610.

[92] H. Muller, S. Marchand-Maillet, and T. Pun, "The truth about corel-evaluation in image retrieval," *Lecture notes in computer science*, pp. 38–49, 2002.

[93] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, p. 91, Nov. 2004.

[94] J. E. Camargo, J. C. Caicedo, and F. A. Gonzalez, "A kernel-based framework for image collection exploration," *J. Vis. Lang. Comput.*, vol. 24, no. 1, pp. 53–67, Feb. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jvlc.2012.10.008

[95] R. J. D. L. J. . W. J. Z. Datta, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40(2), pp. 1–60, 2008.

[96] K. Borner, "Extracting and visualizing semantic structures in retrieval results for browsing," in *DL '00: Proceedings of the fifth ACM conference on Digital libraries.* New York, NY, USA: ACM, 2000, pp. 234–235. [Online]. Available: http://dx.doi.org/10.1145/336597.336672

[97] C. Chen, G. Gagaudakis, and P. Rosin, "Content-based image visualization," *iv*, vol. 00, 2000. [Online]. Available: http://dx.doi.org/10.1109/IV.2000.859730

[98] Y. Chen, J. Z. Wang, and R. Krovetz, "Clue: cluster-based retrieval of images by unsupervised learning," *Image Processing, IEEE Transactions on*, vol. 14, no. 8, pp. 1187–1201, 2005. [Online]. Available: http://dx.doi.org/10.1109/TIP.2005.849770

[99] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, 2003, pp. III–513–16 vol.2.

[100] Z. Harchaoui and F. Bach, "Image classification with segmentation graph kernels," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[101] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2169–2178.

[102] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[103] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1–8.

[104] Y. Zhi-qin, S. Song-zhi, and L. Shao-zi, "Research on branch and bound for pedestrian detection," in *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*, vol. 2, june 2011, pp. 366 –370.

[105] J. Wu and J. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009, pp. 630 –637.

[106] J. Almeida, N. J. Leite, and R. da S. Torres, "Vison: Video summarization for online applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397 – 409, 2012, intelligent Multimedia Interactivity.

[107] S. Boughorbel, J.-P. Tarel, and N. Boujemaa, "Generalized histogram intersection kernel for image recognition," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3, sept. 2005, pp. III – 161–4.

[108] L. Tang, G. Hamarneh, and T. Bressmann, "A machine learning approach to tongue motion analysis in 2d ultrasound image sequences," in *Machine Learning in Medical Imaging*, ser. Lecture Notes in Computer Science, K. Suzuki, F. Wang, D. Shen, and P. Yan, Eds. Springer Berlin / Heidelberg, 2011, vol. 7009, pp. 151–158.

[109] J. Wu, "A fast dual method for hik svm learning," in *Proceedings of the 11th European conference on Computer vision: Part II*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 552–565.

[110] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[111] K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM, 2004, p. 106.

[112] C. Chen, G. Gagaudakis, and P. Rosin, “Similarity-based image browsing,” in *Proceedings of the 16th IFIP World Computer Congress (International Conference on Intelligent Information Processing)*, ser. pp. 206-213. Beijing, China: Proceedings of the 16th IFIP World Computer Congress (International Conference on Intelligent Information Processing), 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.5701

[113] B. Liu, W. Wang, J. Duan, Z. Wang, and B. Shi, “Subsequence similarity search under time shifting,” in *Information and Communication Technologies, 2006. ICTTA ’06. 2nd*, vol. 2, 2006, pp. 2935–2940. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1684881

[114] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance feedback: A power tool for interactive content-based image retrieval,” 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.4695

[115] C. Nastar, M. Mitschke, and C. Meilhac, “Efficient query refinement for image retrieval,” in *CVPR ’98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 1998, p. 547.

[116] K. Porkaew, S. Mehrotra, and M. Ortega, “Query reformulation for content based multimedia retrieval in mars,” in *ICMCS ’99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems*. Washington, DC, USA: IEEE Computer Society, 1999, p. 747.

[117] G. P. Nguyen and M. Worring, “Optimization of interactive visual-similarity-based search,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 1, pp. 1–23, 2008. [Online]. Available: http://dx.doi.org/10.1145/1324287.1324294

[118] M. Y. Chen, M. Christel, A. Hauptmann, and H. Wactlar, “Putting active learning into multimedia applications: dynamic definition and refinement of concept classifiers,” in *Proceedings of the 13th annual ACM international conference on Multimedia*. New York, NY, USA: ACM, 2005, pp. 902–911. [Online]. Available: http://dx.doi.org/10.1145/1101149.1101342

[119] P. G. Matthieu and M. Cord, “Retin al: An active learning strategy for image category retrieval,” in *In Proc. IEEE Conf. Image Processing (Singapore*, 2004, pp. 2219–2222. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.5562

[120] I. Campbell, “The ostensive model of developing information needs,” Ph.D. dissertation, University of Glasgow, 2000.

[121] J. Li and N. M. Allinson, “Long-term learning in content-based image retrieval,” *International Journal of Imaging Systems and Technology*, vol. 18, no. 2-3, pp. 160–169, 2008.

[122] J. Fan, Y. Gao, and H. Luo, “Hierarchical classification for automatic image annotation,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR ’07. New York, NY, USA: ACM, 2007, pp. 111–118.

[123] J. J. Foo, J. Zobel, and R. Sinha, “Clustering near-duplicate images in large collections,” in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, ser. MIR ’07. New York, NY, USA: ACM, 2007, pp. 21–30.

[124] K. Sawase, H. Nobuhara, and B. Bede, “Visualizing huge image databases by formal concept analysis,” in *Human-Centric Information Processing Through Granular Modelling*, ser. Studies in Computational Intelligence, A. Bargiela and W. Pedrycz, Eds. Springer Berlin / Heidelberg, 2009, vol. 182, pp. 351–373.

[125] P. H. Gosselin and M. Cord, “RETIN AL: an active learning strategy for image category retrieval,” *Image Processing, 2004. ICIP ’04. 2004 International Conference on*, vol. 4, pp. 2219–2222 Vol. 4, 2004.

[126] D. Liu, X.-S. Hua, L. Yang, and H.-J. Zhang, “Multiple-instance active learning for image categorization,” in *Advances in Multimedia Modeling*, ser. Lecture Notes in Computer Science, B. Huet, A. Smeaton, K. Mayer-Patel, and Y. Avrithis, Eds. Springer Berlin / Heidelberg, 2009, vol. 5371, pp. 239–249.

[127] T. Berg and A. Berg, “Finding iconic images,” in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, june 2009, pp. 1 –8.

[128] R. Tronci, G. Murgia, M. Pili, L. Piras, and G. Giacinto, “Imagehunter: A novel tool for relevance feedback in content based image retrieval,” in *New Challenges in Distributed Information Filtering and Retrieval*, ser. Studies in Computational Intelligence, C. Lai, G. Semeraro, and E. Vargiu, Eds. Springer Berlin / Heidelberg, 2013, vol. 439, pp. 53–70.

[129] G. Wang, D. Hoiem, and D. Forsyth, “Learning image similarity from flickr groups using fast kernel machines,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2012.

[130] B. Thomee and M. Lew, “Interactive search in image retrieval: a survey,” *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 71–86, 2012.

[131] D. ing Sven Siggelkow, D. Prof, D. T. Ottmann, P. Dr, T. Ottmann, B. Haasdonk, L. Bergen, O. Ronneberger, C. B. S. Utcke, and S. Siggelkow, “Feature histograms for content-based image retrieval,” 2002.

[132] I. E. Sobel, “Camera models and machine perception,” Ph.D. dissertation, Stanford, CA, USA, 1970, aAI7102831.

[133] "A flexible image database system for content–based retrieval," *special issue on contentbased access for image and video libraries*, vol. 75, no. 1/2, pp. 175–195, 1999.

[134] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "Optimizing Kernel Alignment over Combinations of Kernel," Department of Computer Science,Royal Holloway, University of London, UK, Tech. Rep., 2002.

[135] R. Vert, "Designing a m-svm kernel for protein secondary structure prediction," Master's thesis, DEA informatique de Lorraine, 2002.

[136] S. B. Kybernetik, A. Smola, B. Schölkopf, E. Smola, K.-R. Müller, L. Bottou, C. Burges, H. Bulthoff, K. Gegenfurtner, and P. Haffner, "Nonlinear component analysis as a kernel eigenvalue problem," 1998.

[137] J. Rocchio, *Relevance feedback in information retrieval*, ser. G. Salton (ed.), The Smart retrieval system: experiments in automatic document processing. Prentice Hall, 1971, pp. 313–323.

[138] J. Camargo, P. Esseiva, F. González, J. Wist, and L. Patiny, "Monitoring of illicit pill distribution networks using an image collection exploration framework," *Forensic Science International*, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.forsciint.2012.10.004

[139] "World Drug Report 2011," United Nations Office on Drugs and Crime, United States, New York, Tech. Rep., 2009.

[140] L. Aalberg, K. Andersson, C. Bertler, H. Borén, M. D. Cole, J. Dahlén, Y. Finnon, H. Huizer, K. Jalava, E. Kaa, E. Lock, A. Lopes, A. Poortman-van der Meer, and E. Sippola, "Development of a harmonized method for the profiling of amphetamines. I. Synthesis of standards and compilation of analytical data," *Forensic science international*, vol. 149, no. 2-3, pp. 219–29, May 2005.

[141] L. Aalberg, K. Andersson, C. Bertler, M. D. Cole, Y. Finnon, H. Huizer, K. Jalava, E. Kaa, E. Lock, A. Lopes, A. Poortman-van der Meer, E. Sippola, and J. Dahlén, "Development of a harmonized method for the profiling of amphetamines. II. Stability of impurities in organic solvents," *Forensic science international*, vol. 149, no. 2-3, pp. 231–41, May 2005.

[142] K. Andersson, E. Lock, K. Jalava, H. Huizer, S. Jonson, E. Kaa, A. Lopes, A. Poortman-van der Meer, E. Sippola, L. Dujourdy, and J. Dahlén, "Development of a harmonised method for the profiling of amphetamines VI: Evaluation of methods for comparison of amphetamine," *Forensic science international*, vol. 169, no. 1, pp. 86–99, Jun. 2007.

[143] K. Andersson, K. Jalava, E. Lock, H. Huizer, E. Kaa, A. Lopes, A. Poortman-van der Meer, M. D. Cole, J. Dahlén, and E. Sippola, "Development of a harmonised method for the profiling of amphetamines: IV. Optimisation of sample preparation," *Forensic science international*, vol. 169, no. 1, pp. 64–76, Jun. 2007.

[144] E. Lock, L. Aalberg, K. Andersson, J. Dahlén, M. D. Cole, Y. Finnon, H. Huizer, K. Jalava, E. Kaa, A. Lopes, A. Poortman-van der Meer, and E. Sippola, "Development of a harmonised method for the profiling of amphetamines V: Determination of the variability of the optimised method," *Forensic science international*, vol. 169, no. 1, pp. 77–85, Jun. 2007.

[145] M. van Deursen, E. Lock, and A. Poortman-van der Meer, "Organic impurity profiling of 3,4-methylenedioxymethamphetamine (MDMA) tablets seized in the Netherlands," *Science & Justice*, vol. 46, no. 3, pp. 135–152, Jul. 2006.

[146] K. Kuwayama, H. Inoue, J. Phorachata, K. Kongpatnitiroj, V. Puthaviriyakorn, K. Tsujikawa, H. Miyaguchi, T. Kanamori, Y. T. Iwata, N. Kamo, and T. Kishi, "Comparison and classification of methamphetamine seized in Japan and Thailand using gas chromatography with liquid-liquid extraction and solid-phase microextraction." *Forensic science international*, vol. 175, no. 2-3, pp. 85–92, Mar. 2008.

[147] H. Inoue, Y. T. Iwata, and K. Kuwayama, "Characterization and Profiling of Methamphetamine Seizures," *Journal of Health Science*, vol. 54, no. 6, pp. 615–622, 2008.

[148] Y. Iwata, K. Kuwayama, K. Tsujikawa, H. Miyaguchi, T. Kanamori, and H. Inoue, "Seized methamphetamine samples with unique profiles of stable nitrogen isotopic composition documented by stable isotope ratio mass spectrometry," *Forensic Toxicol.*, vol. 28, no. 2, pp. 119–123, 2010.

[149] J. Lee, Y. Park, W. Yang, H. Chung, W. Choi, H. Inoue, K. Kuwayama, and J. Park, "Cross-examination of liquid-liquid extraction (LLE) and solid-phase microextraction (SPME) methods for impurity profiling of methamphetamine," *Forensic science international*, vol. 215, no. 1-3, pp. 175–8, Feb. 2012.

[150] K. Kuwayama, K. Tsujikawa, H. Miyaguchi, T. Kanamori, Y. Iwata, H. Inoue, S. Saitoh, and T. Kishi, "Identification of impurities and the statistical classification of methamphetamine using headspace solid phase microextraction and gas chromatography-mass spectrometry." *Forensic science international*, vol. 160, no. 1, pp. 44–52, Jun. 2006.

[151] F. Bonadio, P. Margot, O. Delémont, and P. Esseiva, "Headspace solid-phase microextraction (HS-SPME) and liquid-liquid extraction (LLE): comparison of the performance in classification of ecstasy tablets. Part 2," *Forensic science international*, vol. 182, pp. 52–6, 2008.

[152] F. Bonadio, P. Margot, O. Delemont, and P. Esseiva, "Optimization of HS-SPME/GC-MS analysis and its use in the profiling of illicit ecstasy tablets. Part 1," *Forensic science international*, vol. 187, pp. 73–80, 2009.

[153] H. Buchanan, N. Daeid, W. Meier-Augenstein, H. Kemp, W. Kerr, and M. Middleditch, "Emerging use of isotope ratio mass spectrometry as a tool for

discrimination of 3,4-methylenedioxymethamphetamine by synthetic route," *Anal. Chem.*, vol. 80, no. 9, pp. 3350–3356, 2008.

[154] H. Buchanan, W. Kerr, W. Meier-Augenstein, and N. Daeid, "Organic impurities, stable isotopes, or both: a comparison of instrumental and pattern recognition techniques for the profiling of 3,4-methylenedioxymethamphetamine," *Anal. Meth.*, vol. 3, no. 10, pp. 2279–2288, 2011.

[155] R. Marquis, C. Weyermann, C. Delaporte, P. Esseiva, L. Aalberg, F. Besacier, J. S. Bozenko, R. Dahlenburg, C. Kopper, and F. Zrcek, "Drug intelligence based on MDMA tablets data: 2. Physical characteristics profiling," *Forensic science international*, vol. 178, no. 1, pp. 34–9, Jun. 2008.

[156] Z. Geradts and J. Bijhold, "Content based information retrieval in forensic image databases," *J. Forensic Sci.*, vol. 47, no. 2, pp. 285–292, 2002.

[157] Y.-B. Lee, U. Park, A. K. Jain, and S.-W. Lee, "Pill-ID: Matching and retrieval of drug pill images," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 904–910, May 2012.

[158] D. Kim and J. Chun, "Drug Image Retrieval by Shape and Color Similarity of the Medication," in *2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering.* IEEE, May 2011, pp. 387–390.

[159] M. Lopatka and M. Vallat, "Surface granularity as a discriminating feature of illicit tablets," *Forensic science international*, vol. 210, no. 1-3, pp. 188–94, Jul. 2011.

[160] L. M. Given, S. Ruecker, H. Simpson, and A. Sadler, Elizabeth (Bess) Ruskin, "Inclusive interface design for seniors: Image-browsing for a health information context," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 11, pp. 1610–1617, 2007.

[161] S. Siggelkow and H. Burkhardt, "Improvement of histogram-based image retrieval and classification," in *Object recognition supported by user interaction for service robots*, vol. 3. IEEE Comput. Soc, 2002, pp. 367–370.

[162] L. G. Berman, Andrew P. , Shapiro, "A flexible image database system for content-based retrieval," in *17TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, 1999.

[163] I. Sobel, "Neighborhood coding of binary images for fast contour following and general binary array processing," *Computer Graphics and Image Processing*, vol. 8, no. 1, pp. 127–135, Aug. 1978.

[164] K. D. Deselaers Thomas , "Features for Image Retrieval - A Quantitative Comparison," in *26th DAGM Symposium*, 2004, pp. 228–236.

[165] L. Fei-fei, “A bayesian hierarchical model for learning natural scene categories,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524–531.

[166] I. Biederman, “Recognition-by-components: A theory of human image understanding,” *Psychological Review*, vol. 94, pp. 115–147, 1987.

[167] A. Barla, F. Odone, and A. Verri, “Histogram intersection kernel for image classification,” in *International Conference on Image Processing ICIP 2003*, 2003, pp. 513–516.

[168] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[169] L. Tan, D. Taniar, and K. A. Smith, “A clustering algorithm based on an estimated distribution model,” *International Journal of Business Intelligence and Data Mining*, vol. 1, no. 2, p. 229, Dec. 2005.

[170] K.-P. Yee, D. Fisher, R. Dhamija, and M. Hearst, “Animated Exploration of Dynamic Graphs with Radial Layout,” in *IEEE Symposium on Information Visualization.* IEEE Computer Society, Oct. 2001, p. 43.

[171] I. Tollis, G. Di Battista, P. Eades, and R. Tamassia, *Graph Drawing: Algorithms for the Visualization of Graphs.* Upper Saddle River, NJ: {Prentice Hall}, 1998.