Abstract

This thesis deals with the problem of multimodal dimensionality reduction (DR), which arises when the input objects, to be mapped on a low-dimensional space, consist of multiple vectorial representations, instead of a single one. Herein, the problem is addressed in two alternative manners. One is based on the traditional notion of modality fusion, but using a novel approach to determine the fusion weights. In order to optimally fuse the modalities, the known graph embedding DR framework is extended to multiple modalities by considering a weighted sum of the involved affinity matrices. The weights of the sum are automatically calculated by minimizing an introduced notion of inconsistency of the resulting multimodal affinity matrix. The other manner for dealing with the problem is an approach to consider all modalities simultaneously, without fusing them, which has the advantage of minimal information loss due to fusion. In order to avoid fusion, the problem is viewed as a multi-objective optimization problem. The multiple objective functions are defined based on graph representations of the data, so that their individual minimization leads to dimensionality reduction for each modality separately. The aim is to combine the multiple modalities without the need to assign importance weights to them, or at least postpone such an assignment as a last step.

The proposed approaches were experimentally tested in mapping multimedia data on low-dimensional spaces for purposes of visualization, classification and clustering. The no-fusion approach, namely Multi-objective DR, was able to discover mappings revealing the structure of all modalities simultaneously, which cannot be discovered by weight-based fusion methods. However, it results in a set of optimal trade-offs, from which one needs to be selected, which is not trivial. The optimal-fusion approach, namely Multimodal Graph Embedding DR, is able to easily extend unimodal DR methods to multiple modalities, but depends on the limitations of the unimodal DR method used. Both the no-fusion and the optimal-fusion approaches were compared to state-of-the-art multimodal dimensionality reduction methods and the comparison showed performance improvement in visualization, classification and clustering tasks. The proposed approaches were also evaluated for different types of problems and data, in two diverse application fields, a visual-accessibility-enhanced search engine and a visualization tool for mobile network security data. The results verified their applicability in different domains and suggested promising directions for future advancements.
The author confirms that the work is his own and that all other work is appropriately referenced.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Acknowledgements

First, I would like to thank my principal supervisor, Prof. Erol Gelenbe, for the opportunity for this research, as well as for his guidance and advice in completing the thesis. I would also like to especially thank my collaborative supervisor at the Information Technologies Institute (ITI), Dr. Dimitrios Tzovaras, without whom this thesis would not be possible, for his supervision and support throughout the whole period of my research. Furthermore, I wish to thank Dr. Anastasios Drosou (ITI) for his invaluable assistance in every aspect, with his knowledge and experience.

Finally, I would like to thank my family, my friends and my colleagues in ITI, for always having been on my side throughout all the difficulties of the past four years.

\footnote{A major part of this thesis was supported by the 7th Framework Programme EU project NEMESYS (FP7-ICT-317888), coordinated by Prof. Erol Gelenbe. The opinions expressed in this thesis are those of the author and do not necessarily reflect the views of the European Commission.}
Contents

1 Introduction ........................................................................... 10
   1.1 Problem formulation .................................................. 15
   1.2 Motivation and contribution ....................................... 19
       1.2.1 Motivation and contribution for multi-objective dimen-
            sionality reduction ............................................. 19
       1.2.2 Motivation and contribution for Multimodal Graph
            Embedding ...................................................... 22
   1.3 List of publications ..................................................... 25
       1.3.1 Journals ......................................................... 25
       1.3.2 Conferences .................................................. 25
   1.4 Thesis outline ............................................................ 26

2 Literature survey .................................................................. 28
   2.1 Unimodal dimensionality reduction ............................... 28
       2.1.1 Global methods .............................................. 28
       2.1.2 Local methods .............................................. 32
   2.2 Multimodal learning .................................................... 39
       2.2.1 Multimodal fusion using a common representation .... 45
       2.2.2 Multimodal fusion using simultaneous learning ......... 55
       2.2.3 Multimodal learning without automatic fusion ........... 57
   2.3 Summary and discussion .............................................. 61

3 Multimodal dimensionality reduction using multi-objective
   optimization .................................................................. 66
   3.1 The multi-objective approach to dimensionality reduction . 67
   3.2 Formulation using graphs ............................................ 72
   3.3 Objective function examples ....................................... 73
   3.4 Summary ............................................................... 76

4 Multimodal dimensionality reduction using Multimodal Graph
   Embedding ................................................................ 79
   4.1 Affinity matrix fusion ................................................. 80
   4.2 Adaptive determination of modality weights ................... 83
   4.3 Summary ............................................................... 91
5 Experimental evaluation

5.1 Datasets
5.1.1 Artificial multimodal dataset
5.1.2 Image dataset
5.1.3 Image-sound dataset
5.1.4 Image-text-video dataset
5.1.5 Image-text dataset

5.2 Multimedia descriptors
5.2.1 Image descriptors
5.2.2 Sound descriptors
5.2.3 Text descriptors
5.2.4 Video descriptors

5.3 Quantitative evaluation measures
5.3.1 Dunn index for visualization
5.3.2 Average Isoperimetric Quotient for visualization
5.3.3 Mean Recognition Rate for object classification
5.3.4 Rand index for clustering

5.4 Experimental evaluation of multi-objective dimensionality reduction
5.4.1 Experiments with artificial data
5.4.2 Comparison of different objective functions
5.4.3 Comparison with existing methods

5.5 Experimental evaluation of Multimodal Graph Embedding
5.5.1 Evaluation in supervised object classification
5.5.2 Evaluation in unsupervised object clustering
5.5.3 Evaluation in unsupervised object visualization

5.6 Summary

6 Applications of multi-objective dimensionality reduction and multimodal graph embedding

6.1 Application of multi-objective dimensionality reduction in accessibility-enhanced search engines
6.1.1 Accessibility feature extraction
6.1.2 Objective functions used in multi-objective dimensionality reduction
6.1.3 Ranking selection using the impairment profile
6.1.4 Experimental evaluation ........................................ 146
6.2 Application of multi-objective and multimodal graph embed-
ing dimensionality reduction in mobile network security . . 152
6.2.1 Mobile network datasets used ................................. 154
6.2.2 Extracted behavioural descriptors ............................ 156
6.2.3 Experimental results for the SMS flood attack scenario 158
6.2.4 Experimental results for the RRC attack scenario . . 163
6.3 Summary .............................................................. 166

7 Conclusions ............................................................ 169
7.1 Summary of the thesis .............................................. 169
7.2 Critical discussion .................................................. 172
7.3 Future work .......................................................... 174

List of Figures

1 Levels of multimodal fusion. ..................................... 41
2 Organization of multimodal learning approaches. ......... 46
3 Example Pareto diagram for two objectives. .................. 70
4 Illustration of solutions discovered by a weighted-sum method, using different weight combinations. ............... 71
5 Example of the potential objective minimization. .......... 77
6 Illustration of multiple combinations of modality weights. . 83
7 Illustration of the notion of graph inconsistency as the existence of triangles ........................................ 84
8 Illustration of the notion of graph inconsistency as triangles of equally weighted edges .............................. 85
9 Example of the result of modality weights optimization. . 89
10 The points of the Pareto front, representing solutions of the multi-objective optimization problem. .............. 104
11 Two-dimensional mappings corresponding to point A, B and C in Fig. 10. .................................................. 105
12 Comparison of two-dimensional mappings of the image dataset, using the minimum angle and potential objective functions. 108
13 Comparison of the Dunn index for the multi-objective method, using the minimum angle and the potential objectives. 109
Comparison of MKL-LPP, MST-FD and the multi-objective method, using the image dataset. .................................................. 110
Comparison of the Dunn index and the AIQ for MKL-LPP, the MST-FD and the multi-objective method, for the image dataset. ................................................................. 114
Comparison of MKL-LPP, MST-FD and the multi-objective method for the image-sound dataset. .................................................. 115
Comparison of the Dunn index and AIQ for MKL-LPP, the MST-FD and the multi-objective method, for the image-sound dataset. ................................................................. 116
Comparison of the Dunn index and the Average Isoperimetric Quotient of the multi-objective, the MKL-LPP and the MST-FD method, for the image-text-video dataset. ............... 117
Comparison of the Dunn index and the Average Isoperimetric Quotient of the multi-objective, the MKL-LPP and the MST-FD methods, for the image-text dataset. ......................... 118
Comparison of the Mean Recognition Rate (MRR) for the separate modalities, the MKL and the MGE methods. ............. 123
Comparison of the Rand Index (RI) for the separate modalities, the MKL and the MGE methods. ......................... 127
The affinity matrices for the unsupervised clustering task, for the image and the image-text-video datasets, using the MKL and the MGE methods. ................................................. 129
Comparison of the Dunn Index (DI) for the MKL, the multi-objective and the MGE methods, for the image dataset. .... 131
Comparison of the Dunn Index (DI) for the MKL, the multi-objective and the MGE methods, for the image-text-video dataset. ................................................................. 132
Example visualizations of the image and the image-text-video datasets using the MKL, the multi-objective and the MGE methods. ................................................................. 133
Overview of accessibility score extraction for vision impairment \( y_m \). ................................................................. 138
Application of the filters for the supported impairments on an example image. ................................................................. 141
Example application of the multi-objective accessibility-based reranking. .............................................. 148
Illustration of the relevance-based, automatic and ground truth rankings, presented in Table 12. ......................... 152
Examples of the behavioural descriptors used in the network security application. ........................................ 159
Two-dimensional mappings of the SMS flood attack dataset using the multi-objective dimensionality reduction method, for the 7th day of the simulation (day of the attack). ........ 160
Multi-objective mappings for each of the 7 days of the period covered by the SMS flood attack dataset. ............. 161
Two-dimensional mappings of multiple modality trade-offs of the SMS flood attack dataset for the 7th day of the simulation. 162
Two-dimensional mapping of the 7th day of the SMS flood attack dataset, using the MGE method. .................. 163
Mappings produced by the MGE method for each of the 7 days of the SMS flood attack dataset. ....................... 164
Mappings of the RRC attack dataset using the multi-objective dimensionality reduction method. ................... 165
Mappings of multiple Pareto-optimal solutions for the RRC attack dataset using the multi-objective dimensionality reduction method. ............................................................. 166
Two-dimensional mappings of the RRC attack dataset, using the MGE method. ......................................... 167

List of Tables

1 Various formulations of graph embedding. .................. 37
2 Various formulations of multimodal graph embedding. .... 81
3 Optimal modality weights for the construction of the multimodal matrix of the example of Fig. 9. ................. 91
4 Notions of cluster distance and diameter used for the Dunn index. ......................................................... 100
5 Values of the evaluation measures for the two-dimensional mappings of Fig. 14. ..................................... 113
Values of the evaluation measures for the two-dimensional mappings of Fig. 16 .................................. 115
Parameters used in the three experiments for the various datasets, in the task of supervised object recognition. ............ 122
Modality weights for the supervised classification task, calculated by MGE, for the two datasets. ...................... 124
Parameters used in the two experiments for the various datasets, in the task of unsupervised object clustering. .......... 126
Modality weights for the unsupervised clustering task, calculated by MGE, for the two datasets. ...................... 128
Parameters used in the two experiments for the various datasets, in the task of unsupervised object visualization. ........ 130
Example of the relevance-based and multi-objective rankings for a user session. ............................................ 151
User groups of the SMS flood attack dataset. ............... 155
User groups of the RRC attack dataset. ...................... 156
Record attribute types. ........................................... 156
Behavioural descriptors used in the network security application. ................................................................. 158
1 Introduction

Dimensionality reduction is a significant part of many machine learning tasks, such as object recognition, classification, regression, clustering, search and retrieval, k-nearest neighbor search and visualization. Dimensionality reduction is the process of using as input a set of points that lie in a space of many dimensions, e.g. larger than 50 or 100, and mapping them on a space of few dimensions, e.g. less than 50, or just 2 or 3, so that the intrinsic properties of the points, such as their pair-wise distances or their participation in various classes, are not altered, or are altered in a small amount.

Although dimensionality reduction has largely been used in the literature, recent advances in machine learning, such as deep Convolutional Neural Networks (CNN) [1], or decision forests [2], successfully solve learning problems, such as classification, without the need for dimensionality reduction as preprocessing. Usually, these approaches, e.g. deep learning methods, internally perform some sort of dimensionality reduction, by learning features that are representative of the input data. However, viewing the objective of dimensionality reduction by itself, i.e. providing a compact representation of the high-dimensional input data, is also greatly useful. Having a low-dimensional representation of a set of data, separated from any further processing that they may undergo, means that only the most important information is kept, that a better intuition in the data characteristics that matter is gained, that the data can be used with any machine learning task that would perform poorly for high-dimensional data, that the data can be visualized more easily, etc.

The usefulness of dimensionality reduction lies, at least, at the following considerations:

- First, not all dimensions may be needed to describe a set of high-dimensional data, at least in the context of a specific application. It is usual that, although the available data are high-dimensional, their intrinsic dimensionality is small. The intrinsic dimensionality is the actual number of dimensions that are needed in order to fully describe the characteristics, variation, similarities and differences of the available data. As an example, let a set of images of size 200x200 be considered, each showing the same grayscale photo of Nicola Tesla, but in various rotations and scales. Although each image could naively be
represented as a vector of 40,000 dimensions, which is the total number of its pixels, the actual values that are needed in order to distinguish one image from the other are only two: the angle of the rotation and the amount of scaling. In formal terms, the data may lie on a low-dimensional manifold inside the high-dimensional space, so that they do not occupy the whole space, but rather a dimensionally limited manifold inside it. Even if the data do occupy the whole space, many dimensions may be rather unimportant, not providing much information about the variance of the data, so that approximating the data as lying on a manifold may be reasonable.

- Second, high dimensionality may prohibit or deteriorate the performance of specific machine learning tasks, such as Nearest-Neighbour (NN) search for classification, or distance-based clustering. Considering a set of points lying in a high-dimensional space, it has been proved [3] that when dimensionality increases, the distance of a sample point to the nearest point becomes very similar to the distance of the sample point to the most distant point. This phenomenon is known as the curse of dimensionality. Since calculation of distances between points is inherent in many machine learning tasks, such as ones involving computation of nearest neighbors or distance graphs, the curse of dimensionality can affect their performance and result in unstable outcomes. In this respect, dimensionality reduction techniques can be an invaluable data pre-processing step, before the application of machine learning algorithms. Of course, many dimensionality reduction methods depend on the calculation of distances themselves, so they should be subject to the curse of dimensionality as well. However, the information of whether a certain distance is larger than another is preserved, even in a high-dimensional space, and dimensionality reduction methods often rely on this information, rather than on the actual values of the distances, thus avoiding the curse.

- Third, an algorithm that uses low-dimensional data as its input usually requires less time and memory resources to process the data, than if the data were high-dimensional. Thus, reducing the dimensionality of a set of data prior to the application of a computationally heavy algorithm, when applicable, can significantly improve the performance. This can
be especially useful if a machine learning task needs to be executed on a mobile platform, such as a mobile phone, where the processing power and the memory resources for short- or long-term storage are often limited.

Dimensionality reduction typically starts by considering the high-dimensional points, and performing a set of actions on them, until a low-dimensional representation is reached. These actions usually involve the calculation of distances among the points. A similar problem is how to generate coordinates for a set of abstract objects for which only distances or similarities are known among them. This problem is not dimensionality reduction in the strict sense, since the dimensionality of the initial space is not known, but it essentially follows the same principles. In this kind of problem, the goal is again that the output space is of low dimensions, in order to have the advantages of time and space efficiency, as well as the avoidance of the curse of dimensionality.

Apart from being an intrinsic processing step for machine learning applications, dimensionality reduction has significant applications in data visualization. The Gestalt laws of visual perception [4] [5] provide guidelines about which visual characteristics assist the viewer to perceive the data most easily. Specifically, the laws of proximity and similarity state that humans perceive a set of objects as similar if they are presented close to each other, or if they are similar in some visual characteristic other than position, e.g. color. However, the visual characteristic that is most easily perceived by humans is position [6], so many visualization methods use relative positioning to map the data features on the screen. In addition, the law of simplicity explains that objects are considered as similar if they form a simple and regular pattern. The Gestalt laws are thus closely related to dimensionality reduction and clustering on the visual plane, by placing similar objects close to each other, while placing dissimilar objects away from each other, forming uniform clusters that can be easily perceived by humans.

Very commonly, the original high-dimensional points provided as input to the dimensionality reduction algorithms are vectorial descriptors extracted from complex objects, such as images, sounds or videos. The descriptors can be used to map the objects as points on a high-dimensional geometric space, in which methods such as classification and clustering can be performed. Dimensionality reduction can then be applied to the original points, to take
advantage of the desired properties of a low-dimensional space, as described above.

However, extracting descriptors from objects such as images, always involves information loss, since they are extracted by focusing on a specific characteristic, e.g. the texture of an image, while ignoring other information. This is the so-called semantic gap between the high-level semantics of an image, such as the subject or the mood it depicts, and the low-level visual characteristics of it, such as the color, the texture, or the shapes in it. This information loss and ambiguity with respect to the reflection of the semantics on the low-level characteristics is later passed to the output of dimensionality reduction methods which use these incomplete original descriptors. The result is that the final mapping of the objects on a low-dimensional space may not follow the semantics of the objects, i.e. the similarities or distances between points in the final space may not correspond to the semantic similarities or distances between the corresponding real objects.

However, information that is lost in one type of descriptor may exist in another type of descriptor. For instance, while color information may not be adequate to separate the image of a leaf from the image of a forest, this separation may be apparent if texture information is considered. Combining multiple sources of information, or multiple channels of information, so-called modalities, has proven to be effective at narrowing the gap between low-level features and high-level semantics [7]. This combination is not a mere concatenation of the features, since some features may be more important than others. The problem of multimodal fusion is a very challenging one and has applications in numerous research areas.

Multimodal fusion is the combined utilization of information originating from multiple sources, or received by different sensors, in order to perceive a particular concept or reach a conclusion. This is the way that the human brain processes information. Information is received simultaneously by all human senses. For each sense, information is of a different form, so called modality, i.e. as a visual image, a sound, a taste etc. However, the processing of this information is not performed independently in each modality. The famous work of McGurk [8], describing the McGurk effect, shows that information of one modality can directly affect the way information of another modality is perceived. If an adult is presented with a video of a person repeating the syllable “ba”, dubbed with the sound of the syllable “ga”, the
adult usually perceives the syllable as “da”, since the brain merges the visual and the auditory channels. If the same adult only listens to the audio channel, without the video, the correct syllable “ga” is perceived. A notable example of this effect is the fact that, in order to recognize speech, humans do not only rely on the sound modality, but also on the visual modality, as they unconsciously read the lips of the speaker in order to clarify ambiguous sounds, which is more apparent when speaking in a noisy environment [9].

Multimodal fusion has been exploited since long ago in the area of the visual arts. In cinema, the use of music simultaneously with the video and the voices of the actors, plays an important role in evoking specific emotions to the audience, or even supporting the plot. A scene of a woman entering a room can be funny if it is accompanied by vivid and happy music, while the same scene can be intense or scary if it is accompanied by high-pitched and intense music. Fusion can occur even within the same sensory channel, if the perceived information is in different forms. In comics, which is a visual-only medium, the drawn pictures are perceived together with the words in the balloons. In [10], seven types of combinations of pictures and words are distinguished. Among them, there are word-specific or picture-specific combinations, where either words or pictures are the most important in a panel; there are duo-specific or parallel combinations, where words and pictures send the same information or very different information; and there is an inter-dependent type, used quite commonly, where the combination of words and pictures conveys an idea that neither could convey on its own.

The observation that the human mind processes information in a multimodal manner has motivated research in machine intelligence. The development of techniques and algorithms that combine multiple sources of information has boosted the performance of machine learning and computer perception approaches, making them approach the way humans perceive information and be closer to the true semantics of the raw information. Inspired by the way humans recognize speech, audio-visual speech recognition systems are being developed [11] [12]. Combining both audio and visual information, these systems can recognize speech more accurately, with larger tolerance to noise, in both the audio and the visual channel. Inspired by the way humans recognize other humans, even when not seeing one’s face, multimodal biometric identification systems are being developed [13] [14]. The combination of multiple modalities, such as the person’s appearance,
voice, gait, iris images, etc., makes identification more robust, more accurate and less obtrusive to the person being identified. Inspired by the way humans communicate with other humans, not only using words, but also the loudness of one’s voice, visual features of one’s face or gestures, novel human-computer interaction approaches are being designed [15] [16]. Combining multiple communication channels, these systems provide a more accurate and natural interaction between human and machine, that is closer to how people communicate with each other. The multimodal nature of the human brain has inspired numerous approaches in the literature, in several other fields, such as multi-sensor fusion, multimedia indexing and retrieval, etc.

In the context of dimensionality reduction, multimodal fusion can be used to merge multiple high-dimensional features extracted from the objects at hand, in order to map them on a low-dimensional space in which there is a better correspondence between distances on the space and semantic similarity than using each feature separately. The calculation of a final low-dimensional space from multimodal data, apart from reducing the number of dimensions, also has the advantage of constructing a unified space, on which all data modalities are mapped and the semantic relationships of the data are more apparent. Such a unified space allows for common unimodal procedures to be performed using multimodal data, for instance calculating distances among multimodal objects, finding nearest neighbors, or visualizing the multimodal objects. The subject of this thesis is to examine ways in which multiple features extracted from the data can be combined for the purposes of dimensionality reduction, with an ultimate goal to perform machine learning tasks on the final space, such as classification, clustering and visualization.

1.1 Problem formulation

The major challenge in dimensionality reduction is how to reduce the data dimensions in such a way that the least amount of information is lost. Early dimensionality reduction techniques that are also commonly used today, such as PCA [17], considered linear transformations of the input space. Although simple to implement, linear methods usually involve a large amount of information loss, since many problems are non-linear in nature. Utilizing the kernel trick [18], linear methods can be transformed to non-linear, thus cov-
ering much more complex cases of data and being more accurate. Using the kernel trick essentially takes the original data and transforms them into a very high-dimensional space, even infinite-dimensional, where certain properties of the data may be apparent, before transforming it back to few dimensions. The intermediate very high-dimensional data are actually never calculated, but instead only a pairwise kernel similarity is known for them, using a specific kernel function.

However, even in this case, the results may not be accurate. In linear and kernel-based methods, information from all the points in the dataset is used to construct a global model for the low-dimensional representation of each point. In cases of largely twisted low-dimensional manifolds embedded in the high-dimensional space, such a global approach may not be accurate, as the geometry of the dataset as a whole may not be characteristic of the geometry at a finer scale. In such cases, finding the local structure of the manifold in a small neighborhood around each point, using the so-called manifold learning methods, such as incremental manifold learning [19] and Riemannian manifold learning [20], is more appropriate in order to describe the whole manifold. Although such an approach is not as efficient as the global methods in terms of speed, it is more efficient in terms of accuracy, and is also capable of handling both simple and complex manifolds. Distances among the original high-dimensional data are calculated and are used to construct $k$-nearest neighbor graphs out of them. Then these graphs are decomposed and used to calculate the low-dimensional points, so that the structure of the neighborhood graph is best maintained. The neighborhood graph constructed from the original data follows the manifold spanned by the data, so that in an essence, the low-dimensional representation is produced by unfolding this manifold onto a low-dimensional space. However, this solution applies only to data that indeed have a manifold structure, and the notion of distance used, the number of neighbors etc, are still challenges.

Another challenging issue with respect to dimensionality reduction is the efficiency of the adopted approach in handling new data, coming from outside the database which is used to calculate the low-dimensional representation. Manifold learning methods, although accurate in several cases, are more difficult in handling new data, since their distances from all other points have to be calculated first. On the other hand, linear methods, although relatively inaccurate, result in a linear transformation of the original dimensions, so
that it is very easy to map new data on the low-dimensional space.

Regarding multimodality, the major challenge is how to properly fuse the multiple available features describing the multimodal objects, so that the resulting low-dimensional representations are close to the semantic relationships among the data.

The first challenge mentioned above, i.e. fine-tuning parameters such as the distance measure used or the number of nearest neighbors, are open issues which generally depend on the dimensionality reduction method or the dataset used. The second challenge, i.e. how to handle data outside the training database, is generally handled differently according to the dimensionality reduction method used. In general, methods that rely on the calculation of nearest neighbors around the points handle data outside the database harder than methods focusing on a global transformation of the input space, since the distances of the new coming data point to all other data points need often to be calculated. The approaches proposed in this thesis are based on nearest neighbours, often relying on existing unimodal techniques, so they share the same difficulty in handling new coming data as the unimodal ones. Although these first two challenges are very interesting and research is being conducted in these directions, the focus of the current thesis is on the third challenge, i.e. how to properly combine the multiple available modalities in order to produce the low-dimensional representations of the input data.

Formally, the problem of multimodal dimensionality reduction can be stated as follows. A set $\mathcal{O}$ of multimodal objects is considered:

$$\mathcal{O} = \{O_1, O_2, \ldots, O_N\},$$

where $N > 0$ is the number of multimodal objects in the dataset. A multimodal object (MO) $O_i$ is a set of $M$ feature vectors:

$$O_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,M}\}, \quad o_{i,m} \in \mathbb{R}^{D_m},$$

$$i = 1 \ldots N, \quad m = 1 \ldots M.$$

where $M > 0$ is the number of modalities and $D_m$ is the dimensionality of the feature vectors of modality $m$.

A modality is hereby considered as a specific type of representation for an object. For example, in a multimedia database, a pair of an image and a sound describing the same concept is considered as a MO having an image
modality and a sound modality. The same holds for e.g. an image described by multiple low-level properties, such as the color and the texture. The $M$ feature vectors of a MO correspond to the multiple representations available for this MO. Each feature vector $\mathbf{o}_{i,m}$ is a descriptor extracted from the object and is of dimensionality $D_m$, which is generally different for each modality.

Hereby, each multimodal object is considered to consist of a fixed number of $M$ feature vectors, each of a different modality. In a more general case, a multimodal object can consist of a different number of feature vectors, where some modalities may be missing or others be present multiple times, i.e. as multiple feature vectors of the same modality. For instance, if a multimodal object is considered as a bag of multimedia describing a specific semantic concept, it could consist of a video, a sound and three images, while another object of the same dataset could consist of a video, an image and no sound items. The problem hereby is restricted to cases where the number of modalities in a multimodal object is fixed and each feature vector corresponds to another modality. This restriction is imposed due to the nature of the approaches proposed in this thesis, which are based on the definition of unimodal distances between the multimodal objects. A unimodal distance between two objects considers only the distance between their feature vectors of a specific modality. The restriction of having one feature vector per modality simplifies the definition of the unimodal distance and puts the focus on the combination of the distances, which is the purpose of this thesis. However, if different notions of unimodal distances are used, which handle more than one feature vectors per modality or missing feature vectors for some modality, the proposed approaches can readily be used to handle such cases as well.

The goal of multimodal dimensionality reduction is to represent each multimodal object $O_i$ as a point $\mathbf{y}_i \in \mathbb{R}^d$ in a low-dimensional space of dimensionality $d << D_m, \forall m \in \{1 \ldots M\}$. The output of multimodal dimensionality reduction can thus be formulated as a matrix $\mathbf{Y} \in \mathbb{R}^{N \times d}$, where the $i$-th row represents the low-dimensional representation of multimodal object $O_i$. Moreover, the mapping $f : \mathcal{O} \rightarrow \mathbb{R}^d$ is often desired, in order to map points coming from outside the initial dataset onto the low-dimensional space. The challenge, which is being addressed by existing methods, is how to efficiently combine the multiple modalities. Another challenge, which has
not been given much focus, is how to provide such effective combination mechanisms that are also generic, so that they can easily extend unimodal dimensionality reduction methods into multimodal ones. This thesis is targeting both these objectives.

1.2 Motivation and contribution

This thesis proposes two methods for multimodal dimensionality reduction:

- the multi-objective dimensionality reduction method, and
- the Multimodal Graph Embedding dimensionality reduction method.

Both methods target the goals of efficient multimodal dimensionality reduction, in terms of accurately performing machine learning tasks on the low-dimensional space, as well as allowing existing unimodal dimensionality reduction methods to be extended in multiple modalities easily. This section outlines the motivation behind the two proposed methods and the contribution of the thesis to the state-of-the-art, with respect of each of them.

1.2.1 Motivation and contribution for multi-objective dimensionality reduction

Motivation  Many instances of the unimodal dimensionality reduction problem are handled as optimization problems. Starting with the high-dimensional points, an objective function is defined in such a way that its minimization or maximization results in a low-dimensional representation of the points that has specific desired properties which ensure that it is a proper representation for the original points. This objective function is often defined using the pair-wise distances between the original points. If multiple feature types can be extracted from the original data, multiple unimodal dimensionality reduction objective functions can be defined, resulting in multiple optimization problems. The ideal solution to these problems would be one which simultaneously minimizes/maximizes them all. However, since the optimization problems for the different modalities are generally conflicting, this solution does not exist.

This multi-objective view of the problem of multimodal dimensionality reduction is generic and includes many of the existing techniques as subcases. In order to solve the multi-objective optimization problem, many
methods follow the *scalarization* technique. They merge the multiple objectives into one single objective and then solve a single optimization problem. The merging of the objectives involves certain preferences with regard to one objective or another, or setting specific importance weights to the objectives. However, how to set these preferences is not straightforward. It usually means that some information is lost, since one modality is preferred over another. In the same way as a sum does not preserve all information of the parts that are summed, fusion cannot preserve all information of the modalities that are being fused. Even if the trade-off is selected accurately, according to the data, the fact that the result is a single solution means that some information from the original separate modalities is lost. Nevertheless, although in many cases the fused result suffices for the solution of a specific problem, the information that is lost could be invaluable for the understanding of a specific dataset and its internal structure, since viewing a dataset from different views can provide more insight in the data.

An example of an approach that combines multiple features of the available data, without fusing them and without concealing any information is provided by the field of data visualization: using *glyphs*. A glyph is a visual representation of a multimodal record, where multiple different attributes of a data object are mapped on different visual characteristics of the glyph. Such visual characteristics include the position, the color, the size and the shape of the glyph. In such a representation, all the available information is simultaneously presented to the viewer. Instead of somehow fusing the information before it is presented to the viewer, it is up to the viewer to perceptually fuse the available information. Mentally, the viewer can focus on one visual cue or another, or combine them, and switch from a specific view of the data to another. Through this freedom, the viewer can gain much more insight in the data. This is related to the above problem of having multiple objective functions to be simultaneously minimized, while not combining them in a way that information is lost.

Nevertheless, as the alternative of scalarization techniques, problems of multiple conflicting objectives have been addressed in the literature by *multi-objective optimization* methods [21]. Multi-objective optimization has been effectively applied in fields such as economics and engineering [22], for solving problems of conflicting objectives, e.g. maximizing the profit while minimizing the risk of an investment. Instead of computing a single solution to the
problem, based on a combination of the multiple modalities, these methods compute a set of optimal solutions. This set is called the Pareto set and its solutions correspond to various trade-offs among the modalities, from which one can be selected by a user or an automated process. Since the problem of multimodal dimensionality reduction involves simultaneously minimizing conflicting objectives formed from different modalities, the use of multi-objective optimization techniques for multimodal dimensionality reduction seems promising.

Apart from computing the Pareto set, there are various scalarization techniques attempting to solve the multi-objective optimization problem by computing a single solution. As described later, in Section 2.2.3, these techniques, such as the weighted sum, or the $\epsilon$-constraint method, correspond to existing multimodal fusion methods. It is interesting to investigate how multimodal dimensionality reduction can be handled by the class of multi-objective optimization methods that, instead, calculate the whole set of Pareto-optimal solutions.

**Contribution** The contribution of the multi-objective method to the state-of-the-art is twofold:

- A novel perspective to the problem of multimodal dimensionality reduction is proposed, where the problem is viewed as a multi-objective optimization problem. In contrast to existing methods, this new perspective results in a number of Pareto-optimal solutions/mappings for a user to select, including ones which cannot be discovered by existing weighted sum-based methods, even if all weight combinations are considered. The weighted sum-based approach may compute solutions where one of the objectives has a large value, so long as the weighted sum of the objectives is minimum. In this way, other solutions which may have small values for all objectives, but they do not minimize any weighted sum, would be ignored, although they may constitute useful trade-offs. In this respect, the contribution of the proposed multi-objective dimensionality reduction approach is that it considers techniques designed for such generic multi-objective optimization problems, instead of relying on weighted sum-based methods, in order to discover such solutions.

- A novel graph-based multimodal dimensionality reduction method is
proposed, which utilizes the multi-objective framework, using graph-based objective functions, and outperforms existing techniques.

The multi-objective dimensionality reduction method constitutes Chapter 3 of this thesis.

1.2.2 Motivation and contribution for Multimodal Graph Embedding

Motivation A large number of dimensionality reduction techniques consider that the high-dimensional data lie on a low-dimensional manifold and attempt to unravel this manifold. As proposed by the authors of [23], many manifold learning dimensionality reduction methods can be considered as special cases of a general graph embedding framework. The graph embedding framework relies on the use of a neighborhood graph, and its associated affinity matrix, which acts as a “target” towards which the embedding is performed. Although random forests [24] have also been used in the literature as an alternative to common nearest neighbors-based techniques, they have mostly been used for supervised classification purposes, with a few exceptions [25]. In this thesis, neighborhood graphs constructed using measures of distance between data points will be considered, an approach that is widely adopted in the related research area.

The neighborhood-based approach to dimensionality reduction is closely related to label propagation [26], which is a common technique for semi-supervised learning. In label propagation, labels are known for a small subset of the high-dimensional data. These labels, acting as seeds, are propagated to the rest of the data, in the same way that heat originating from a set of sources is propagated to fill the whole space. The propagation is guided by a neighborhood graph constructed from the data, so that a data point essentially only affects its nearest neighbors.

In order for the low-dimensional mapping to describe the semantic relationships of the original data more accurately, for approaches such as dimensionality reduction and label propagation, it is expected that areas of the data space in which the data are similar to each other are represented by areas of the graph which contain many edges among the similar points. This desired property of the graph is hereby referred to as graph consistency. However, since the descriptors extracted from the data may not be accurate
enough to encode the semantic relations among the data, such graphs constructed from unlabeled data may not be sufficiently consistent, leading to inaccurate embeddings.

This semantic gap can be narrowed if more than one modalities are considered. Hereby, the assumption is made that the descriptors of each modality have an amount of discriminating power for describing a certain view of the data relationships. In other words, it is assumed that the individual modalities are not completely uncorrelated with each other and that each is capable to convey at least some information about the semantic similarities among the data. Assuming the opposite, i.e. that there are modalities that do not convey any information about the semantic relationships among the data would render them useless to begin with. Following this assumption, fusing the affinity matrices of multiple graphs, e.g. by summation, is expected to contribute to enriching those areas of the final graph containing semantically similar data with more edges. Although edges would also be added in other, inappropriate, areas of the graph, those redundant edges would normally be fewer. Thus, the final graph would be more consistent than each of the unimodal graphs, with a better balance between valid edges, which are increased, and redundant ones, which are decreased. However, the graph embedding framework of [23] does not handle cases of multiple modalities. Recent attempts to include multiple modalities in the framework, such as [27] which addresses the related problem of spectral clustering, calculate the optimal combination of the modalities jointly with the optimal parameters of the low-dimensional embedding, which makes them difficult to be applied to existing unimodal dimensionality reduction techniques.

Specifically for dimensionality reduction, the most relevant technique is the work of [28], in which the Multiple Kernel Learning Dimensionality Reduction (MKL-DR) framework is proposed. MKL-DR has been the basis of several recent multimodal learning techniques, such as [29], which makes use of training data to learn the weights of the modalities, or [30], which assesses the discriminative power of each modality in order to compute the weights. MKL-DR is indeed a multimodal extension of the Graph Embedding framework. It is based on the kernelization formulation of graph embedding and considers, instead of one, multiple kernel matrices among the multimodal objects, one per modality. Multimodal fusion in MKL-DR is performed by fusing the multiple kernel matrices using a weighted sum. However, graph
embedding also requires a target affinity matrix, which guides the dimensionality reduction process. If this target matrix is not based on ground truth knowledge, as in supervised methods, but needs to be extracted from the data as well, then the same fusion problem arises again, as a different affinity matrix is extracted from each modality. These multiple affinity matrices need to be combined before the dimensionality reduction can be performed. In [28], the focus is on the fusion of the kernel matrices, while this affinity matrix fusion is resolved by merely considering the average of the multiple unimodal affinity matrices as the final multimodal affinity matrix. However, this naive approach may not lead to accurate results. One unimodal graph may be more consistent than another, which means that it should be considered as more important for the determination of the final graph. Merely using an average of the unimodal graphs may not stress enough the affinities of an important graph, while it may over-stress the affinities of another, relatively inaccurate, graph. The optimal trade-off among the unimodal graphs needs to be calculated, which leads to the most consistent final graph.

A general method that extends existing unimodal dimensionality reduction approaches by optimally combining multiple graphs constructed from the multiple data modalities would be expected to have even superior performance. However such a method is missing from the existing literature. There are proposed methods, such as [31] and [25], to construct affinity graphs that are more suitable than ones constructed using simple nearest neighbors. However, most of these methods consider a single feature space as input, from which the affinities are to be computed. Methods to combine multiple affinity matrices to improve the characteristics of the combined affinity matrix are missing from the literature. The hereby proposed Multimodal Graph Embedding (MGE) approach is an attempt to fill this gap. Extending the graph embedding framework to multiple modalities means that all the dimensionality reduction methods that are its special cases are extended as well.

**Contribution** The contribution of the proposed Multimodal Graph Embedding method to the state-of-the-art can be summarized in the following:

- A novel measure of graph inconsistency is introduced for the evaluation of the suitability of a neighborhood graph for dimensionality reduction,
measuring how much the edges of triangles formed in the graph differ from each other.

- An extension of the graph embedding dimensionality reduction framework to the multimodal case is proposed, in which a multimodal neighborhood graph is constructed as the weighted sum of the unimodal graphs. The weights of the sum are computed by minimizing the proposed graph inconsistency measure.

The Multimodal Graph Embedding dimensionality reduction method constitutes Chapter 4 of this thesis.

1.3 List of publications

The following publications have been produced during the study for the completion of the current thesis.

1.3.1 Journals


1.3.2 Conferences


2. Kalamaras, I., Drosou, A., Tzovaras, D., “A multi-objective clustering approach for the detection of abnormal behaviors in mobile net-
works”, 2015 IEEE International Conference on Communication Workshop (ICCW), IEEE, 2015.


1.4 Thesis outline

The current thesis consists of 7 chapters. Chapter 1 introduces the reader to the problem of multimodal dimensionality reduction, to its necessity and its applications. After the problem to be addressed is formally stated, the motivation for conducting research in the area of multimodal dimensionality reduction and for addressing it through the proposed multi-objective and multimodal graph embedding techniques is stated.

Next, Chapter 2 constitutes a detailed review of the existing literature that is related to the subject of the thesis. Specifically, literature related to unimodal dimensionality reduction, multimodal fusion in general and techniques of multimodal fusion developed specifically for dimensionality reduction and related tasks, is covered. This literature survey facilitates the identification of those areas that have a potential for further investigation, motivating the work proposed in the current thesis.

Chapter 3 introduces and describes in detail the proposed multi-objective dimensionality reduction method. The method constructs multiple unimodal distance graphs and their minimum spanning trees from the multiple features of the multimodal objects and uses them to define multiple objective functions. Each objective function is defined with the goal that its minimization leads to a placement of the objects that best reveals the structure of the underlying tree. Then, multi-objective optimization techniques are employed in order to handle the existence of multiple objective functions that need to be simultaneously minimized. The multi-objective optimization results in a Pareto front of optimal trade-offs among the modalities, which contains solutions that could not be calculated by conventional weighted-sum based methods.

Chapter 4 describes the second proposed method, namely Multimodal Graph Embedding dimensionality reduction (MGE). This method is a mul-
timodal extension of the existing unimodal graph embedding dimensionality reduction framework, which contains many popular unimodal dimensionality reduction methods as its instances. Since the graph embedding framework relies on the use of appropriately defined affinity matrices, the proposed MGE method uses a multimodal affinity matrix, defined as the weighted sum of the unimodal affinity matrices constructed by the individual modalities. The weights of the sum are automatically learned using the available data, so that the resulting matrix maximizes an introduced notion of consistency of the underlying graph.

The experimental evaluation of the two proposed methods is included in Chapter 5. For both methods, the evaluation was conducted using existing and constructed multimodal multimedia datasets, using image, sound, video and text descriptors extracted from the multimedia as modalities. The proposed methods were compared to state-of-the-art multimodal dimensionality reduction methods and they seemed to outperform the existing methods in tasks such as data visualization, clustering and classification.

Apart from this experimental evaluation, which illustrates the potential of the proposed methods over the state-of-the-art, further applications of them in diverse application fields have been included in Chapter 6, in order to illustrate the generic nature of the methods and their applicability to many forms of data and application types. Specifically, two application areas are investigated, accessibility-enhanced search engines for visually impaired users and data visualization for mobile network security.

The final chapter summarizes the work and results of the current thesis. This chapter provides a critical discussion of the outcomes, stressing the advantages and disadvantages of the proposed methods, and suggests possible future enhancements and directions for research.
2 Literature survey

Multimodal dimensionality reduction is a research area derived from two large families in the literature: dimensionality reduction and multimodal learning. Significant amount of work has been conducted in these two areas, for a wide range of applications. The two areas have also been merged in a number of recent works. This section provides an overview of the literature in these two areas, focusing on the most representative approaches. The section is divided into three parts: first, an overview of unimodal dimensionality reduction methods, second an overview of multimodal learning techniques, not restricted to the dimensionality reduction area, and finally an overview of the methods that have been proposed to tackle specifically the problem of multimodal dimensionality reduction, or similar ones.

2.1 Unimodal dimensionality reduction

Most existing work on dimensionality reduction considers that the data are unimodal, i.e. there exists only one, high-dimensional, data source, from which a mapping to a low-dimensional space is desired. Recently, a number of papers have considered the use of multiple modalities, or sources of information, for the problem of dimensionality reduction, or similar problems. These methods will be discussed below, in Section 2.2.1. This section deals with existing work addressing the problem of unimodal dimensionality reduction, i.e. the transformation of original data of high dimensionality, e.g. in the order of 100 or 1000, to data of much lower dimensionality, so that the geometry of the final data resembles the geometry of the original data as much as possible.

In this section, the methods of unimodal dimensionality reduction have been roughly divided into two categories: global methods, where the relationship of a data point to all other points of a dataset is used, and local methods, where neighbors around a data point are used. Representative methods of the literature for these two categories are presented in the following sub-sections.

2.1.1 Global methods

The first attempts for dimensionality reduction were methods that consider the relationships of a data point to all data points in a dataset, in order
to compute the embedding in the low-dimensional space (global methods). The oldest and most commonly used method is Principal Component Analysis (PCA) [17]. PCA is a linear method that computes an orthonormal set of principal axes describing the directions of maximum variance found in the high-dimensional data. The direction along which the original high-dimensional data have their maximum variance is selected as the first principal axis. Then, the data are projected to a sub-space, of dimensionality one less than the original space, that is perpendicular to the first principal axis. Using this projection as the original data, the next direction of maximum variance is selected as the second principal axis. This procedure is repeated until the zero dimension has been reached, or a sufficient number of dimensions have been computed. The principal axes are the eigenvectors corresponding to the largest eigenvalues of the covariance matrix constructed from the data, so the problem of PCA reduces to the solution of an eigenvalue problem. Projecting the data to the subspace spanned by the principal axes leads to a low-dimensional representation of the data, which preserves the data variance. PCA is a very straightforward and easy to implement technique. It results in a linear mapping from the original high-dimensional data to the low-dimensional space. Since the mapping is linear, it is easy to handle points outside the original dataset. However, being linear is also a disadvantage of the method, since it cannot handle data that lie on non-linear manifolds of the high-dimensional space.

In order to overcome this limitation, kernel PCA has been proposed [32]. This approach utilizes the kernelization technique commonly known as the kernel trick [33], which is widely used to promote linear machine learning techniques, such as Support Vector Machines (SVM), to non-linear ones. In order to apply the kernel trick to PCA, the method is first reformulated so that, instead of using the covariance matrix, a matrix containing the pairwise Euclidean distances between the original points is used. The method is then modified so that the distance matrix is replaced by a kernel matrix, constructed from the data, where each element is computed by applying a kernel function to the corresponding distance. The solution is again obtained by computing the eigenvectors and eigenvalues of the kernel matrix, and the low-dimensional representation is reconstructed from the principal eigenvectors. Commonly used kernel functions are the polynomial and the heat kernels. Kernel PCA can handle several cases of non-linear data, but
has the disadvantage that the size of the kernel matrix is usually large, hence it is computationally inefficient to compute the eigenvectors.

A similar distance-based technique is Multidimensional Scaling (MDS) [34]. MDS is a dimensionality reduction method that tries to preserve the pairwise distances among the original data. MDS takes as its input a distance matrix constructed by considering the pairwise distances among the original high-dimensional points. As an output, it produces new coordinates for these points, on a space of a specified dimensionality, so that the distances between the points are best preserved. Similar to PCA and its variants, and similar to many other dimensionality reduction methods, MDS is formulated as an eigenvalue problem. If the original dimensionality of the data is $D$, then the $D$ largest eigenvalues are non-zero, while the rest are zero. Considering the $d$ largest eigenvalues, where $d < D$, one can approximate the original data, with increasing accuracy as $d$ approaches $D$. As a side note, since the starting point of MDS is the distance matrix constructed from the data, it can also be used in cases where, instead of the original high-dimensional features of the data, only the distances between them are known.

MDS has been used in a variety of applications, and several variations have been proposed. A discussion of the use of MDS for visualization purposes can be found in [35]. MDS has also been used for graph drawing [36], where the goal is to place the nodes in such positions, so that target distances, defined in terms of path lengths, are best preserved. A variation of MDS, namely Local Multidimensional Scaling, has also been proposed [37], where neighborhood relationships among the data are also taken into account.

PCA, MDS and their variants consider only the available data, in order to compute the low-dimensional embedding. However, if supervision information about the classes that the data belong to is available, it can be exploited in order for the low-dimensional embedding to better describe the underlying data classes. Linear Discriminant Analysis (LDA), also known as Fisher Discriminant Analysis (FDA) [38] [39] is a popular technique that follows this principle. The idea behind LDA is similar to PCA, i.e. it computes an optimal set of directions such that the projection of the data on them describes the data in the best way. However, instead of computing these directions so that they constitute the directions of maximum variance in the data, as in PCA, the directions are hereby calculated as those that achieve maximum separation among the classes that the data
belong to. Two scatter matrices are defined, the within-class scatter matrix, which encodes the within-class distances among the points, and the between-class scatter matrix, which encodes the between-class distances among the points. The goal of LDA is to find projection directions that minimize the within-class distances, while maximizing the between-class distances. After the projection axes are calculated, the first few of them are selected, as in PCA, to define the low-dimensional projection space. New, unlabeled, data can then be projected on this low-dimensional space, and then undergo further machine learning processing, such as classification.

Probabilistic formulations of PCA and MDS have also been developed in the literature. In [40], the parameters of PCA are considered as random variables modeled as Gaussian processes, and the whole process of determining them is viewed as fitting the Gaussian process to the data. The authors of [41] propose a modification of this formulation, in order to introduce supervision and reach more accurate results when it comes to more complex and highly non-linear datasets. Formulating dimensionality reduction methods in a probabilistic framework allows for easy extension to non-linear cases, by adjusting the type of the underlying distribution. It also provides a principled way for designing new dimensionality reduction methods.

Neural networks have also been used for dimensionality reduction. The recent advances in deep neural networks and their success in tasks such as object recognition, classification, etc., demonstrate their applicability in many machine learning fields. Deep learning networks have a supreme ability to learn useful and compact representations from the original data [42]. This characteristic has been exploited for designing dimensionality reduction techniques, mostly using auto-encoders. A multilayer autoencoder [43] [44] is a neural network composed of an odd number of hidden layers. The network is constructed so that the input and output layers have $D$ nodes, where $D$ is the dimensionality of the original high-dimensional data, and the middle layer has $d$ nodes, where $d$ is the desired dimensionality of the low-dimensional output space. The original data are provided as input to the network and the training is performed so that the error between the input and the output points, both of dimensionality $D$, is minimized. After the network is trained, the low-dimensional representation of the data can be computed by using the original data points as input and extracting the output values of the middle layer, which are of dimensionality $d$. A single set
of transformation coefficients between the input and middle values, as well as between the middle and the output values, are learned, using all the training data, thus this method is hereby categorized as a global method. Both linear and non-linear activation functions can be used, allowing the mapping between the high-dimensional and the low-dimensional spaces to be either linear or non-linear. Drawing on the auto-encoder idea, the authors of [45] pre-trained a deep network in order to learn binary encodings for speech spectrograms and then duplicated the encoders in reverse order to create a complete encoder-decoder scheme. This auto-encoder was fine-tuned so that the output resembled the input as much as possible. In [46], a variation of auto-encoders, namely Contractive Auto-Encoders, is proposed, where the input space is enforced to be contracted around the training samples, so that the learned features are invariant to local changes in direction around the training points, thus learning the low-dimensional manifold spanned by the data.

2.1.2 Local methods

The assumption implicitly made by all dimensionality reduction methods is that the high-dimensional data are of low intrinsic dimensionality, i.e. they lie on, or near, a low-dimensional manifold existing in the high-dimensional space. The global methods, presented in the previous section, further assume that this manifold is linear, i.e. a hyperplane, and does not contain any curves and twists. In order to handle cases of non-linear manifolds, another large family of methods employ the use of graphs of local neighborhoods around the data points.

One of the first such methods is Isomap [47], from the description of which, a representative use of neighborhood graphs can be illustrated. The first step of Isomap is to construct a neighborhood graph from the original high-dimensional data. This graph has the data points as its vertices and there are edges between two points, if one of them is among the $k$ nearest neighbors of the other. The neighbors are calculated with respect to a selected distance measure, usually Euclidean distance. Ideally, the structure of the constructed graph follows the low-dimensional manifold, even if its geometry is highly non-linear. Isomap proceeds by calculating the geodesic distance between two points, as the shortest path, on the neighborhood graph, between the two points. Multidimensional Scaling is then used, with
these geodesic distances as its input, in order to find the low-dimensional mapping. This mapping is equivalent to unfolding the manifold spanned by the data.

The ability of neighborhood graphs to follow the geometry of the non-linear low-dimensional manifolds has been exploited in numerous dimensionality reduction methods. In **Locally Linear Embedding (LLE)** [48], the assumption is made that, in the near neighborhood of each point, the geometry of the low-dimensional manifold is linear. Each point is thus approximated as a linear combination of its $k$ nearest neighbors. After the coefficients of this linear combination are found, the method proceeds by computing the points of the low-dimensional output space, in such a way that the linear relationships of the neighborhoods are preserved.

In another graph-based dimensionality reduction method, namely **Laplacian Eigenmaps** [49], the low-dimensional projection preserves the point neighborhoods. Points which are close to each other in the high-dimensional space are mapped to points that are close to each other in the low-dimensional space as well, while no care is taken for points that are away from each other. A neighborhood graph, encoded in the corresponding affinity matrix, is formed from the data, where each point is connected to its $k$ nearest neighbors, with edges whose weights are inversely proportional to the distance between the corresponding points. The low-dimensional projections of the original points are computed as the eigenvectors of a generalized eigenproblem involving the Laplacian of the neighborhood graph.

A drawback of Laplacian Eigenmaps is that it can only handle the training data which were used for the construction of the graph and not new test data. The **Locality Preserving Projections (LPP)** method [50] overcomes this problem by considering that the output points are linear projections of the original points and seeking this linear mapping, instead of directly computing the output points. The solution to this problem is computed from a slight modification of the eigenproblem of Laplacian Eigenmaps. If the linear transformation is not enough for the final projection to be adequate, or if the original data are not absolutely known, but only their pairwise distances, the kernel trick can be applied. This leads to the kernel LPP method [50], where the output points $Y$ are considered to be linear transformations of the entries of a kernel matrix constructed from the original high-dimensional data.
The methods described above perform in an unsupervised manner, by considering only the available data, with no extra information about their ground truth class labels, if any. As already mentioned for the global methods for dimensionality reduction, utilizing information about the classes in which the data belong, if such information exists, can produce low-dimensional embeddings which facilitate further processing, such as classification, of the data. A local method following this principle is **Local Discriminant Embedding (LDE)** [51]. Two graphs are formed from the data, an affinity graph, encoding the within-class data neighborhoods, and a penalty graph, encoding the between-class data neighborhoods. The affinity graph is constructed similar to the neighborhood graph of Laplacian Eigenmaps [49], by connecting each point $x_i$ to its $k$ nearest neighbors, but considering as neighbors only points that belong to the same class as $x_i$. The penalty graph is similarly formed by connecting each point $x_i$ to its $k$ nearest neighbors, but considering only points that belong to classes different than $x_i$. The goal is to find a low-dimensional embedding which minimizes the within-class distances between the points, while maximizing the between-class distances. This embedding is calculated by computing the largest eigenvalues in a generalized eigenvalue problem involving the Laplacian matrices of the penalty and the affinity graphs, at each side of the eigenvalue equation, respectively.

Kernel-based extensions of many of the above methods also exist [52], [53], [50], [23], [51]. Apart from extending these methods to non-linear ones, the use of kernels also allows their application to cases where data are in non-vectorial form, and only distances or similarities among them are known.

As already mentioned in the introduction, dimensionality reduction is often used as a data pre-processing step, in order for machine learning tasks, such as classification and clustering, to be performed more efficiently. There are many cases where the data of a single class or cluster consist themselves of multiple sub-clusters, i.e. they contain multiple modes. In order for dimensionality reduction methods to capture this aspect and provide embeddings where the intra- and inter-class relationships are best preserved, supervision information can be used. In [54], a variation of LDA is proposed, where, instead of learning a global data transformation, the data are split into a number of clusters and a different transformation is learned locally for each cluster. This method, namely the LDA mixture model, formulated as an extension of a mixture of local PCA transformations, manages to sep-
arate classes even when they consist of multiple clusters. However, it does not consider the alignment of the transformations in the different clusters, so that the final transformation of the whole data space may not be accurate. The Locally Linear Discriminant Analysis (LLDA) proposed in [55] overcomes this problem, by constructing the between-class and within-class scatter matrices so that alignment of the local transformations is enforced. The method presented in [56] is a combination of FDA [38] and LPP [50] and makes use of labeled points in order to handle data with multiple modes per class. In [57], the authors use the LPP dimensionality reduction method [50] in combination with supervision in the form of pair-wise constraints. These constraints are either must-link constraints, determining pairs of points that should be considered to be in the same class, or cannot-link constraints, determining points that should not be considered in the same class.

Several of the above mentioned graph-based methods share a common pattern: A neighborhood graph is constructed from the high-dimensional data and then an optimization problem is formulated, involving the affinity matrix of the neighborhood graph. The solution to the optimization problem is usually computed as the solution to a generalized eigenvalue problem. The authors of [23] proposed a common framework, namely the **Graph embedding** framework, which includes a large range of dimensionality reduction methods as its sub-cases. In the unimodal graph embedding dimensionality reduction framework of [23], an affinity graph is formed among the data and the goal is to map the data on a low-dimensional space, so that the affinities of the graph are best preserved. A penalty graph is also often used, which determines which points should not be put close to each other in the low-dimensional space. Different dimensionality reduction methods, such as LDA, Isomap, LLE, Laplacian Eigenmaps, even PCA, can be modeled by modifying the way that the affinity and penalty matrices are computed.

Graph embedding dimensionality reduction will be used in this thesis as the background for the proposed Multimodal Graph Embedding approach, so hereby some more formal notation is introduced, for later reference. In order to create the affinity graph, a set of $N$ input points $x_1, x_2, \ldots, x_N$, $x_i \in \mathbb{R}^{D_1}$ is considered, where $D_1$ is the dimensionality of the input points, which is usually high. The points are grouped in a matrix $X \in \mathbb{R}^{D_1 \times N}$, whose $i$-th column is $x_i$. A graph $G = \{X, W\}$ is formed as follows: Two data points $x_i$ and $x_j$ are considered to be connected if $x_i$ is among the $k$
nearest neighbors of $x_j$, or $x_j$ is among the $k$ nearest neighbors of $x_i$. The elements $w_{ij}$, $i, j \in 1 \ldots N$ of the affinity matrix $W \in \mathbb{R}^{N \times N}$ are defined as:

$$w_{ij} = \begin{cases} s(x_i, x_j), & \text{if } x_i \text{ and } x_j \text{ are connected} \\ 0, & \text{otherwise} \end{cases}$$

where $s$ may be dependent on the pairwise distance between the points, e.g. $s(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma^2}$, or may be defined simply as $s(x_i, x_j) = 1$.

The elements $w_{ij}$, $i, j \in 1 \ldots N$ of the affinity matrix $W \in \mathbb{R}^{N \times N}$ are defined as:

$$w_{ij} = \begin{cases} s(x_i, x_j), & \text{if } x_i \text{ and } x_j \text{ are connected} \\ 0, & \text{otherwise} \end{cases}$$

where $s$ may be dependent on the pairwise distance between the points, e.g. $s(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma^2}$, or may be defined simply as $s(x_i, x_j) = 1$.

The goal of the graph embedding dimensionality reduction methods is to map the input points $x_1, x_2, \ldots, x_N, x_i \in \mathbb{R}^{D_1}$ on output points $y_1, y_2, \ldots, y_N, y_i \in \mathbb{R}^{D_2}$, where $D_2$ is the dimensionality of the output points, which is significantly lower than the dimensionality of the input points, $D_2 << D_1$.

The graph embedding dimensionality reduction problem can be formulated in the three formulations presented in Table 1. In the direct formulation, the output is the low-dimensional points $Y$. In linearization, the output points are considered as linear transformations of the input points, so the output is the transformation coefficients $V$. In kernelization, the output points are considered as transformations of a kernel matrix constructed from the input data, so the output is the transformation coefficients $A$. The linearization and kernelization formulations allow data points outside of the original dataset to be mapped on the low-dimensional space as well. Methods such as Laplacian Eigenmaps and LLE fall in the direct formulation. Methods such as LPP fall in the linearization formulation, while methods such as kernel LPP fall in the kernelization formulation of the graph embedding framework. In any case, the solution is formulated as a generalized eigenvalue problem involving the Laplacian of the affinity graph, $L$, and a constraint, or penalty, matrix $B$. The Laplacian of a graph $G(X, W)$ is defined as:

$$L = D - W,$$

where $D$ is the diagonal matrix with the row (or column) sums of $W$. The constraint matrix $B$ can be the Laplacian matrix of a penalty graph $G'$, containing information about which objects should be kept apart, or the diagonal matrix $D$.

The introduction of the graph embedding framework leads to a better understanding of how dimensionality reduction (DR) methods work, as the various DR methods are expressed simply as different definitions of the neighborhood graphs. It also facilitates the design of new DR techniques, by merely altering the way that the neighborhood graphs are constructed.
Continuing with the examination of the related literature, since the construction of the neighborhood graph plays an important role in the application of the graph embedding framework, methods have been proposed to construct optimal affinity graphs, which are suitable for machine learning tasks. In [31], the optimal affinity graph is constructed based on the consensus of multiple runs of the k-nearest neighbors procedure, for various neighborhood sizes. This results in more robust relations of similarity or dissimilarity between the vertices. In [25], random forests [24] are used in an unsupervised manner, in order to construct the affinity matrix, leading to supreme performance.

Local dimensionality reduction methods, such as Laplacian Eigenmaps, Isomap and Locally Linear Embedding have also been formulated in a probabilistic framework [58]. The framework is formulated in terms of maximizing the entropy of the data while keeping neighbouring data nearby, thus unfolding the underlying manifold. This approach, namely Maximum Entropy Unfolding (MEU), can describe other existing local dimensionality reduction methods as its sub-cases, by altering the error objective function considered. Furthermore, the fact that it is probabilistic in nature allows it to better handle missing data, to easily be extended to mixture models and to be combined with other probabilistic models.

A different approach for dimensionality reduction, still using graphs, is

<table>
<thead>
<tr>
<th>Formulation</th>
<th>Minimization problem</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>( \min_Y \sum_{i,j=1}^N</td>
<td></td>
</tr>
</tbody>
</table>
| Linearization| \( \min_V \sum_{i,j=1}^N ||V^T x_i - V^T x_j||^2 w_{ij} = \text{tr}(V^TXLX^T V) \) \(\text{s.t.} \ \text{tr}(V^TXBX^TV) = 1 \) | \(XLX^Tv = \lambda XBX^Tv\) \( (4) \)
| \( y = V^Tx \) |
| Kernelization| \( \min_A \sum_{i,j=1}^N ||A^T k_i - A^T k_j||^2 w_{ij} = \text{tr}(A^T K LKA) \) \(\text{s.t.} \ \text{tr}(A^T KBKA) = 1 \) | \(KLKA = \lambda K LKA\) \( (6) \)
| \( y = A^T k \) |
to use graph aesthetic measures in order to properly place the graph vertices in the low-dimensional space. Graph aesthetic measures are mostly used to draw an arbitrary graph on two dimensions, so that the structure of the graph is apparent in an aesthetically pleasing manner, with as little clutter as possible. Such measures consider, for instance the number of edge crossings in the final drawing or the angles formed by the edges around each vertex [59]. A prominent category of graph drawing algorithms considering such measures, which can be easily extended to more than two dimensions, is the force-directed placement algorithms or spring-embedding algorithms. Force-directed algorithms [60] [61] proceed by considering the vertices and the edges of the graph as the elements of a mechanical system. In particular, vertices are considered as repelling charges, each one imposing a repelling force to all others, while edges are considered as attractive springs attached to pairs of vertices. This mechanical system is allowed to behave freely until it reaches an equilibrium. The repelling forces among the particles are expanding the graph, while the attractive forces of the springs are keeping connected particles close to each other. This leads to a placement of the graph that is more comprehensible and represents the underlying structure adequately, by preserving the point neighborhoods. Variations also consider forces between edges, as in [62]. Although such methods are commonly used to project graphs on just two dimensions, allowing the particles to move in a higher-dimensional space can lead to projections with more dimensions, which may be more accurate.

Neural network-based local dimensionality reduction methods also exist. The Self-Organizing Map (SOM) [63] [64] is a popular nonlinear dimensionality reduction method that maps high-dimensional data to two dimensions. A SOM is a two-dimensional grid of nodes, where each node is associated with a model vector. A model is a vector of the same dimension as the high-dimensional input data, representing a local subset of the data. After an iterative process, the models are tuned to the data, i.e. they take such values that models which are nearby on the map represent data areas that are close to each other in the original space. In other words, the two-dimensional map is stretched so as to follow the (assumed two-dimensional) manifold spanned by the high-dimensional data. The models are initialized to random values (or values that roughly represent the data manifold, for faster convergence). At each iteration, the first step is to select an input
point and find the model closest to it, which is called the winner model. The second step is to update the winner model, as well as its neighbors on the map, in order to be brought closer to the input point. These steps are repeated until the model values reach a steady state. After this procedure, points that are close to each other in the original space are represented by models that are close together on the map. The procedure followed by the SOM resembles the \(k\)-means clustering method. The map’s models can be considered as the cluster centers, which are tuned to the data. The novelty of the SOM is that it not only finds the cluster centers, but also puts them in such positions on the map, so that the topology of the original data is preserved (i.e. similar clusters are put close to each other on the map). In [65], a modification of the original SOM algorithm is presented, which allows the original data to be non-vectorial, or to be such that only distances among them are known. Another variant is proposed in [66], where an input vector is matched not to one model, but to a linear combination of models.

2.2 Multimodal learning

Most existing dimensionality reduction methods perform by considering a single data feature, e.g. the color histogram of images [67], or the frequency spectrum of sounds [68]. However, there are many occasions where multiple representations are available for a set of data. For instance, multiple descriptors may be extracted from an image, describing properties, such as the textures or the shapes in the image [28]. Another case is multimedia of different forms describing the same semantic concept, such as the images, sounds and videos contained in a single web page, covering a single semantic topic [69]. Visualization methods can utilize the multiple available data descriptions (the so-called modalities), in order to provide outcomes which more effectively address the human perception and reveal the semantic relationships among the data.

Generally, in the literature, the problem of combining different representations of the same concept is handled by techniques of multimodal fusion. Multimodal fusion is used in cases where information is available in more than one information channels. These information channels, so-called modalities, usually represent different sensory paths, for example the visual channel, the audio channel, the touch channel etc. This multitude of information sources provides complementary or redundant information, which
can be exploited by learning methods, in order to make better judgments.

Multimodal fusion methods can be characterized by the level at which fusion is done, during the information processing. They range from early fusion ones to late fusion ones. In the following, a brief description of early and late fusion methods will be presented. A comprehensive presentation of fusion techniques can be found in [7].

In early fusion methods, fusion happens early during the information processing, at the feature level. The features which are extracted from multimedia are merged into one multimodal feature, which is afterwards used in a learning task, as if it originated from just one source. The goal of early fusion techniques is that this multimodal feature captures all the necessary information of all the available modalities and at the same time it is in a compact form that can be easily used by a learning algorithm. Figure 1a shows the early fusion method graphically. An important issue that arises in early fusion is the compatibility among the unimodal features. Multimedia of different types, or different characteristics of the same media type, are described with very different descriptors, differing in dimensionality, value range and value meaning. A common approach to deal with the problem of compatibility is to try to project the unimodal features onto a space that is common for all modalities. Examples of early fusion methods can be found in [70] and [69].

In late fusion methods, fusion is performed at a late stage during the information processing. Analysis or learning tasks (e.g. classification) are performed separately for each modality and the results of these unimodal procedures are then fused in order to produce a final result or decision. Figure 1b depicts the late fusion procedure. In late fusion methods, fusion is performed in a higher, semantic, level and thus problems of compatibility do not arise, as in early fusion. There are many unimodal learners, some of which may even not be reliable enough (e.g. in the presence of noise) and the goal is to combine them in order to produce a robust decision. The works in [71] and [72] are examples of the late fusion technique.

Early fusion methods have the advantage that, using the raw features, correlations between modalities can be discovered and exploited for the learning task. This cannot be accomplished in late fusion techniques, where fusion is done after the raw features have been processed. On the other hand, as already mentioned, the compatibility problem of early fusion methods does
Figure 1: Levels of multimodal fusion. (a) Early fusion: Unimodal features $F_i$ are fused to produce a multimodal feature $F$, which is afterwards used in a learning process ($L$) in order to produce a final decision $D$. (b) Late fusion: Separate learners $L_i$ are applied to each unimodal feature $F_i$ to produce separate decisions $D_i$. These unimodal decisions are then fused to produce the final decision $D$.

not arise in late fusion ones. In addition to this, another advantage of the late fusion approaches over the early fusion ones is that the unimodal learning processes that precede fusion can be performed by learners of different types, each one most appropriate for each modality [7].

The multimodality of the human brain has inspired researchers and efforts have been made to apply it in designing systems that can more accurately recognize concepts and make decisions. Atrey et al. in [7] present a review of methodologies used for multimodal fusion in various application areas. Potential applications of multimodal fusion are numerous. Some representative work that has already been done is briefly mentioned in this section.

**Multi-sensor fusion**  A common application of multimodal fusion is in cases where there is a multitude of sensors and the information gathered by them must be integrated. Such cases emerge in application fields such as robotics [73], biomedical monitoring [74], environment monitoring [75] and transportation systems [76]. Luo et al. in [77] present a review of uses of multimodal fusion techniques in many application areas. In all such applications, the complementary or redundant information of the different sensor types result in predictions that are more accurate and less prone to noise.
**Audio-visual speech recognition** One of the most prominent indications of multimodality in the human brain is speech recognition, where both the audio and the visual modalities are used. This paradigm has been used by researchers to build robust speech recognition systems (e.g. [78] [79] [80] [11] [12]). In such systems, audio units of information, called phonemes are combined with simultaneously occurring visual ones, called visemes, to reduce the uncertainty of the recognized speech. This is especially helpful in cases where one of the two modalities is not reliable enough, for example in the presence of acoustic noise.

**Biometric identification** Biometric characteristics, such as a person’s image, voice, gait etc., are currently used for person identification. A significant increase in the accuracy of biometric identification systems can be accomplished if such characteristics are combined. Multimodal biometric systems are potentially more robust and less obtrusive. Example applications of multimodal biometric identification systems and methods are presented in [81] [82] [83] [84] [13] [14]. Existing work has also used measures of the quality of the data acquired from a modality, in order to promote the most accurate modalities during fusion [85], while other work considers not only the reliability of each modality but also time-related criteria, in order to provide continuous verification [86].

**Multimodal human-computer interaction** A lot of research has been done on designing intuitive and easy to use interfaces for human-computer communication. More straightforward and comfortable user interfaces can be implemented if a combination of standard unimodal communication ways, such as using a keyboard, a pen, a touch screen or a microphone, is used. For example the user could simultaneously write something with the pen and say something to the microphone, in order to convey some message [87]. Multimodal communication is more important nowadays, with the wide availability of mobile devices that integrate many types of inputs (camera, microphone, touchpad, keyboard, accelerometer) and outputs (screen, speakers, vibration). Other higher-level types of input, such as gaze or gesture recognition can also be used as additional communication channels [88]. Methods to accomplish multimodal human-computer interaction can be found in [89] [88] [15] [16].
Multimedia indexing and retrieval  Another area with great potential for multimodal fusion, and which is directly related to the present research, is multimedia indexing for the purposes of retrieval, classification etc. Very often, media of different types occur at the same time or in the same place or context. For instance, in video processing, visual information of the video sequence is accompanied by audio information and maybe text information (if text appears in the video, or if there are subtitles). Combination of all these modalities can lead to more accurate results in tasks such as automatic video annotation or event detection. Simultaneous co-occurrence of media of different types also happens, for example, in webpages, where text, images, audio and video are contained in the same page, thus having common semantics. Correlations among these modalities can be learned in order to enhance media annotation and retrieval. Some example applications of the above principles are presented in [90], [91] and [92].

In this section, state-of-the-art methods for fusing information of many modalities, are presented, focused on methods which can be applied for the task of mapping a database of multimodal objects on a common space and performing dimensionality reduction, or related tasks, such as visualization and clustering.

In application areas such as audio-visual speech recognition, biometric identification, multimodal video indexing or multimodal computer interfaces, multimodal fusion is based on the fact that information of different modalities is received simultaneously in time. For example, a speech recognition system fuses an audio phoneme and a simultaneous visual viseme in order to decide about the actual linguistic information [78]. The fact that these two pieces of information occur simultaneously in time is an indication that they represent the same higher-level concept, so the system can safely correlate them.

When handling multimodal multimedia databases, there is a database of multimedia items of various modalities and some kind of analysis needs to be done on them, such as indexing, retrieval, classification, clustering, dimensionality reduction or visualization. In this type of problems, time cannot be used as a context for the correlation of different types of media. What is commonly done is that the same label is attached to media items of different types, as a form of supervision, to indicate that these media represent the same underlying concept.
A common way to provide such supervision is by considering that media items representing the same concept are grouped in higher-level collections. Zhang and Weng, in [69], introduce the concept of a multimedia bag. It is a set of unlabeled media items, grouped in a bag, which is labeled by a semantic concept, common for all the contents of the bag. For example, a bag labeled as “car” could consist of images of cars, sound files relative to cars or text information about cars. In [93], Zhuang et al. use the similar notion of a multimedia document, which is a collection of media items of various modalities that share a common semantic concept. The concept of multimedia bags is similar to the bags of instances in Multiple Instance Learning (MIL) [94] [95], where positive or negative labels are assigned to entire data collections, rather than single instances. However, a multimedia bag is different in that the “instances” in the bag are generally of different modalities, thus of different types, size, etc., and, in multimodal learning, the content of the bag is characterized by utilizing all the available modalities, rather than the single positive instance that characterizes a multiple instance bag.

Collections of conceptually similar media of different modalities are a convenient way of describing a multimodal database and they will also be used in the present research. In general, low-level features are extracted from the media contained in such objects. These features are usually numerical vectors encoding characteristics of the media, such as the colors or the textures of an image, or the acoustic frequency spectrum of a sound. Since the representation of the media is in the form of vectors, there is no distinction as to the type of media that the vector is extracted from. Regardless of the media contained in a collection being e.g. an image or a sound, the treatment is the same. Other forms of descriptors for a modality are also possible, such as a collection of vectors, as e.g. in the popular SIFT image descriptor [96], where a collection of vectors, corresponding to various points in an image, describe a single image. Since most dimensionality reduction methods are based on distances between descriptors, rather than on the descriptors themselves, if distance measures are defined between such other kinds of descriptors, then they can be easily used as well. Hereby, the term multimodal object will be used to describe a collection of media descriptors, extracted from media of the same or of diverse types.

The existing approaches to handling multimodal data in the field of di-
mensionality reduction, as well as in related fields, such as clustering, can be organized in three broad categories:

- Multimodal fusion using a common representation
- Multimodal fusion using simultaneous learning
- Multimodal learning without automatic fusion

Methods of the first category attempt to combine the multiple modalities in a single common representation, e.g. by projecting modalities to a common space or merging unimodal distances between objects, and then use this common representation for dimensionality reduction, clustering, etc. On the other hand, methods of the second category attempt to find a solution by simultaneously learning across all modalities, e.g. by iteratively alternating among the multiple modalities to fine-tune the solution. Finally, methods in the third category do not attempt to combine the multiple modalities, but instead preserve as much information from the individual modalities as possible, leaving the ultimate combination be performed by an external actor, such as a human user.

Each of these three categories can be further divided into more sub-categories, corresponding to families of related methods. In the following sections, each of the three broad categories and their sub-categories will be presented, providing references to representative work for each family of methods. The overall organization of the covered approaches to multimodal learning is depicted in Fig. 2.

2.2.1 Multimodal fusion using a common representation

**Fusion using projection to common space** For problems related to analysis of a multimodal dataset, an important task that usually needs to be solved is being able to compare items of different types, such as the image of a tiger with the sound of a dog. As already mentioned, an issue regarding early fusion methods is the compatibility of the different unimodal media types. Being able to compare items of different modalities allows applications such as multimodal retrieval (which needs to compare items to find the most relevant ones to a query) or multimodal dimensionality reduction (which needs to compare items to decide how they should be mapped on the low-dimensional space) to be implemented.
Figure 2: Organization of multimodal learning approaches. The approaches proposed in this thesis are noted in bold, illustrating where they belong in this organization.

Two different approaches have generally been used to handle the problem of compatibility. In one approach, the effort is to compare single unimodal media items of different types, while in the other approach, the effort is to compare whole multimodal objects. The first approach leads to defining a common space for all modalities, where an item can be mapped, irrespectively of its modality, while the second one leads to methods that map whole multimodal objects in a multimodal space.

The idea behind the approach of a common-space, where each modality can be mapped, is the following: Since media items of different types cannot be directly compared (for example one cannot compare the color features extracted from an image to the spectral features of a sound), it would be convenient if the different unimodal features could be projected to a space that is common for all modalities. Then comparisons between items of any type could be made in this common space and distances between them could be defined, allowing the implementation of a large range of learning algorithms. Many works have already dealt with this problem and the most prominent solution for finding these projections is with the use of Canonical Correlation Analysis.

Canonical Correlation Analysis (CCA) [97] [98] is a statistical method that considers two sets of random variables and finds linear combinations of the components of each set, so that the correlation between the linear combinations of the two sets is maximized. The linear combinations are computed by solving an eigenproblem. The eigenvectors of this problem define an axis system for each variable. The projections of the variables
onto these defined spaces are maximally correlated. Hence, these spaces can be used as a common space on which data of both variables can be projected onto, and at the same time preserving their correlations as much as possible. CCA can be extended for non-linear projections, by using the kernel trick.

In the context of multimodal data, data of two different modalities can be considered as the two random variables and the semantic labels of the data can be used to link together pairs of data from the two different modalities, so that CCA can be applied. After the CCA projections are found, they can be used to project the data of the two modalities on the common space. Comparisons can then be done between data of any modality in this common space and distance metrics can be defined for them.

In [98], this method is used to correlate images with their associated text, for the purposes of cross-media retrieval (retrieving images with a text query, when no text tag is associated with the images). The authors of [69] consider more than two modalities and calculate projections for each pair of them. After calculating pairwise distances between data of the different modalities in the common space, they take an extra step and find the manifold spanned by the data in this common space, in order to better analyze the data. In [99], kernel CCA is used to learn the projections and a graph-based method is used to discover the manifold structure of the common space. The authors of [100] use the CCA-based common subspace method to accomplish retrieval of text using image queries and vice-versa.

CCA has been formulated in a probabilistic manner [101], which allows for easier handling of missing data, as well as for combination of multiple probabilistic models in a mixture, in order to describe more complex datasets. In this formulation, the input data are considered to be drawn from a Gaussian distribution. In order to cope with the poor performance of Gaussian-based CCA to handle outliers, the authors of [102] proposed using the Student-t distribution, instead.

Several extensions of CCA have been proposed in order to provide more accurate representations in various scenarios. CCA considers just two views of the data. In order to handle multiple views, one common practice is to consider the multiple views in pairs and perform CCA for each pair. However, this approach may not be sufficient in many cases. In [103], the Multi-view CCA method is proposed, where multiple views are considered simultaneously in order to construct the common projection space. CCA has also been
extended with regard to discriminant analysis, so that the common space fulfills certain criteria, such as maximum class separation. The authors of [104] propose the Correlation Discriminant Analysis (CDA) method, where correlation metrics are used to maximize intra-class correlation, while minimizing the inter-class correlation. Recently, the authors of [105] proposed the Multi-view Discriminant Analysis method for constructing a common projection space. This method again considers intra- and inter-class correlations, but extends them to multiple views.

The other type of common space methods are ones which attempt to map whole multimodal objects on a common multimodal space. The naive approach would be the space defined by the concatenation of the multiple feature vectors comprising a multimodal object. However, this approach is not so flexible, as some modalities may be more important than others, or there may be correlations among the modalities, which should be useful. A more flexible approach is using tensors to describe multimodal objects. In [106], the multiple modalities are considered as the orders of a tensor, comprising the so-called TensorShot representation of a multimodal object. These tensor representations are then used to compute distances among objects and perform dimensionality reduction. In the recent work of [107], tensor representations are used with regard to multiple features extracted from videos, in order to utilize the correlations among the features, leading to improved video fingerprinting. Tensor-based approaches require that the multimodal objects consist of the same number and same types of modalities. More flexible techniques consider only distances among multimodal objects, without requiring an explicit representation in some high-order space. Such methods are discussed below, in the “Late fusion approaches” section.

**Modality selection** The aim of projecting multimodal data on a common space, as described in the previous section, is to be able to compare single media items which are of different modalities. For this purpose, multimodal objects are used as a form of supervision, to state that some objects of different modality have the same semantics. Another approach to analyzing multimodal databases is to consider that the basic units are the multimodal objects themselves and to try to make comparisons between whole multimodal objects. This problem setting is close to the so-called multi-view setting. In multi-view learning, data are considered as consisting of many
representations, or views, and the task is to make use of all of them in order to enhance data analysis. In multimodal databases, multimodal objects can be seen as concepts and each of these concepts has many different views, which are the media items that a multimodal object consists of. Therefore, techniques relevant to multi-view learning can be used to exploit the multitude of modalities.

An important aspect introduced by multimodal fusion methods is the determination of the importance of each modality for a specific task, otherwise referred to as feature selection or modality selection. This importance is usually expressed as a numerical weight assigned to a modality. For a specific learning task, some modalities of a multimodal object may be more informative than others, so these modalities should be considered as more important during the learning method. For example, for the task of separating objects related to the sea from objects related to the forest, the image modality may be more informative than the sound modality, as separating images of seas and forests may be straightforward, while separating their associated sounds may be confusing.

Modality weighting deals with promoting the best modalities for the specific task, either by using only the important ones, or by assigning importance weights to each modality. In multiple kernel learning [108] [28], for instance, weights are assigned to each unimodal kernel matrix, before they are added. These weights represent the importance of each modality for the learning task, with large weights meaning that the respective modalities are important.

In [109], a feature vector is transformed into a set of most informative modalities. Linear dimensionality reduction techniques, such as Principal Component Analysis [110] and Independent Component Analysis [111] are used to discover modalities that are least correlated, and thus most complementary to each other, hence most informative. Each modality is a combination of the feature values, where each value is given a specific weight, relative to its importance.

Another discipline related to automatic modality weighting is subspace clustering, or projected clustering. The problem here is clustering high-dimensional data. An observation is made that some dimensions of the data may be irrelevant for specific clusters, and thus they should not be taken into account during the clustering. Subspace or projected clustering tries
to find subspaces of the original space, so that more clear clusters emerge by projecting the data on these subspaces. The works of [112] and [113] are two examples of projected clustering and subspace clustering respectively. A difference and an advantage of such subspace-based methods over feature-selection methods which are based on a naive weighted sum is that in subspace clustering, the weights assigned to each modality are not global, for all the data, but they are different for each cluster. Thus they are more flexible in encoding the different characteristics of the data.

Instead of automatically discovering the best modalities, feature selection can also be performed in a user-supervised way. In many cases, the system alone cannot determine which modalities are most important for a specific task. In such cases, user supervision can assist the system to adapt itself, by changing the weights assigned to each feature, so that they better reflect the higher-level concepts. In [114], an image retrieval system is described, where several image descriptors are considered and the most relevant ones for the user are selected, based on user supervision in the form of relevance feedback. A similar concept is used in [115], where the user selects pairs of objects to inform the system that they are conceptually similar, and the weights of features are adjusted accordingly. Another similar example is [116], where the user interactively supplies the system with more labeled data.

**Late fusion approaches** The works presented in the previous two sections handle the multiple modalities by considering the low-level features of each modality themselves, either by computing a projection to a common feature space, or selecting among the multiple features. This is an instance of early fusion, since the multiple modalities are handled early in the learning process, before any further application, such as classification or clustering is applied. Another class of multi-view methods are ones that, instead of combining the raw features, combine higher-level characteristics of them, such as similarities or distances among them. This is an instance of late fusion, since the multiple modalities are handled in a later step in the whole learning procedure. Late fusion approaches are closely related to boosting [117], or ensemble learning [118], where a set of weak classifiers are merged in order to produce a strong one. In the same way, the unimodal similarities or distances between the media items of multimodal objects can be seen as weak measures of relationship, which cannot adequately reflect
the underlying conceptual relations. A multimodal similarity or distance measure, defined by combining these unimodal ones, can hopefully better represent the underlying semantics. The calculated multimodal relations can afterwards be used in learning tasks, as if they originated from a single source. They can be used directly for purposes such as retrieval, or they can be used to map the multimodal objects as points in some multimodal space, for purposes such as dimensionality reduction and clustering.

There are many means to combine the unimodal (dis)similarities into a multimodal one. Rule-based methods exist, where the combination is based upon certain rules. The authors of [91] construct a similarity matrix, containing the similarities between every pair of multimodal objects, by considering the $k$-nearest neighbors of them. The set of nearest neighbors of a multimodal object is a concatenation of the multimodal objects whose media items are nearest neighbors of the media items of the first object. Some additional rules are applied to limit the number of neighbors to $k$. In their approach, the produced similarity matrix contains just 0s and 1s, thus having eliminated the need for compatibility transformations. As a final step, Laplacian Eigenmaps [49] are used to map the multimodal objects as points in space. A late fusion approach is presented in [119], where similarities between videos with respect to text and visual attributes are combined, in order to detect an event structure.

However, the most common way to fuse unimodal distances is using a weighted sum. The weights of this sum represent the importance of each modality in the determination of the final distance. For instance, if some modalities are irrelevant for a specific task, they can be assigned low weights. In [120], unimodal distances are defined between two multimodal objects, one for each modality. The constituting media items of each modality are used to find the unimodal distance in the respective modality. Then the unimodal distances are fused into a multimodal one, using a normalized weighted sum. The weights of this sum are calculated by the precisions of retrievals which are based on each modality separately. In other words, they represent how accurate each modality is for the specific task. In [121] and [122], similarity graphs encoding the unimodal similarities are constructed and then a weighted sum is used, along with supervision information, to fuse the unimodal graph Laplacians. The resulting graph is used in a regularization framework for the purpose of label propagation. The modality weights are
experimentally set to fixed values in [121], while in [122] they are automatically calculated by including them as additional variables in the optimization. The approach of [122] is similar in concept to the proposed Multimodal Graph Embedding approach, described in Section 4. However, in [122], the modality weights are learned together with the dimensionality reduction coefficients, which restricts this approach to a specific dimensionality reduction technique. On the other hand, in the hereby proposed approach, the weights are learned separately, which allows them to be used with any dimensionality reduction method thereafter.

The weighted sum approach is closely related to multiple kernel learning [108]. In machine learning, kernels allow linear learning methods (such as Support Vector Machines) to be extended into non-linear ones. In such methods, a kernel matrix is commonly used, encoding the pairwise similarities between the data points. In multiple kernel learning, multiple notions of similarity are considered between two data points and thus many kernel matrices are constructed. The multiple kernels are then merged into one, usually by a weighted sum. This setting is closely related to the fusion setting. The unimodal similarities between the items of two multimodal objects can be seen as multiple notions of similarity between the objects. Multiple unimodal similarity matrices are constructed and then merged into a multimodal one, in a similar way as in multiple kernel learning.

In [123] and [28], many descriptors are considered for each data point and a kernel matrix is constructed for each descriptor. These kernel matrices are merged via a weighted sum into a global kernel matrix, which is later used for dimensionality reduction, extending the graph embedding framework of [23] for many modalities. This method, namely Multiple Kernel Learning Dimensionality Reduction (MKL-DR), is based on Multiple Kernel Learning [108], where multiple kernels are combined via a weighted sum.

Some formal notation about MKL-DR is introduced hereby, since MKL-DR is used within the proposed Multimodal Graph Embedding approach. In MKL-DR, there are $M$ kernel matrices $K_1, K_2, \ldots, K_M$, $K_m \in \mathbb{R}^{N \times N}$, $m \in 1 \ldots M$, each corresponding to a different notion of similarity among the data. The total kernel matrix $K_{\beta}$, which is ultimately used in the dimensionality reduction process, is a weighted sum of the multiple kernel matrices,

$$K_{\beta} = \sum_{m=1}^{M} \beta_m K_m,$$
where $\beta_m \geq 0$ are the weights of the sum. The subscript $\beta$ in $K_\beta$ indicates that the total kernel matrix is calculated using the specific weight configuration $\mathbf{\beta} = [\beta_1 \ \beta_2 \ \ldots \ \beta_M]^T$.

Dimensionality reduction is formulated as an optimization problem and the kernel weights are introduced as additional optimization variables, in order to automatically discover the optimal combination of kernels. The problem is similar to the kernelization formulation of graph embedding of Table 1, where the kernel matrix $K$ is replaced by the multimodal kernel matrix $K_\beta$. The variables of minimization are both the transformation matrix $A$, which includes the coefficients for mapping the data described in the kernel matrix on the low-dimensional space, and the vector $\mathbf{\beta}$, which contains the modality weights for the construction of the multimodal kernel matrix:

$$
\min_{A,\beta} \quad \text{trace}(A^T K_\beta L K_\beta A)
$$

subject to $\quad \text{trace}(A^T K_\beta B K_\beta A) = 1 \quad (7)$

In this way, the performance of the various unimodal dimensionality reduction methods used is boosted, by exploiting the multiple modalities available. In [124], multiple kernel learning is used for image retrieval. Images are described with many descriptors and a kernel matrix is constructed for each of them. Boosting is then used to merge the multiple kernels. Another application of MKL-DR is presented in [125], where Local Fisher Discriminant Analysis (LFDA) is used as the dimensionality reduction method.

Variations of the MKL-DR technique have also been proposed, in order to improve the efficiency of the method in specific problems. In [29], a method for classifying image sets is proposed. The method uses Multiple Kernel Learning in order to learn a distance metric between image sets. Multiple statistics of different order are computed from each image set and are used as the multiple features, or modalities. The distance between two image sets is defined as a weighted sum of the distances between the various statistical features between the two image sets. The weights of the sum are learned in a localized multiple kernel learning fashion, by solving an optimization problem which uses training data and tries to maximize inter-class distance while minimizing intra-class distance. The technique presented in [126] adds an extra step to Multiple Kernel Learning, by selecting some of the available kernels, in an Ensemble Learning fashion, for the purpose of classification. The selection of kernels is performed in a supervised manner, by selecting
those kernels that achieve higher discrimination and diversity. Recently, in the work of [30], a variation of multiple kernel learning is proposed, where each kernel is weighted according to the discriminative power of the corresponding feature vector. Entropy-based measures are used to assess this discriminative power, resulting in increased performance for image and object recognition.

The basic MKL-DR technique of [28] focuses on the determination of the weights for the weighted sum of the multiple kernels. However, the technique is based on graph embedding dimensionality reduction, which relies upon the definition of appropriate affinity and penalty matrices among the data. The existence of multiple modalities leads to multiple notions of affinity and neighborhood among the multimodal objects, thus to multiple affinity and penalty matrices. Recent works of the literature have focused on the combination of these affinity matrices, apart from the kernel fusion. The authors of [27], consider the problem of spectral clustering, which is in general similar to graph embedding dimensionality reduction. Multiple affinity matrices are considered and a weighted sum of them is used for spectral clustering. The weights of the sum are extra parameters in the optimization problem of spectral clustering and are learned together with the spectral clustering coefficients, in an alternating Expectation-Maximization fashion. In one step, the modality weights are fixed, rendering the problem as a standard spectral clustering one. In the other step, the spectral clustering coefficients are fixed, and the optimization is performed for the modality weights. A different approach is adopted in [127], in order to find the optimal fused affinity matrices to use. An affinity and a penalty matrix are used, as in the graph embedding framework of [23], where the affinity matrix is defined using intra-class similarities and the penalty matrix is defined using inter-class distances. The two matrices are initialized using the average of the multiple kernels as a basis for the data similarities and are subsequently learned using a two step procedure: First, the MKL-DR approach is followed to find the optimal kernel coefficients. Second, these coefficients are used to re-define the affinity matrices, by computing each element in the affinity matrices as the inner product of the corresponding learned kernel coefficient vectors. Although this method is conceptually similar to the proposed Multimodal Graph Embedding approach, presented in Section 4, in the latter, the process of computing the modality weights is detached from
the process of determining the dimensionality reduction parameters, so that any dimensionality reduction technique can be used, not only MKL-DR.

2.2.2 Multimodal fusion using simultaneous learning

**Fusion using co-training**  One of the first attempts to learn from multiple views of data is the co-training setting of Blum and Mitchell [128]. Their motivation is labeling a set of unlabeled data, using a small set of labeled ones. The data consists of two views. Two separate classifiers are used, one for each view and data which have their labels learned using either of the two classifiers are added to the labeled set for a next classification. In this way, information from both views is gradually used to classify the data.

In [129], the authors propose an algorithm that combines the co-training algorithm of [128] with the Expectation-Maximization algorithm [130]. Their algorithm, called co-EM, considers that the data consist of two views (or that their features can be split in two views) and then the EM algorithm is performed for classifying the data. The novelty is that the iterations of the EM algorithm are alternated between the two views, so that information of one view assists in classifying data of the other view.

The co-EM algorithm has been used by researchers to adapt several learning methods to cases where two views of the data are available. In [131], the authors use it to extend the standard Support Vector Machine (SVM) classification method to the multi-view case. The work of [132] is an application of co-EM to partitioning and agglomerative clustering and [133] uses co-EM in combination with a semi-supervised clustering method.

A drawback of the co-training setting is that it is based on the assumption that each of the two views is sufficient for learning, so that the high-level information extracted from it (e.g., labels) can indeed be used as knowledge when learning in the other view. However, this assumption does not always hold.

An idea similar to the co-training approach is to formulate the problem as an optimization problem for one of the modalities and using information extracted from the other modalities as constraints for the optimization. In [134], visual similarity between frames of videos is used as a constraint in a clustering procedure based on the text description associated to the videos.
Fusion using co-clustering  

Dimensionality reduction techniques are usually based on the computation of distances or similarities between the objects considered. In this respect, dimensionality reduction is closely related to clustering, since clustering techniques are also based on the computation of distances or similarities. Therefore, ideas and approaches used for clustering multimodal data can be adapted to address the problem of dimensionality reduction as well.

For the problem of clustering multimodal data, co-clustering is a common approach, where clustering of the data occurs simultaneously for two or more modalities. In [135], a corpus of documents is clustered both for the documents and for the words appearing in them, using the co-occurrence of words and documents. The final clustering is the one which maximizes the mutual information between separate clusterings of documents and words. Another method for co-clustering is presented in [136], where a bipartite graph is formed from the videos of two different sources and co-clustering is accomplished using bipartite spectral clustering.

The problem of co-clustering documents and words is closely related to probabilistic Latent Semantic Analysis (pLSA) [137]. In [137], the co-occurrence of documents and words has been modeled by the joint probability distribution of documents and words, which is modeled using the topics covered in the documents as latent variables. The authors of [138] use the pLSA approach in the area of image classification, by considering images as documents consisting of visual words, such as grass, roads, etc., taken from a vocabulary of visual words. Latent Dirichlet Allocation (LDA) [139] is a similar approach, utilizing a Dirichlet prior distribution for the topics, which generally leads to more accurate mixtures of topics.

Information-theory measures, such as mutual information, have also been used in order to “align” the results of unimodal learning methods. In [140], the authors consider multiple views of multimedia and set all the possible clusterings of the data in these views as random variables in a Markov Random Field (MRF). There is one variable for each modality. These variables are connected in the MRF and the connections show the correlations between the modalities. The purpose of this MRF is to assign values to its variables (i.e. find unimodal clusterings) such that they conform to the training data and at the same time minimize the MRF’s energy. The potential function used in the MRF, representing the disagreement between the clusterings of
two modalities, is the mutual information of the two clusterings. As another example of such methods, the authors of [141] again use mutual information to maximize the agreement between clusterings of different modalities. They propose an extension of the information bottleneck clustering method [142] to the multi-view setting, utilizing co-training, in order to calculate the final clustering in a way that it uses all available information from all the data views.

**Multi-task learning** A related family of methods is the so-called multi-task learning methods. They include techniques for simultaneously handling multiple learning tasks, either by using multiple sources of information to enhance the learning process, or for utilizing information of one source in order to enhance learning from another source. In [143], multi-task learning is formulated as an extension of existing single-task regularization-based learning techniques, such as Support Vector Machines, and is used in classification problems with multiple training sources. In [144], classifiers for multiple tasks are combined using a Dirichlet process-based statistical model to learn the similarities between the multiple classifications. As a paradigm, multi-task learning is related to both co-training and co-clustering techniques.

2.2.3 **Multimodal learning without automatic fusion**

**Perceptual fusion** Combining the multiple modalities, either in a common representation or through simultaneous learning, involves a certain amount of information loss, as some compromise among the modalities is ultimately made, in order to reach a single solution. However, there are approaches in the literature, where such a compromise is not made and the information from all modalities is preserved. The approach of using of glyphs in the field of data visualization is such a paradigm.

Visualizing a set of data is generally performed by mapping characteristics of the data on specific visual cues. A natural extension of visualization techniques in cases where there are multiple representations of the data is mapping the features of each modality on a different visual characteristic. For this purpose, specific data markers, called *glyphs*, are used [145]. Each data item is represented by a glyph and each of the multiple features extracted from the item is mapped to a visual characteristic of the glyph, such as its size, color, shape, position, etc. [145]. Examples of glyphs are Chernoff
faces [146], arrows [147] and stick figures [148].

In glyph-based visualizations, the available information is mapped to different visual characteristics and it is the visual perception of the human viewer which fuses this information in order to extract the semantics from the raw data. In other words, the individual modalities of the multimodal objects of the dataset are transformed into suitable visual representations and are then mapped on the “space” of visual perception, where fusion is performed by the human mind.

A drawback of glyphs is that some visual characteristics are more easily perceived by humans than others. For instance, relative position is easier to perceive than relative size [6]. As a result, those data features which are mapped on easily perceived characteristics are favored over features mapped on characteristics that are more difficult to perceive. The purpose of the automatic fusion techniques presented in the sections to follow is to mimic the way that the human brain fuses information originating from multiple sources, so that there is no need to rely on human perception.

**Multi-objective optimization** Very often, unimodal dimensionality reduction methods are expressed as an optimization problem, where a suitable objective function is minimized. When many modalities are considered, many objective functions can be defined. *Multi-objective optimization* is a field of optimization dealing with problems having not one but many objectives to be simultaneously minimized [21] [149]. In multi-objective optimization, the multiple objectives are usually conflicting, so that there is no solution minimizing all of them simultaneously. This conflict is generally resolved either by scalarizing the objective functions, or by computing multiple solutions, instead of one, which represent different trade-offs among the objectives.

In scalarization techniques, the multiple objectives are combined in a single one, which is then minimized with traditional single-objective optimization methods. The most common scalarization method is minimizing a weighted sum of the objectives [21]. The weights of the sum represent a priori preferences about the different objectives. Another common method for scalarizing the objectives is the $\epsilon$-constraint method, where only one of the objectives is minimized, while the others are transformed to constraints [21]. Other scalarization methods rely on the definition of achievement functions.
that measure the distance of a solution from a reference one [150]. Scalarizing the objectives always involves the introduction of preferences for the multiple objectives. For example, in the weighted sum approach, the preferences are encoded in the weights of the objectives. In the \( \epsilon \)-constraint approach, each of the objectives which are transformed to constraints is assigned an \( \epsilon \) value, and is constrained to be below this value. The smaller the \( \epsilon \) value for an objective, the higher the preference for it. In the achievement function-based approaches, the preferences are encoded in the reference point used as the target. Using preferences for the multiple objectives is essential in order for a single final solution to be found. However, setting the values of these preferences automatically is not a trivial task and human users are often introduced to determine them.

Another class of multi-objective optimization methods do not consider any preferences at all. Instead of providing a single solution, they result in multiple solutions [21]. Such methods utilize the notion of dominance of a solution over another, meaning that a solution surpasses another in all objectives simultaneously. The result of optimization is a set of solutions that dominate all other solutions, but which do not dominate each other. This set is called the Pareto set and the solutions contained in it are all optimal for the specific problem, corresponding to different trade-offs among the objectives. Constructing the Pareto set is the most generic approach of handling multi-objective optimization problems, and contains scalarization techniques as its sub-cases. This has the advantage that all optimal trade-offs among the modalities are computed, even ones which cannot be discovered by scalarization techniques (see Fig. 4, later). On the other hand, the Pareto set consists of many solutions, from which one needs to be selected, which is not usually a trivial task.

Multi-objective optimization has been used extensively in the literature, mostly for solving problems emerging in economics and engineering, where multiple criteria need to be considered simultaneously. The work of [151] provides a survey of multi-objective optimization methods addressing the problem of portfolio optimization, where there is a need to select a portfolio with a large amount of profit, while at the same time reducing the risk. In [152], multi-objective optimization techniques are used in order to design a decision support system for supply chain management. Pareto sets of solutions are used to select e.g. strategies for fast deliveries, while at the
same time keeping the delivery cost low. In the field of manufacturing, the authors of [153] proposed a multi-objective approach to select milling parameters, in order to simultaneously achieve high milling quality, high production rate and low energy consumption. Multi-objective optimization has generally been used in several diverse fields. However, it has not been much used in machine learning applications and specifically for multimodal dimensionality reduction.

The Pareto set resulting from multi-objective optimization often consists of an infinite number of solutions, hence it is usually approximated by a finite set of representative ones. The Pareto set is commonly calculated using genetic algorithms, since the fact that they maintain a population of solutions, instead of a single one, at each generation, suits the goal of calculating multiple solutions [149] [154]. For the determination of the fitness function of the genetic process, different approaches have been followed [154], including using a weighted sum of the objectives with varying weights [155], alternating among the objectives [154] and using dominance relations [156].

There is a relation of multi-objective optimization to multi-task learning and related approaches, such as co-clustering and co-training, since all involve optimizing with multiple sources of information. However, the multi-objective approach differs from multi-task learning in that in multi-task learning the effort is to use the available sources of information cooperatively, in order to provide an optimal combined outcome, while in multi-objective optimization, the available modalities “compete” with each other, resulting in a set of all possible trade-offs among them. Existing multimodal approaches, such as using weighted sums and co-training can be used to calculate different trade-offs among the modalities, by altering, in each method, those parameters which correspond to the modality weights. However, there are limitations. In weighted sum-based methods, and in cases of non-convex Pareto sets, not all Pareto points can be discovered, even if all weight combinations are considered, as will also be shown later (Fig. 4). Co-training methods also have limitations. Co-training is related to the $\epsilon$-constraint scalarization method, where, at each iteration step, the modality to be considered for the optimization function and the modality to be considered for the constraints are alternated. According to [21], using the $\epsilon$-constraint method with varying values of the $\epsilon$ parameter results in different points of the Pareto set. All Pareto-optimal trade-offs can be discovered in this way,
however multiple runs of the algorithm are needed, which are also computationally expensive, since many single-objective optimization problems need to be solved [157].

The field of multi-objective optimization has not been used explicitly for multimodal learning, although the concepts of the two fields are relevant. In the research of this thesis, multi-objective optimization techniques are used for multimodal dimensionality reduction.

2.3 Summary and discussion

In this section, a comprehensive survey of the existing literature about the topic of multimodal dimensionality reduction has been presented. This topic combines the task of dimensionality reduction with that of multimodal fusion, in order to handle the problem of reducing the dimensionality of multimodal data. Thus, the survey was split into two parts, covering techniques for unimodal dimensionality reduction, and techniques for multimodal learning, with a focus on dimensionality reduction and related fields.

In the first part, several methods for unimodal dimensionality reduction were presented. Unimodal dimensionality reduction aims at reducing the dimensionality of a set of high-dimensional data, so that the low-dimensional representation preserves as much information of the original data as possible. Unimodal dimensionality reduction techniques were themselves split into two categories, namely global and local techniques. Global techniques are among the oldest approaches to dimensionality reduction. They consider the relationships of a data point to all the other data points of the dataset in order to produce the low-dimensional embedding. Their goal is to produce a low-dimensional projection space, which maintains as much of the variance and discrimination between the original points as possible. On the contrary, local methods consider small neighborhoods around the data points and aim at preserving the relationships of the data points only to their neighbouring points. The neighborhood relationships among the points are encoded in neighborhood graphs, which are utilized as guides for the calculation of the low-dimensional embedding. Global methods are often simpler and more computationally efficient than local methods, especially in handling new data, outside of the original training dataset. On the other hand, it is not that easy to handle data outside the training dataset with local methods, since usually the distances of the new data to all data of the
training dataset are needed. However, the advantage of local methods is that they can discover and unfold highly non-linear low-dimensional structure embedded in the high dimensional data.

In the second part, techniques for multimodal learning, with a focus to dimensionality reduction and related fields, were outlined. Multimodal fusion is the process by which information originating from multiple, often incomplete or unreliable, sources is combined, in a way that exploits the complementarities and redundancies of the information sources, in order to produce outcomes that are more robust and reliable than using each source independently. The multiple sources of information can be data having multiple views, namely multiple modalities. Multimodal fusion can be performed either at an early stage or at a late stage. Early fusion methods combine low-level features of the multimodal data before high-level decision-making processes are performed. Late fusion methods combine high-level decisions computed after performing decision-making processes on low-level features of each individual modality. Multimodal fusion has successfully been used in research areas as diverse as multi-sensor fusion, speech recognition, biometric identification, human-computer interaction and multimedia retrieval.

Most of the multimodal learning methods that are designed specifically for dimensionality reduction and related problems consider that the available data are in the form of multimodal objects, which are collections of data sharing a common semantic concept. In other words, a multimodal object contains multiple modalities, each of which may be a feature vector of a different type.

The existing literature on multimodal learning was split into three broad categories. In the first category, methods that combine the multiple modalities into a common representation have been examined. An early approach is to try to map the data of different modalities on a common space, so that further processing, such as clustering or retrieval can be performed on this common space, independently of the data modality. A large family of multimodal learning methods are based on the concepts of modality selection and modality weighting. In modality selection, the most informative of the available modalities are discovered and are combined in order to perform learning tasks using the combined modalities. Modality weighting works in a similar manner, by assigning weights to each modality, encoding its importance or discrimination ability for the specific problem or dataset. Multimodal
learning problems are formulated in the form of weighted combinations of features extracted from the individual modalities, such as weighted sums of feature vectors or of neighborhood graphs. The modality weights are learned so that the combination is as much informative and discriminative as possible. Multiple Kernel Learning is a significant approach employing this principle. In the second broad category, methods that find a solution by simultaneously learning in multiple modalities have been examined. An early approach is the co-training approach and its variants, where machine learning tasks, formalized as iterative processes, are performed alternatively at each of the available modalities, so that learning in one modality is assisted by information from the other modalities. Similar is the concept of co-clustering methods, where clustering of multimodal data is performed simultaneously on each of the available modalities, with one modality assisting the other. Multi-task learning techniques have also been used to simultaneously handle multiple learning tasks. Finally, in the third broad category, methods that handle multiple modalities without automatically fusing them are examined. Using glyphs for visualization is an early attempt to map map different modalities on different visual characteristics, in which case multimodal fusion is performed in the mind of the human observer. Another approach is multi-objective optimization, which has been used effectively in fields such as economics and engineering, but has not been utilized in the context of multimodal dimensionality reduction yet. In this approach, learning tasks for the individual modalities are formulated as unimodal optimization problems. The existence of multiple modalities leads to multiple optimization problems, which are handled by multi-objective optimization techniques. These techniques, instead of resulting in a single solution, result in a set of solutions representing optimal trade-offs among the modalities, from which one can be selected using pre-defined criteria, or by incorporating the human decision maker.

This revision of the recent literature in the field of multimodal dimensionality reduction has revealed two open challenges. First, multimodal fusion methods inherently attempt to fuse the available modalities. The process of fusing involves, either directly or indirectly, the consideration of the relative importance of one modality over the other modalities. This is mostly prominent in modality selection methods or methods based on weighted combinations of modalities. In such methods, some modalities are considered more
important than others and are selected or assigned larger weights. Even in methods such as co-training or co-clustering, where there is an alternation between the modalities, there is an underlying assignment of importance, by considering all modalities equally important. This consideration of the relative importance of the modalities in order to combine them renders the related methods non-trivial, since they first need to compute which modalities to select or how to distribute the weights. On the other hand, problems of multiple conflicting objectives have already been handled in areas such as finance and engineering, without considering, at least at the most part, any importance for the various objectives. Such problems have been handled by multi-objective optimization techniques, which, instead of arriving at a single solution, representing a single trade-off among the multiple objectives, produce a set of solutions to select from. The large computational part of these techniques is performed without assuming any relative importance among the modalities, but using only relations between different solutions that can be quantified independently of any importance assumption. Only after a minimal set of optimal trade-offs has been reached must a decision making process be utilized in order to select one among the different trade-offs. However, this decision is now performed from a limited set of solutions, and it can even be performed by a human decision maker, after exploring various of these trade-offs. Nevertheless, multi-objective optimization techniques have not been used in the context of multimodal dimensionality reduction, or related problems, yet, leaving an challenging open area for research.

The second challenge concerns methods that do consider the relative importance of modalities. As discussed in the literature survey, a large number of unimodal dimensionality reduction techniques can be collectively described as instances of the graph embedding framework of [23]. Extending this framework so that it can handle data of multiple modalities would allow any dimensionality reduction method that can be formulated in this framework to be extended to multiple modalities as well. This, in turn would allow the design and selection of multimodal dimensionality reduction techniques that work best for a particular type of dataset. Attempts have already been made to extend the graph embedding framework to multiple modalities, such as the work of [28] and its variants. However, while the graph embedding framework relies largely on the definition of the affinity and penalty matrices, [28] focuses mostly on the fusion of the kernel matrices to be used
in the framework, while a naively fused affinity matrix is used. The works of [27] and [127] are more promising attempts to fuse the affinity matrices, but they focus on classification and clustering. The very limited works on the direction of affinity matrix fusion that can be used for extending the graph embedding framework to multiple modalities leaves an open area for research.

In this thesis, two approaches for multimodal dimensionality reduction, namely Multi-objective dimensionality reduction and Multimodal Graph Embedding, are proposed. As depicted in Fig. 2, the Multi-objective DR approach belongs to the approaches that do not perform modality fusion, making use of multi-objective optimization techniques. On the other hand, the Multimodal Graph Embedding approach belongs to the approaches attempting to compute a common representation of the modalities, specifically using late fusion approaches, relying on the fusion of affinity graphs. These methods attempt to progress research in the two open areas described above, so as to fill the gaps in the literature and provide steps for further research in the area of multimodal dimensionality reduction.
3 Multimodal dimensionality reduction using multi-objective optimization

In this section, it is proposed that multimodal dimensionality reduction be formulated as a multi-objective optimization problem [21], where the goal is to simultaneously minimize a set of suitable cost functions, each corresponding to a single modality. Each cost function is defined so that its minimization leads to a mapping of the data on a low-dimensional space where the similarities and dissimilarities among the data are best preserved, while considering that the data only consist of the respective modality. For the definition of the unimodal cost functions, graph-based functions, inspired from graph aesthetic measures [59] [60], are utilized. Graph aesthetic measures provide an unsupervised way to evaluate the effectiveness of a low-dimensional mapping in revealing the structure of a graph. The multi-objective approach is generic enough to allow for any type of measures to be used for the definition of the multiple cost functions, so measures used in graph embedding [23] or Multidimensional Scaling [34] could also be used. Graph aesthetic measures were preferred in order for the final mapping to be more appealing to the human eye, since the experimental evaluation of the multi-objective approach was conducted in data visualization use cases.

Instead of providing a single solution, as existing fusion-based techniques do, multi-objective optimization typically results in a set of optimal solutions, namely the Pareto set, which represents different trade-offs among the various objectives, i.e. modalities. In fusion-based methods, a compromise is made among the various modalities, in order to produce a single solution. On the other hand, in multi-objective optimization, the whole range of possible optimal solutions is calculated, without employing any fusion scheme, thus maintaining the original information of the separate modalities. This range of solutions includes ones which cannot be discovered by fusion-based techniques, even by altering the fusion parameters. After determining the Pareto set, one solution can be selected according to user preferences, or the whole set can be presented to a human operator, in order to explore different mappings of the same dataset.

The remainder of this section is divided in three parts: The first part contains the proposed approach for solving the problem, based on multi-objective optimization. In the second part, a graph representation of the
data is presented, which facilitates the definition of objective functions. In the final part, two examples of objective functions are presented, which will be used in the experimental sections.

3.1 The multi-objective approach to dimensionality reduction

As mentioned in the introduction, the desired output of multimodal dimensionality reduction is a matrix $Y \in \mathbb{R}^{N \times d}$, where the $i$-th row represents the low-dimensional representation of multimodal object $O_i$. Hereby multimodal dimensionality reduction is approached as a multi-objective optimization problem, resulting in a set of solutions, instead of a single one. Given, for instance, a dataset where each object consists of an image and a sound modality, the problem hereby is to find a low-dimensional representation of the objects where the similarities of their images are apparent, while at the same time the similarities of their sounds are also apparent. These conflicting objectives are hereby formulated as appropriate cost functions and the goal is to minimize all of them simultaneously.

Let a unimodal case be considered, where the objects $O_i, i = 1 \ldots N$, consist of just one modality, denoted hereby as “1”. As far as this modality is concerned, the objects of the dataset have certain similarities and dissimilarities among them. The desired mapping $Y$ of the objects is one in which similar objects are placed close to each other and dissimilar ones are placed away from each other. In case each object belongs to one of a number of distinct classes, the desired mapping $Y$ would be one where clusters of similar objects are apparent. Let

$$J_1: \mathbb{R}^{N \times d} \rightarrow \mathbb{R}_{\geq 0},$$

be a cost function (objective function), evaluating the capability of a mapping $Y \in \mathbb{R}^{N \times d}$ for revealing the similarities and dissimilarities among the data, as far as only modality 1 is concerned. $\mathbb{R}_{\geq 0}$ is the set of non-negative real numbers. The smaller $J_1(Y)$ becomes, the more appropriate $Y$ is for revealing the data structure. In such a setting, the goal is to find the optimal mapping that minimizes $J_1(Y)$:

$$Y_{opt,1} = \arg \min_{Y \in \mathbb{R}^{N \times d}} J_1(Y).$$

67
If a different modality, denoted as “2”, is considered, then the similarities and dissimilarities among the objects will also be different, resulting in a different optimal mapping. For example, modality 1 could stand for the image modality, described by color feature vectors, while modality 2 for the sound modality, described by sound-related features. Since similarity with respect to color does not necessarily mean similarity with respect to sound spectrum, the optimal placements calculated using the two different objective functions are generally different.

Each modality has an associated objective function evaluating mappings as far as only this modality is concerned. However, in a multimodal setting, all modalities need to be taken into account at the same time. The goal is to find a mapping which simultaneously minimizes all objective functions:

\[
Y_{\text{opt, multimodal}} = \arg \min_{Y \in \mathbb{R}^{N \times d}} J(Y),
\]

\[
J(Y) = (J_1(Y), J_2(Y), \ldots, J_M(Y))
\]

where \(J(Y)\) is the vector of all objective functions and \(\min_{Y \in \mathbb{R}^{N \times d}} J(Y)\) means simultaneously minimizing all objectives. If such a mapping could be found, then the data structure as far as all modalities are concerned would be simultaneously preserved and all information from all modalities would be present in the low-dimensional space. However, as mentioned above, it is usually not possible to find such an ideal mapping, because each cost function is minimized at a different mapping.

In the literature, such problems of optimizing multiple conflicting objectives are handled effectively by multi-objective optimization approaches [21]. Multi-objective optimization methods result in a set of solutions, instead of a single one. These solutions are those for which no other solution can be found which dominates them in all objective functions simultaneously. However, as mentioned in Section 2.2.3, multi-objective optimization has mostly been used to solve problems in the fields of economics and engineering, such as increasing the profit while simultaneously reducing the risk, while it has not been much used in machine learning.

Herein, it is proposed that the problem of multimodal dimensionality reduction should be expressed in terms of multi-objective optimization, by considering the mapping \(Y\) as the optimization variable and the cost functions \(J_m, m = 1 \ldots M\), as the multiple objectives.
In order to calculate the most efficient solutions, there is a need to compare between different objective function vectors \( J(Y) \), resulting from different values of the variables \( Y \). Such comparisons are commonly performed using the notion of \textit{Pareto dominance}. A feasible solution dominates another one, if it has a smaller value for at least one objective and there is no objective for which it has a larger value. Between the two solutions, the dominant one is preferred, since it corresponds to better values for all objectives, without sacrificing any one of them. Formally, an objective vector \( J(Y_1) = (J_1(Y_1), J_2(Y_1), \ldots, J_M(Y_1)) \), resulting from variable \( Y_1 \), is said to dominate another objective vector \( J(Y_2) = (J_1(Y_2), J_2(Y_2), \ldots, J_M(Y_2)) \), resulting from variable \( Y_2 \), if

\[
J_m(Y_1) \leq J_m(Y_2), \; \forall m \in \{1, \ldots, M\}, \text{ and } \exists k \in \{1, \ldots, M\} : J_k(Y_1) < J_k(Y_2)
\]

Pareto dominance is denoted as \( J(Y_1) \prec J(Y_2) \), meaning that \( J(Y_1) \) dominates \( J(Y_2) \). Similarly, a variable \( Y_1 \) is said to dominate another variable \( Y_2 \) \( Y_1 \prec Y_2 \), if \( J(Y_1) \prec J(Y_2) \).

If \( Y_1 \prec Y_2 \), then \( Y_1 \) is a better solution than \( Y_2 \), since at least one of the objective function values for \( Y_1 \) is smaller than the respective value for \( Y_2 \), without any other objective of \( Y_1 \) having a larger value than the respective of \( Y_2 \). If two objective function vectors mutually do not dominate each other, they are said to be \textit{incomparable}, since there can be no impartial judgment as to which is better than the other. The goal of multi-objective optimization is to find the set of all solutions which dominate all other solutions, but are mutually incomparable. This set is called the Pareto set, and the corresponding values of the objective function vectors is called the Pareto front. An example Pareto front for a problem of two objectives is illustrated in Fig. 3. The gray-shaded area represents the set of all feasible solutions, while the bold border in the lower left of the feasible area is the Pareto front. Three example points are presented. Point \( Y_2 \) dominates \( Y_1 \) since both objectives have smaller values at \( Y_2 \). Similarly, \( Y_2 \) dominates all points within the hatched area. On the other hand, points \( Y_2 \) and \( Y_3 \) are incomparable, since none dominates the other. None of the solutions of the front dominates each other, since decreasing one objective leads to increasing the other one. The goal of multi-objective optimization is to find all those points which are not dominated by others, i.e. the Pareto front,
Figure 3: Example Pareto diagram illustrating the Pareto front for a problem of two objectives, $J_1$ and $J_2$. The gray area represents the set of all feasible solutions, while the bold border is the Pareto front. Solution $Y_2$ dominates $Y_1$, as well as all solutions within the hatched area. Solutions $Y_2$ and $Y_3$ are incomparable. An example Pareto diagram produced from a real application is depicted in Fig. 10, in the experimental section below.

in order to present the decision maker with a minimal set of solutions from which to select. As an example of a Pareto diagram produced from a real application, the reader is referred to Fig. 10, in the experimental section below.

It should hereby be stressed that the above multi-objective approach to dimensionality reduction is different and more generic than the common weighted sum-based scalarization approaches. In a weighted sum-based approach, the multiple objectives are combined in a single objective, using a weighted sum:

$$J(Y) = w_1 J_1(Y) + w_2 J_2(Y) + \ldots + w_M J_M(Y).$$

Solutions to the problem are thus found by minimizing the combined objective function $J(Y)$. By varying the weights $w_1, \ldots, w_M$, one can put more focus to one objective or the other, and thus compute a different final solution. As mentioned in [158], the solutions computed with this approach are indeed Pareto-optimal solutions, considering the notion of dominance described above, i.e. they lie on the Pareto front. However, not all Pareto-optimal solutions can be discovered this way, even if all the possible weight combinations are considered. In a case of two objectives, the weighted sum approach is geometrically interpreted by a line which is tangential to the
Figure 4: Illustration of solutions discovered by a weighted-sum method, using different weight combinations. Points in the concave part of the Pareto front are missed out.

Pareto Front [21], as depicted in Fig. 4. By using different weights for the objectives, different tangential points are found, representing different trade-offs between the objectives. However, this approach fails when the Pareto Front is non-convex, as the points in the concave part cannot be discovered. Thus, solutions in the concave part are missed out, although they may correspond to significant and useful solutions. This illustrates that the proposed approach, based on multi-objective optimization, is more general than commonly used weighted-sum-based methods and can discover more solutions that are suitable.

The Pareto front is commonly computed using genetic algorithms [149] [154], which maintain a population of solutions, instead of one, thus they fit with the multiple solution nature of multi-objective optimization. The computational bottlenecks of the Pareto calculation are the calculation of the fitness function and the environmental selection step. If dominance relations are used for the definition of the fitness function, as in [156], the computational complexity of the fitness function is \( O(Q^2 \log Q) \), \( Q = P + \bar{P} \), where \( P \) is the population size and \( \bar{P} \) is the size of the archive used in the genetic algorithm. The environmental selection step of the algorithm also has a complexity of \( O(Q^2 \log Q) \), resulting in a total complexity of \( O(Q^2 \log Q) \). The ultimate time complexity of the algorithm then depends on the complexity...
of the computation of the objective functions, which are used to compute the dominance relations.

Pareto diagrams such as the one in Fig. 3 are useful for illustrating the Pareto optimal solutions and the trade-offs among the various objectives. However, when more than two objectives are considered, higher-dimensional diagrams need to be constructed, which are difficult to depict graphically. It is also difficult for a human operator to select one of the Pareto-optimal solutions in such high-dimensional diagrams, although navigation in them could be assisted by one-dimensional navigation for each of the objectives, selecting smaller or larger values for it, while staying on the Pareto front. For purposes of simplicity and without loss of generality, in the examples to be presented in the following sections, two-dimensional Pareto diagrams will be considered.

Following the above formulation, objective functions need to be defined, one for each modality, evaluating the various mappings, so that the optimal mapping to be found is the one which simultaneously minimizes all objective functions. In the following section, graph representations of the data are used to define such objective functions.

3.2 Formulation using graphs

Using graphs to encode the pairwise relationships between objects is a common technique in dimensionality reduction. Although most commonly graphs have been used to consider neighbors around the objects and discover the low-dimensional manifold spanned by the data, hereby they are used in a different manner. Graphs are hereby constructed from the available data, in order to encode the data structure and allow this structure to be apparent in the low dimensional space. Graphs are used to allow the definition of objective functions, which, when minimized, lead to a such a placement of the vertices, within the low-dimensional space, that makes the graph structure apparent in an aesthetically pleasing way. Although the goal of aesthetically pleasing vertex positioning is borrowed from two-dimensional graph drawing problems, the same principles are hereby applied to spaces with more dimensions.

The set $O$ of multimodal objects is again considered. A distance function
$d_m$ is defined between two multimodal objects, with respect to modality $m$:

$$d_m : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}_{\geq 0}$$

$$d_m(O_1, O_2) = ||\mathbf{o}_{1,m} - \mathbf{o}_{2,m}||, \; m = 1 \ldots M,$$

where $||\mathbf{o}_{1,m} - \mathbf{o}_{2,m}||$ denotes the euclidean distance between feature vectors $\mathbf{o}_{1,m}$ and $\mathbf{o}_{2,m}$. This distance function is the distance between two multimodal objects, if just their $m$-modality components are considered.

Considering modality $m$ and the respective distance measure $d_m$, a complete weighted graph $G_m(\mathcal{O}, E_m)$, $E_m \in \mathcal{O} \times \mathcal{O}$, can be constructed for the set of multimodal objects $\mathcal{O}$, having the objects as its vertices and edges (set $E_m$) between every pair of them, weighted by the distances between the respective objects, according to $d_m$. In order for the graph structure to be less cluttered, the minimum spanning tree (MST) $T_m(\mathcal{O}, E'_m)$, $E'_m \subseteq E_m$, of the complete graph is calculated. The MST is able to reveal the similarities and dissimilarities among the objects, as objects that are similar to each other are connected through small paths on the tree [159]. Since there are $M$ modalities, $M$ minimum spanning trees are constructed, one for each modality. All $M$ trees have the same set of vertices (the multimodal objects), while the set of edges is different for each modality.

These MSTs can be used to define the objective functions $J_m$. The objective function for a specific modality is defined by first embedding the MST in the low-dimensional space, where the vertices are placed according to mapping $Y$, and then measuring the suitability of $Y$ in terms of how aesthetically pleasing the final positioning is, i.e. how easily the tree’s structure can be perceived. Multi-dimensional graph aesthetic measures can be used for this purpose, inspired from two-dimensional graph aesthetic measures, such as the ones in [59]. Examples of graph aesthetic measures are presented in Section 3.3 below.

### 3.3 Objective function examples

Any metric that quantifies the suitability of a vertex positioning for revealing the structure of a graph or a tree in an aesthetically pleasing way can be used to define the multiple objective functions. For the purposes of this thesis, the following two graph aesthetic measures are used.
Minimum angle objective The minimum angle aesthetic measure [59] is based on the idea that, in an aesthetically pleasing graph drawing, the angles between the subsequent edges originating from a vertex are equal. The metric measures the deviation of the edge angles of a graph drawing from the ideal angles. The minimum angle objective function is defined as:

$$J_{\text{angle},m}(p) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\theta_{i,\text{ideal}} - \theta_{i,\text{min}}}{\theta_{i,\text{ideal}}} \right|, \quad (9)$$

where $\theta_{i,\text{min}}$ is the actual minimum angle, according to placement $p$ and $\theta_{i,\text{ideal}}$ is the ideal (maximal) minimum angle of the edges connected to $o_i$. For a 2-dimensional mapping, $\theta_{i,\text{ideal}}$ is computed using the simple formula:

$$\theta_{i,\text{ideal}} = \frac{2\pi}{\text{degree}(o_i)}, \quad (10)$$

with $\text{degree}(o_i)$ being the number of edges connected to vertex $o_i$. For mappings in more dimensions, the computation of the ideal angle is not so straightforward and is out of the scope of this thesis. Nevertheless, in the experimental evaluation of Section 5.4, the use of the minimum angle objective will be restricted to two dimensions, so the above simple formula will be utilized.

The minimum angle objective function takes values in the $[0, 1]$ range, with 0 meaning that placement $p$ is aesthetically pleasing for modality $m$, according to the criterion used, while 1 means that the placement is not aesthetically pleasing. Thus, minimizing this objective for each modality would lead in an aesthetically pleasing placement of all unimodal trees.

The minimum angle objective will be used in the experiment with the artificial dataset of Section 5.4.1, in the experimental evaluation section below.

Potential objective In [60], it is stated that an aesthetically pleasing graph drawing is produced by considering the graph’s vertices as repelling charges and the edges as attractive springs attached to pairs of vertices. Starting at a random initial placement of the vertices, this dynamical system is let to run until convergence. The final result is an aesthetically pleasing and easily perceivable drawing of the graph, where vertices connected with lighter edges (smaller distances) are drawn close to each other.
The purpose of [60] was drawing the graphs on a 2-dimensional screen, thus the focus of the force-directed algorithm was on 2-dimensional mappings. The authors of [60] also examine the case of 3-dimensional graph drawings, with the ultimate purpose of projecting them again to two-dimensions. In the latter case, the quality of the drawing was highly dependent on the viewing angle, which should be selected by the user. However, this limitation of the final mapping to be 2-dimensional is imposed only if the purpose is the visualization of the graph on a screen. If the purpose is classification or clustering, where there is no need to present the graphs necessarily in two dimensions, higher-dimensional graph drawings would be sufficient, or even more efficient than two-dimensional ones.

Thus, hereby, based on this force-directed method, and extending it to more than two dimensions, an objective function can be defined for the drawing of a unimodal MST, \( T_m(\mathcal{O}, E_m) \), as the potential of the mechanical system of the vertices:

\[
J_{\text{potential},m}(\mathbf{Y}) = \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} \frac{q_i^2}{||\mathbf{Y}_i - \mathbf{Y}_j||} + \sum_{i,j, (\mathcal{O}_i,\mathcal{O}_j) \in E_m} k||\mathbf{Y}_i - \mathbf{Y}_j||^2. \quad (11)
\]

\( \mathbf{Y}_i \) is the \( i \)-th row of \( \mathbf{Y} \), which is a \( d \)-dimensional vector and corresponds to multimodal object \( \mathcal{O}_i \). The \( q \) parameter is the electric charge of the vertices, \( k \) is the spring constant and \( ||\mathbf{Y}_i - \mathbf{Y}_j|| \) denotes the euclidean distance between points \( i \) and \( j \) in the mapping \( \mathbf{Y} \).

The first term corresponds to the potential energy of the repelling electric forces, summing the potential at each vertex. The electric potential is calculated according to Coulomb’s law. Similarly, the second term corresponds to the potential energy of the attractive spring forces, according to Hooke’s law. The minimization of the electrical term tends to push the vertices far from each other, since the electrical force is repelling. On the other hand, the minimization of the spring-related term tends to pull the vertices that are connected by edges close to each other, since the spring force is attractive. The repelling electrical force is stronger when the vertices are close to each other, while the attractive spring force is stronger when the vertices are far from each other. Thus, minimizing the whole potential, i.e. the sum of the electrical and the spring-related terms, is guaranteed to converge to an equilibrium between the electrical and the spring forces, leading to a low-energy graph positioning in the low-dimensional space. Furthermore,
since the repelling forces are applied to all the vertices, while the attractive ones only to those connected by edges, the minimization of the potential “unfolds” the graph and leads to a positioning where the structure of the graph, or hereby the tree, is apparent. However, this minimization may lead to local minima. For instance, in the case of trees, as hereby, where the final positioning should contain no edge crossings, the minimization of the potential may lead to placements with edge crossings, which, in order to be resolved, would require the system to pass through a state of higher energy, such as vertices not connected by an edge passing very close to each other.

As an illustration of how the potential objective proceeds, Fig. 5 depicts the case of a small three-dimensional dataset that is mapped on a two-dimensional space. In this dataset, each item corresponds to a different color, encoded as a three-dimensional vector. A complete graph is formed from these data, using the euclidean distance between the three-dimensional colors for the edge weighting. The vertices are initially placed on random locations on the two-dimensional plane, as depicted in Fig. 5(a). The minimum spanning tree of this graph is depicted in Fig. 5(b), where it can be observed that similar colors are connected in the MST. By minimizing the potential objective function for this tree, the positioning depicted in Fig. 5(c) is obtained. The vertices are repelled from each other and only those that are connected by an edge are kept close. This leads to an aesthetically pleasing positioning of the tree, where its structure is easily apparent and the vertex similarity is preserved.

The computational complexity of the potential objective function is $O(N^2)$, which affects the time needed for the Pareto-optimal solutions to be calculated. In a computer with an Intel i7 processor, the execution of the multi-objective approach for about 5000 objects needs about 10 seconds to stabilize.

The potential objective will be used in the comparison of the multi-objective approach to the state-of-the-art, using real-world datasets, presented in Section 5.4.3, in the experimental evaluation section below.

3.4 Summary

In the approach presented in this section, multimodal dimensionality reduction was handled as a multi-objective optimization problem. Using graph representations of the data, unimodal objective functions were defined for
Figure 5: Example of the potential objective minimization. (a) The objects and their similarities form a complete graph. The graph’s vertices are placed at random positions. (b) The minimum spanning tree of the graph is calculated. (c) By minimizing the potential objective, the vertices are moved so that the tree’s structure is apparent.

each modality, quantifying the suitability of a particular mapping of the multimodal objects for revealing the similarities and dissimilarities among them. The unimodal objective functions were based on existing graph aesthetic measures of the literature, which were used in a novel manner, as the multiple objectives of a multi-objective optimization approach for dimensionality reduction. Multi-objective optimization techniques were used to simultaneously optimize the objectives of all modalities, producing a set of Pareto-optimal solutions, which represent different trade-offs among the various modalities.

The fact that a set of optimal solutions, instead of a single one, is provided, is an advantage of the proposed method over fusion-based methods. The solutions of the Pareto set are calculated in an objective manner, by considering only the available data and some suitable objective functions and by making no assumption as to how the modalities should be fused. The Pareto-optimal solutions are objectively the best solutions, according to the selected objective functions. Of course, the task of dimensionality reduction is to compute a single data mapping, so a decision regarding which of the available solutions should be presented must be taken at some point. However, in the proposed method, this decision is postponed as a last step before the presentation of the results to a human operator. The search space for the final solution is reduced to the Pareto front, which is usually approximated by a set of discrete points. Hence the final solution is selected from a limited
set of optimal solutions. The selection of one of the solutions of the Pareto front can be accomplished either automatically, using some predefined preferences, or interactively, by employing the human user in the procedure. Such a selection strategy involves encoding the user preferences as a point on the Pareto diagram and selecting the solution that is closer to the preference point. The selection of a solution is generally out of the scope of this thesis, whose goal is to investigate the applicability of using multi-objective optimization for multimodal dimensionality reduction and to present that it can discover solutions that existing methods cannot. However, the reader can find an example selection strategy in the accessibility-related application of Section 6.1. Further developing and fine-tuning such a strategy for multimodal dimensionality reduction is a direction for future research.
4 Multimodal dimensionality reduction using Multimodal Graph Embedding

In this section, the proposed Multimodal Graph Embedding (MGE) dimensionality reduction method is described. MGE is an extension of the graph embedding framework of [23]. In graph embedding, there is a “target” affinity matrix, which guides the embedding procedure, so that in the final embedding, the affinities are preserved. This affinity matrix can be constructed in a supervised, a semi-supervised, or an unsupervised manner. In the supervised case, the construction of the affinity matrix is straightforward, as the affinities are determined by external knowledge, coming from supervision. In the semi-supervised or unsupervised cases, the construction of the affinity matrix is also based on the characteristics of the data themselves, such as neighborhoods or distances among them. This difference between purely supervised approaches and those utilizing the data themselves has an impact on how graph embedding is handled in case of multiple modalities. Extending the supervised case to multiple modalities does not have any impact on how the affinity matrix is constructed, since it is again determined solely by supervision. However, in the semi-supervised and unsupervised cases, the affinity construction is affected, since now there are multiple notions of neighborhood or distances among the data. This suggests that a fusion strategy needs to be adopted, in order to combine the multiple representations into a single affinity matrix. The proposed Multimodal Graph Embedding approach is an attempt towards this direction and will be considered in the following only for semi-supervised or unsupervised cases.

In MGE, there is an effort to exploit the multiple available modalities, by focusing on the fusion of the unimodal affinity matrices via a weighted sum. This fusion is meaningful only in cases where the construction of the affinity matrices involves the use of the unlabeled multimodal data. Thus, the focus hereby is on affinity matrices for unsupervised or semi-supervised dimensionality reduction methods. The weights of the affinity matrix sum are calculated adaptively, using the available unimodal data, so that the resulting multimodal affinity matrix is as consistent as possible. In order to assess the graph consistency, a cost function based on a rational assumption concerning the similarities among the objects in the graph is introduced. By focusing on the affinity matrix fusion, the optimal “target” for the embedding
is calculated, so that the dimensionality reduction methods used afterwards lead to more accurate results. Furthermore, by fusing the modalities at the stage of affinity matrix calculation, any formulation of graph embedding, i.e., direct, linearization or kernelization, can be used afterwards, so that other dimensionality reduction methods, apart from the ones based on kernels, as in the MKL-DR method [28], can be boosted by the combination of modalities.

In this section, the proposed MGE method is detailed, by describing the affinity matrix fusion process, followed by the approach for the automatic determination of the fusion weights. MGE is related to the unimodal graph embedding and the Multiple Kernel Learning frameworks. The reader is referred to Sections 2.1.2 and 2.2.1 for background information about unimodal graph embedding and Multiple Kernel Learning, respectively.

4.1 Affinity matrix fusion

The setting of the problem is similar to the unimodal graph embedding setting described in Section 2.1.2. However, hereby there are multiple affinity matrices \( W_1, W_2, \ldots, W_M, \) where \( M \) is the number of modalities. The affinity matrices are constructed from the intrinsic similarities of the data, either in combination with supervision, in semi-supervised approaches, or not, in unsupervised approaches. Different dimensionality reduction techniques construct the affinity matrices differently. It is hereby assumed that the distinct modalities are not random in nature, but convey at least some amount of information about the semantic relationships among the objects, so that the multiple affinity matrices are not completely different from each other, but instead have some amount of redundancy. Although this assumption does not generally hold for any set of data, it is necessary for multimodal learning that the multiple modalities are different views of the same hidden semantic information, which needs to be uncovered. In this respect, there is some correlation among the multiple modalities, which is exploited in multimodal learning. If the modalities were completely random and independent in nature, there would be no meaningful semantic information to uncover.

A weight \( b_m \in [0, 1] \) is introduced for each modality. The modality weights are grouped in a vector \( \mathbf{b} = (b_1, b_2, \ldots, b_M)^T \). The multimodal affinity matrix \( W_{\mathbf{b}} \) is then defined as the weighted sum of the unimodal
Table 2: Various formulations of multimodal graph embedding.

<table>
<thead>
<tr>
<th>Formulation</th>
<th>Minimization problem</th>
</tr>
</thead>
</table>
| Direct       | \[
\begin{align*}
\min_Y & \quad \operatorname{tr}(Y^T L_b Y) \\
\text{s.t.} & \quad \operatorname{tr}(Y^T B_b Y) = 1
\end{align*}
\] (13) |
| Linearization| \[
\begin{align*}
\min_V & \quad \operatorname{tr}(V^T X L_b X^T V) \\
\text{s.t.} & \quad \operatorname{tr}(V^T X B_b X^T V) = 1
\end{align*}
\] (14) |
| Kernelization| \[
\begin{align*}
\min_A & \quad \operatorname{tr}(A^T K L_b K A) \\
\text{s.t.} & \quad \operatorname{tr}(A^T K B_b K A) = 1
\end{align*}
\] (15) |

affinity matrices $W_m$:

$$W_b = \sum_{m=1}^{M} b_m W_m, \quad \sum_{m=1}^{M} b_m = 1. \quad (12)$$

The subscript $b$ in $W_b$ indicates that this unified kernel matrix results from a weighted sum of the $M$ kernel matrices, using the specific weight configuration $b = [b_1 \ b_2 \ldots \ b_M]^T$. The multimodal Laplacian $L_b$, the diagonal $D_b$, and the constraint $B_b$ matrices are similarly calculated as weighted sums of the corresponding unimodal ones.

The optimization problem for the hereby proposed multimodal graph embedding framework is an extension of the graph embedding framework of Table 1, by replacing the Laplacian and constraint matrices with their multimodal counterparts, as presented in Table 2.

The proposed technique is independent of the graph embedding dimensionality reduction method used, so it could be applied with any method of the direct, linearization or kernelization formulations. The kernelization methods do not require the data themselves as input, but only the kernel matrix among them. Thus they have the advantage of being able to handle many different types of data, even data that cannot be represented by feature vectors. When multiple modalities are present, multiple kernels exist. Instead of arbitrarily combining the multiple kernels, the Multiple Kernel Learning (MKL) method of [28] was selected, precisely because it also han-
dles multiple kernels by automatically calculating the kernel weights. Within
the context of the proposed scheme, the MKL method is extended as:

$$\min_{A, b, \beta} \text{trace}(A^T K_{\beta} \mathbf{L}_b K_{\beta} A)$$

subject to \( \text{trace}(A^T K_{\beta} \mathbf{B}_b K_{\beta} A) = 1 \) \hfill (16)

The difference between Eq. (16) and the original MKL-DR framework
of [28] (Eq. (7)), is that modality weights are introduced to the Laplacian
and constraint matrices \( \mathbf{L}_b \) and \( \mathbf{B}_b \). Thus, there are two sets of weights, \( b \) and \( \beta \), which fulfill different purposes. Using the \( \beta \) weights, of the MKL-DR
framework, a multimodal kernel matrix \( K \) is formed, as

$$K = \sum_{m=1}^{M} \beta_m K_m.$$ 

This resulting kernel matrix is the one for which the transformation matrix
\( A \) is “easier” to calculate, i.e. it results to the least cost, while trying to
conform to the input affinity matrix \( W_b \). On the other hand, the \( b \) weights
are used to calculate the multimodal affinity matrix \( W_b \). This is the optimal
affinity matrix, which, if used in a dimensionality reduction framework, leads
to the optimal dimensionality reduction for the input data. Using the MKL-
DR framework to conform to the multimodal affinity matrix calculated with
the \( W_b \) weighs, is expected to result in better performance, compared to
the average matrix used in [28].

As an illustration of the effect of the \( \beta \) and \( b \) weights, Fig. 6 depicts the
results of multiple combinations of the weights for an object recognition task,
using the image dataset described in Section 5.1.2. For the sake of illustration,
two modalities are considered, namely the SIFT and GIST descriptors,
described in Section 5.2.1, thus the weights are vectors of two elements. The
elements of both the \( \beta \) and \( b \) are altered from 0 to 1, with a step of 0.1,
while constraining their sum to be 1. Thus, a weight of 0 means that all the
weight is given to the GIST modality, while a weight of 1 that all the weight
is given to the SIFT modality. This convention is used in both the \( \beta \) and
the \( b \) weights. For each weight combination, the performance of the MGE
method for an object recognition task is measured in terms of recognition
accuracy. It can be observed that the recognition rate depends highly on the
selection of the MKL-DR \( \beta \) kernel weights, with the highest scores achieved
when giving a weight of 0.8 to the SIFT modality. Using either modality
Figure 6: Illustration of multiple combinations of modality weights for an object recognition task. Two modalities are considered, namely SIFT and GIST. The weights are altered between the two modalities, for the two sets of weights considered, i.e., $\beta$ and the $b$. For each weight combination, the object recognition rate after the application of the MGE approach to an image dataset is measured. The highest score is achieved at about 0.8 SIFT for the $\beta$ weights and 0.7 SIFT for the $b$ weights. This suggests that the average affinity matrix used by MKL-DR does not lead to the optimal point, which lies at a different point than the average.

on its own does not lead to high performance. However, performance also depends on the selection of the $b$ affinity matrix weights. The optimal point of the highest score lies at around 0.7 SIFT weight. This means that a specific combination, other than the mere average, of the affinity matrix weights leads to supreme performance. The purpose of the proposed MGE approach is to discover such an optimal point. The two sets of weights act as two different “dimensions”. The $b$ weights determine the most consistent target affinity matrix, while the $\beta$ weights determine the most appropriate kernel combination in order to conform to the target affinity matrix.

### 4.2 Adaptive determination of modality weights

In the Multiple Kernel Learning framework of [28], in order to merge the multiple available affinity and penalty matrices, average affinity and penalty matrices were used. In this case, the weight vector $b$ is:

$$b = (b_1, b_2, \ldots, b_M)^T, \ b_i = \frac{1}{M}.$$  

In this work, instead, an optimal weight vector $b$ is searched for, which
Figure 7: Illustration of the notion of graph inconsistency as the existence of triangles. (a) The graph of one modality (inconsistent). (b) The graph of a second modality (inconsistent). (c) The multimodal graph, as a weighted combination of the two previous graphs (more consistent).

should lead to a more accurate low-dimensional embedding. Intuitively, in order for an affinity matrix to be suitable for embedding, it should describe the similarities and dissimilarities among the data as consistently as possible. This notion of graph consistency is hereby introduced. The intuition behind it is based on the assumption that the input data are semantically organized into different classes. Thus, the ideal affinity graph would connect objects of the same class, while not connecting objects of different classes. If the data were sorted by their class labels, this ideal matrix would be block-diagonal in nature.

In the ideal case, if two vertices are connected to each other through a path passing from a single other vertex, e.g. as the configuration of Fig. 7a, then the two vertices should also be connected directly via an edge between them. In other words, if three nodes of the graph are semantically similar to each other, there should be edges between all three pairs of them. If one of these edges is missing, this constitutes an inconsistency. For instance, the configurations of Figs. 7a and 7b are inconsistent. If the graphs of Figs. 7a and 7b are considered as the unimodal graphs of two modalities, then assigning weights to them and merging them leads to a configuration similar to Fig. 7c. In this figure, the more bright an edge is depicted, the smaller its affinity value. Edges BC and AC have intermediate affinity values, since they do not exist in both unimodal graphs. However, the configuration of Fig. 7c is more consistent than the other two, since all three edges appear.

A different way to express this notion of consistency is that in any triangle
formed by three vertices, the affinity values of the three edges should be similar. The graph of Fig. 8a is inconsistent, since edges AB and AC have very different values, while in an ideal situation, it would be expected that all three vertices are similar in the same way, thus the edges have equal values. The graph of Fig. 8b, with all edges having equal values, is consistent.

Figure 8: Illustration of the notion of graph inconsistency as triangles of equally weighted edges. (a) An inconsistent setting, due to much different edge weights. (b) A consistent setting. The weights of all edges are equal to each other.

This description of consistency allows for a formal way to quantify the inconsistency of the whole graph. Let $E^*$ be the set of all edges appearing in any of the unimodal graphs:

$$E^* = \{\{i, j\}, i, j \in \{1 \ldots N\} | \exists m \in \{1 \ldots M\} : w_{m,ij} > 0\}.$$

Then, the following measure of graph inconsistency is hereby proposed:

$$f(b) = \sum_{\{i,j\} \in E^*} \sum_{k=1}^N (w_{b,ik} - w_{b,jk})^2,$$  \hspace{1cm} (17)

where $w_{b,ij}$ is the $(i,j)$ element of the final affinity matrix $W_b$, which results by merging the unimodal affinity matrices with weights $b$. Intuitively, each vertex pair $\{i,j\}$ that is connected by an edge in the final affinity matrix $W_b$ is considered. For each vertex pair, every other vertex $k$ of the graph is considered and the values of the edges connecting $k$ to the two vertices of the pair $\{i,j\}$ are found. If these values are equal, the $k$ vertex is consistently connected to both the vertices of the pair $\{i,j\}$, while if they are not equal, this means that the $k$ vertex is “more connected” to one of the vertices of the
pair than the other, which constitutes an inconsistency. The difference between these values, squared, contributes to the inconsistency measure, which is the sum of these differences over all pairs \( \{i, j\} \) and vertices \( k \). Thus, the higher the value of the inconsistency measure, the larger the inconsistencies in triplets of vertices.

Let also \( w_{ij} \) be the vector containing the \( (i, j) \) values of all the unimodal affinity matrices:

\[
w_{ij} = (w_{1,ij}, w_{2,ij}, \ldots, w_{M,ij})^T.
\]

Then \( w_{b,ij} \) can be written as

\[
w_{b,ij} = \sum_{m=1}^{M} b_m w_{m,ij} = b^T w_{ij}.
\]

Eq. (17) can thus be written as:

\[
f(b) = \sum_{\{i, j\} \in E} \sum_{k=1}^{N} (b^T w_{ik} - b^T w_{jk})^2.
\]

In the example of Fig. 8, \( f(b) \) measures the squared difference between edges AB and AC, summed with the differences for any other pair of edges originating at a vertex and ending at another edge, so that the configuration of Fig. 8a has a higher value than the configuration of Fig. 8b, i.e. the former is less consistent than the latter.

Minimizing this inconsistency measure favors matrices approximating the ideal affinity matrix, which, for clarity, would be block-diagonal in structure, if the data were sorted by their class labels. In this respect, the optimal weight configuration \( b_{opt} \) leading to the most consistent final graph can be found by minimizing the objective function \( f(b) \):

\[
b_{opt} = \arg \min_{b \in \mathbb{R}^M} f(b)
\]

subject to \( b_m \geq 0, \sum_{m=1}^{M} b_m = 1, \ m = 1 \ldots M. \)

The first constraint ensures that the modality weights are not negative. The second constraint, forcing the modality weights to sum to 1, has been introduced in order to avoid the trivial solution of all weights being zero. As a note, a constant affinity matrix, connecting any object to all other objects, is
also a solution to the above minimization problem. However, a constant matrix cannot result from the weighted sum of the unimodal affinity matrices, since it is assumed that the affinity matrices are not completely uncorrelated to each other.

The objective function of Eq. (19) can be rewritten as:

\[
 f(b) = \sum_{(i,j) \in E^*} \sum_{k=1}^{N} (b^T(w_{ik} - w_{jk}))^2 \Rightarrow 
\]

\[
 f(b) = \sum_{(i,j) \in E^*} \sum_{k=1}^{N} b^T(w_{ik} - w_{jk})(w_{ik} - w_{jk})^T b = b^T Q b,
\]

where

\[
 Q = \sum_{(i,j) \in E^*} \sum_{k=1}^{N} (w_{ik} - w_{jk})(w_{ik} - w_{jk})^T. 
\]

Thus, the optimization problem of Eq. (20) can be written as:

\[
 b_{opt} = \arg \min_{b \in \mathbb{R}^M} b^T Q b 
\]

subject to \( b \geq 0_M, \ 1_M^T b = 1 \),

where \( 0_M \) is the column vector of size \( M \) with all its elements being 0, and \( 1_M \) is the column vector of size \( M \) with all its elements being 1. The optimization problem of Eq. (22) is a quadratic optimization problem and can be solved efficiently using existing solvers. Regarding the computational complexity of the weight calculation, quadratic optimization is generally NP-hard. However, the size of the above problem is small, since the \( Q \) matrix is of size \( M \times M \), and \( M \), i.e. the number of modalities, is usually small, e.g. less than 10. The bottleneck is in the computation of matrix \( Q \), which is \( O(N^3) \). The application of graph embedding dimensionality reduction techniques also involves solving an eigenvalue problem, which also has \( O(N^3) \) complexity. The total time needed for the weights calculation in an Intel i7 processor, for 600 objects, is about 20 seconds. This is a rather large time, the reduction of which is a consideration for future improvements.

As an example of the optimization results, Fig. 9 illustrates a case of 5 modalities, using an image dataset, described later, in Section 5.1.2, and using the different image descriptors, described in Section 5.2.1 as modalities. The affinity matrix for each modality is depicted, as a gray-scale image, with
brighter pixels denoting larger affinity values. For illustration purposes, the data are ordered according to the ground truth labels, so that the data classes are square blocks of equal size on the main diagonal. The multimodal affinity matrix is also depicted. This matrix results as a weighted sum of the 5 unimodal affinity matrices, using the optimal weights $b$, calculated as the solution to the optimization problem of Eq. (20).

It is apparent that the affinity matrix of the SIFT modality is close to the ideal affinity matrix, however there are several points missing from the main diagonal, while there is also noise in the areas outside the main diagonal. This means that there are edges in the graph connecting data which are semantically dissimilar. In the PHOG and GIST modalities, the clusters in the main diagonal are more clearly outlined than in SIFT, but there is much noise in the areas outside the diagonal. Finally, the GB-DIST and GB modalities are quite noisy, with little resemblance to the ideal affinity matrix. The multimodal affinity matrix manages to gather the best characteristics of all modalities, combining the small amount of noise of SIFT and the clearer clusters in the diagonals of PHOG and GIST. The multimodal matrix is thus smoother and much closer to the ideal matrix, than any of the modalities. The modality weights resulting from the optimization and used for the construction of the multimodal matrix are presented in Table 3. The values of the weights mirror the importance of the modalities, i.e. how close they are to the ideal case. Larger weight is given to the SIFT modality, which is the most accurate, while smaller weights are given to the other modalities, with GB receiving the smallest weight.

Overall, given the available unimodal affinity matrices $W_m$, the minimization problem of Eq. (20) is solved and an optimal weight vector $b_{opt}$ is calculated. Then the final affinity and penalty graphs are calculated as a weighted sum of the unimodal affinity matrices, according to Eq. (12), and similarly the constraint matrix, using vector $b_{opt}$ as the weights. The resulting affinity and penalty matrices can then be used in a state-of-the-art dimensionality reduction method, in order to map the multimodal data to a unified low-dimensional space. Hereby, the resulting matrices are used within the multiple kernel learning framework of [28], for comparison, so the
Figure 9: Example of the result modality weights optimization. Five unimodal matrices are depicted as gray-scale images, with the brighter colors denoting larger affinity values. The affinity matrices are of various accuracies. These matrices are provided as input to the optimization of Eq. (20) and the optimal weights are listed in Table 3. Using these weights, the multimodal affinity matrix (last) is calculated, which is closer to the ideal case than any of the unimodal matrices.

The following optimization problem is finally formed:

\[
\begin{align*}
\min_{A, \beta} & \quad \text{trace}(A^T K_\beta L_b K_\beta A) \\
\text{subject to} & \quad \text{trace}(A^T K_\beta B_b K_\beta A) = 1
\end{align*}
\] (23)

Using the resulting \( \beta \) weights, the multimodal kernel matrix \( K_\beta \) is formed:

\[
K_\beta = \sum_{m=1}^{M} \beta_m K_m
\] (24)

Then, the output points can be calculated using the multimodal kernel matrix \( K_\beta \) and the transformation matrix \( A \), as:

\[
Y = A^T K_\beta.
\] (25)
Algorithm 1 Pseudocode of the Multimodal Graph Embedding procedure.

**Input:** Unimodal affinity matrices $W_m$, unimodal penalty matrices $B_m$, unimodal kernel matrices $K_m$.

**Output:** Multimodal kernel matrix $K$, transformation matrix $A$.

1. Form unimodal affinities vectors $w_{ij}$, as in Eq. (18).
2. Solve the optimization problem of Eq. (20), using the objective function of Eq. (19), to find $b_{opt}$.
3. Use $b_{opt}$ to calculate the multimodal affinity matrix $W_b$ from Eq. (12) and similarly the multimodal constraint matrix $B_b$.
4. Calculate the Laplacian matrix $L_b$.
5. Use $L_b$ and $B_b$ in the Multiple Kernel Learning framework of Eq. (23), to calculate the multimodal kernel matrix $K_\beta$ and the transformation matrix $A$.
6. Use $K_\beta$ and $A$ to map training or testing data to low-dimensional points, according to Eq. (25).

The overall procedure is summarized in Alg. 1. As outlined above, the computational complexity of the MGE algorithm is $O(N^3)$. On a computer with an Intel i7 processor, the time needed for the algorithm to run with an input of 500 objects, is about 20 seconds.

It should be noted that the optimization with respect to $b$ is separated from the optimization with respect to the weights $\beta$ of the MKL framework. These two weight vectors could be learned together in the optimization problem of Eq. (16), or, in general, the modality weights $b$ could be learned together with the parameters of any dimensionality reduction method used. Such an approach could result in more accurate results, tailored for the specific dimensionality reduction method used. However, this would limit the application of the MGE method to the MKL algorithm, or to any specific dimensionality reduction method. At the same time, the algorithm of the DR method used would need modifications in order to include the modality weights. On the other hand, calculating the $b$ weights separately allows the application of MGE with any out-of-the-box graph embedding dimensionality reduction method, which facilitates the extension of unimodal DR methods to multiple modalities.
Table 3: Optimal modality weights for the construction of the multimodal matrix of the example of Fig. 9.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>0.5870</td>
</tr>
<tr>
<td>PHOG</td>
<td>0.1498</td>
</tr>
<tr>
<td>GIST</td>
<td>0.1234</td>
</tr>
<tr>
<td>GB-DIST</td>
<td>0.0753</td>
</tr>
<tr>
<td>GB</td>
<td>0.0646</td>
</tr>
</tbody>
</table>

4.3 Summary

In this section, an extension of the graph embedding framework for dimensionality reduction has been presented, which utilizes the multiple modalities available for the data. The multiple modalities are described as a set of unimodal neighborhood graphs and their affinity matrices, constructed from the unimodal data.

The method proceeds by constructing a multimodal affinity matrix, as a weighted sum of the unimodal affinity matrices. The weights of the sum are calculated based on a rational notion of consistency, regarding the relations between similar objects, as described in Section 4.2. The multimodal affinity matrix can then be used with any dimensionality reduction method which can be described within the graph embedding framework. The result of dimensionality reduction is a mapping of the high-dimensional data to a unified low-dimensional space, where supervised and unsupervised learning tasks, such as object recognition, clustering and visualization, can be performed with increased accuracy.

The weights of the sum of the unimodal affinity matrices for the construction of the multimodal one are determined as the solution to an optimization problem, which minimizes an introduced measure of graph consistency. Hereby the notion of graph consistency means that two graph vertices indirectly connected through another vertex, should also be directly connected with an edge. Thus the optimal modality weights are determined as those for which the resulting multimodal graph has the least inconsistency.
5 Experimental evaluation

The multimodal dimensionality reduction approaches proposed in the previous sections, i.e. the multi-objective dimensionality reduction and the multimodal graph embedding, can be applied to any kind of multimodal data, provided that vectorial descriptors can be extracted from them, or a certain notion of distance can be defined between data objects. The proposed approaches have been experimentally evaluated in a number of diverse cases and for various machine learning tasks, in particular classification, clustering and visualization. The evaluation of the proposed multimodal dimensionality reduction approaches has been performed in terms of how accurate the results of these learning tasks are, when the outputted low-dimensional data are used. This is an indirect way to evaluate the information loss between the multiple high-dimensional input spaces and the low-dimensional output space after multimodal fusion and dimensionality reduction. Instead of measuring the amount of information that is lost, the characteristic that is measured is whether the low-dimensional data manage to maintain the important information that is necessary to accurately perform the various machine learning tasks. Information loss measures could also possibly be used in order to directly compare the amount of information contained in the original and final spaces. However, such an approach could be misleading, since a lot of information contained in the original space may be irrelevant to the data semantics. The evaluation approach followed hereby considers the problem from the application point of view, where the semantic information is most important, and evaluates whether the low-dimensional data preserve these semantics.

After the description of the used datasets, descriptors and evaluation measures, in Sections 5.1, 5.2 and 5.3, the results of the experimental evaluation are presented in Sections 5.4 and 5.5.

5.1 Datasets

In the context of this thesis, a multimodal object is any abstract object which can be considered as consisting of multiple representations, of different kinds. In this respect, examples of multimodal objects could be a circular disk, consisting of a color modality and a size modality, an image, described using multiple image descriptors, such as SIFT and PHOG, and a collection
of multimedia items, containing an image, a sound and a video, describing together a specific semantic concept. In order to evaluate the proposed methods in various multimodal settings, a number of diverse datasets have been used. These datasets are described in the following.

5.1.1 Artificial multimodal dataset

A set of $N = 7$ multimodal objects is considered, each consisting of $M = 2$ modalities. The dataset is artificially generated and its purpose is to illustrate cases of non-convex Pareto fronts, where the multi-objective dimensionality reduction method discovers more solutions than weighted sum-based methods. The two modalities are herein considered to be the size and the color of the objects. Here the objects are considered to be circles with various radii and colors. The size modality describes the radius of the object and hence is described by a 1-dimensional vector, while the color modality is described by a 3-dimensional vector, containing the RGB color components. The values of the vectors describing the objects in the dataset are generated randomly, by a uniform distribution in the $[0, 1]$ range.

5.1.2 Image dataset

The image dataset used for the experiments on multimodal dimensionality reduction is a subset of the Caltech-101 dataset [160]. The Caltech-101 dataset consists of images of various objects split into 101 categories, with an additional category of background images. Each category contains about 40 to 800 images. Each image is about $300 \times 200$ pixels and the subject of each image is centered in most cases.

Following the unsupervised image clustering setting of [28], a subset of the Caltech-101 dataset is used, by selecting 10 image categories, namely brain, car side, faces, garfield, leopards, pagoda, snoopy, stop sign, windsor chair, yin yang. From each category, 30 images are randomly selected, and a final dataset of 300 images is formed, which will be used for the experiments. In order for the images to have similar sizes, they are resized to about 60000 pixels, while preserving their aspect ratio.

93
5.1.3 Image-sound dataset

Apart from the above image dataset, another multimodal dataset of objects consisting of an image and an associated sound has also been used in another set of experiments. In the literature, there are only a few benchmark multimodal datasets of images and sounds, and these are mostly limited to biometric applications, such as [161], for biometric identification. In the literature regarding multimodal dimensionality reduction, clustering, etc., there are no benchmark multimodal multimedia datasets, which leads the authors to creating ad hoc multimodal datasets, such as those used in [69] and [162]. In this thesis, a similar approach has been followed and the image-sound dataset has been created by combining an image dataset and a sound dataset.

For the images, a subset of the Caltech-UCSD Birds-200-2011 dataset [163] is used. This dataset consists of 11788 images of birds, split into 200 categories. A subset of the dataset is used, by selecting 10 categories, namely yellow-headed blackbird, indigo bunting, red-faced cormorant, vermilion flycatcher, frigatebird, rose-breasted grosbeak, California gull, mallard, hooded merganser and white pelican. From each category, 50 images are taken, resulting in a final dataset of 500 images. Segmentation masks available in the Caltech-UCSD Birds-200-2011 dataset have been used to remove the background of the images for the extraction of color descriptors.

In order to associate each image to a sound, sound files related to the above bird species have been downloaded from the Xeno-Canto online database [164], which contains recordings of bird sounds. A sound file has been downloaded for each of the 500 images of the above described image dataset. Since many sound files are long in duration and contain irrelevant sounds, the sounds have been cropped to about 1-2 seconds around the part containing the bird sound. The images and the sounds are combined to form a multimodal dataset, which is available at [165].

5.1.4 Image-text-video dataset

For further experimentation, a dataset consisting of image, text and video modalities has also been used. The dataset was constructed by using the 2995 YouTube videos of the EVVE video event dataset [166], enhanced with their thumbnail images and textual descriptions, obtained from YouTube.
The textual descriptions were filtered from any stop words. The videos of the EVVE dataset are split to 13 categories of variable size, and this information has been used as the ground truth for the evaluation of the proposed method.

5.1.5 Image-text dataset

In order to test the applicability of the proposed method to larger datasets, the Web Queries dataset [167], containing objects of two modalities, namely images and text, has also been used. The dataset consists of 71478 images, collected using web search engines, accompanied by text metadata. The text metadata used hereby are the words surrounding the images in the web pages in which they were found. The images of the Web Queries dataset are manually labeled as relevant or not to the text query used to retrieve them. This information can be used as the ground truth for the evaluation of the proposed techniques. The objects are split into 352 classes, each of variable size.

5.2 Multimedia descriptors

In order for the proposed methods to be applied to any of the above datasets, the modalities need to be expressed in vectorial form, or in the form of distance or kernel matrices, constructed from the data. The vectorial descriptors extracted from the images, sounds, videos and text of the datasets of Section 5.1 are described in the following. For each descriptor type, the distance measure used to calculate the distance between the descriptors of two objects is also presented. The distance measures are generally different for the various descriptor types, since a different notion of distance is more appropriate at each case. In the following, the distance measures used are the ones that are most commonly used in the literature for each descriptor type. Details about some of the descriptors and their corresponding distance measures have also been set according to [28].

5.2.1 Image descriptors

Each image in the datasets is described with various image features, serving as the multiple modalities in the proposed multimodal approach. Hereby, a subset of the image descriptors used in [28] is employed, along with a
color descriptor, although extending the evaluation to use more descriptors is straightforward. The features extracted from the images are used to calculate unimodal distances between them (Section 3.2), to be used in the proposed methods. The image descriptors and details for the corresponding distance measures are presented below:

- **SIFT**: SIFT features [96] are extracted from salient keypoints detected in each image. Then distances between each pair of images are calculated using the distance measure presented in [168]. For two images $I_1$ and $I_2$, this distance measure is defined as follows: Let the number of keypoints detected in $I_1$ be $P_1$. A feature vector $x_{1,i}$, $i = 1 \ldots P_1$, is extracted from each keypoint of $I_1$. Moreover, the position of each keypoint on image $I_1$ is denoted by $r_{1,i}$. Similar notation is considered also for image $I_2$. Then, the distance measure is defined as:

$$d(I_1, I_2) = d(I_1 \rightarrow I_2) + d(I_2 \rightarrow I_1),$$

where

$$d(I_1 \rightarrow I_2) = \frac{1}{P_1} \sum_{i=1}^{P_1} \min_{j=1 \ldots P_2} \left( ||x_{1,i} - x_{2,j}||^2 + \frac{\lambda}{r_0} ||r_{1,i} - r_{2,j}|| \right)$$

and similarly for $d(I_2 \rightarrow I_1)$. Here $||\cdot||$ denotes the Euclidean norm, $r_0$ is the average image size and $\lambda$ is a trade-off parameter between the feature vector distance and the keypoint distance on the image.

- **PHOG**: PHOG descriptors [169] are extracted from the images. The pyramid level is limited to 2. The $\chi^2$ distance is used to calculate distances between each pair of images. The $\chi^2$ distance is a modification of the Euclidean distance, where a weight is assigned to each dimension:

$$d(x, y) = \sqrt{\sum_{k=1}^{D} \frac{1}{c_k} (x_k - y_k)^2}.$$  

For the calculation of the $\chi^2$ distance, the two input vectors are considered to originate from a population of vectors. Hereby, $c_k$ is the sum of all elements of dimension $k$ in the population, normalized by the total sum of all elements in all dimensions in the population.
• **GIST:** The images are resized to 128 × 128 pixels and then GIST descriptors \([170]\) are extracted from them. Euclidean distances are calculated between each pair of images:

\[
d(x, y) = \sqrt{\sum_{k=1}^{D} (x_k - y_k)^2}. \tag{28}
\]

• **GB-DIST:** Geometric blur descriptors \([171]\) are extracted from the images and then the distance measure of \([168]\) is used for the pairwise distances, as for the SIFT features.

• **GB:** Again the geometric blur descriptor is used and the distance measure of \([168]\) is used, but the spatial distance between keypoints is ignored in the calculation of the distances, i.e. the distance measure is calculated as:

\[
d(I_1 \rightarrow I_2) = \frac{1}{P_1} \sum_{i=1}^{P_1} \min_{j=1 \ldots P_2} ||x_{1,i} - x_{2,j}||^2 \tag{29}\]

• **COLOR:** The histograms of the RGB color values of the images are also used as descriptors. Three histograms are constructed per image, for the three color channels separately. The Bray-Curtis dissimilarity is used as a distance measure between the histograms of a single channel:

\[
d(x, y) = \frac{\sum_{k=1}^{D} |x_k - y_k|}{\sum_{k=1}^{D} (x_k + y_k)}. \tag{30}\]

The Bray-Curtis dissimilarity takes values in the range \([0, 1]\), where 0 means that the two vectors are the same, while 1 that they do not have any bin in common. The average of the dissimilarities in the three channels is used as the distance measure for the color descriptors.

### 5.2.2 Sound descriptors

For the description of the sound files of the image-sound dataset, **MFCC** descriptors \([172]\) are extracted from the sounds. The sounds are split to timeframes of 10ms each and MFCC features are extracted from each timeframe. The minimum, maximum, mean and variance of the timeframes for each of the 13 Mel bands are used as the final descriptors, resulting in descriptors of 52 dimensions. As a measure of distance between two MFCC descriptors, the
Bray-Curtis dissimilarity is used, similarly to the color features above. The Bray-Curtis dissimilarity considers that the two input vectors are histograms of $D = 52$ bins.

### 5.2.3 Text descriptors

As a text descriptor, for numerically describing the text in the datasets, a **term histogram** has been used. A histogram of the terms appearing in the text, over a dictionary of 45382 words, is constructed from the words appearing in the accompanying text of all the images in the related datasets. The dictionary is constructed from all the words appearing in the datasets.

As a measure of distance between two text descriptors, the *cosine dissimilarity* is used. Given two histograms $x$ and $y$, the cosine dissimilarity is defined as follows:

$$d(x, y) = 1 - \frac{x \cdot y}{||x|| ||y||},$$

where $x \cdot y$ is the inner product of the two vectors and $||x||$ denotes the Euclidean norm of vector $x$.

### 5.2.4 Video descriptors

For the description of videos, the **Mean-Multi VLAD (MMV)** descriptor [166] has been used. The extraction of MMV descriptors internally employs a bag-of-visual words model, utilizing SIFT descriptors and a dictionary of visual words. In this respect, the nature of the MMV descriptor is similar to the term histogram descriptor described above. For this reason, as a measure of distance between two MMV descriptors, the cosine dissimilarity is used, similar to the text descriptors.

### 5.3 Quantitative evaluation measures

The datasets presented in Section 5.1 contain objects that are semantically split into distinct classes. Thus the desired low-dimensional mapping would be one in which clusters of similar objects would be apparent. The results of the proposed multimodal dimensionality reduction techniques are evaluated in terms of how well separated the ground truth semantic data classes are in the low-dimensional mapping. This mapping is ultimately used hereby for three purposes/tasks:
The success of the proposed dimensionality reduction methods is evaluated in terms of its success in the above three tasks, so the quantitative evaluation measures used are specific for each task. Due to the desired property of the low-dimensional mapping to consist of clusters of similar data, the evaluation measures presented below are often based on data clustering concepts.

5.3.1 Dunn index for visualization

One of the measures used for the performance evaluation in visualization tasks is the Dunn Index (DI) [173]. The Dunn Index evaluates the visual separation of the data classes, by measuring how much the objects of the same class are placed close to each other, while objects of different classes are separated. This conforms to the Gestalt law of proximity [4] [5], which states that semantically similar objects should be put close to each other in order for a visualization to be better perceived.

Given a set of points on the two dimensional plane, each classified as belonging to one of a number of classes, the Dunn index considers that maximal class separation is achieved if there is a large distance between different classes of points, while each class has a small diameter. The Dunn index is defined as the ratio of the minimum distance between the classes of points over the maximum class diameter. A large minimum class distance and a small maximum class diameter result in a large value for the Dunn index, indicating a clear class separation. Comparing two low-dimensional mappings, the one with the larger Dunn index separates the classes more clearly.

For a dataset $X$ of $N$ row vectors $x_1, x_2, \ldots, x_n$, $x_i \in X$, hereby the points of the low-dimensional space, split into $m$ classes $C_1, C_2, \ldots, C_m$, $C_i \subset X$, $\bigcup_{i=1}^{m} C_i = X$, the Dunn index (DI) is defined as

$$
DI = \frac{\min_{1 \leq i < j \leq m} \text{dist}(C_i, C_j)}{\max_{1 \leq k \leq m} \text{diam}(C_k)},
$$

where dist$(C_i, C_j)$ is a measure of the spatial distance between classes $C_i$ and $C_j$ and diam$(C_k)$ is a measure of the diameter of class $C_k$. 

99
Table 4: Notions of cluster distance and diameter used for the Dunn index.

**Cluster distance**

| Distance of closest points | \( \text{dist}(C_i, C_j) = \min_{x_q \in C_i} \min_{x_r \in C_j} ||x_q - x_r|| \) (32) |
|----------------------------|------------------------------------------------------------------|
| Distance of furthest points| \( \text{dist}(C_i, C_j) = \max_{x_q \in C_i} \max_{x_r \in C_j} ||x_q - x_r|| \) (33) |
| Distance of centroids      | \( \text{dist}(C_i, C_j) = \left\| \frac{1}{|C_i|} \sum_{x_q \in C_i} x_q - \frac{1}{|C_j|} \sum_{x_r \in C_j} x_r \right\| \) (34) |

**Cluster diameter**

| Maximum distance           | \( \text{diam}(C_i) = \max_{x_q, x_r \in C_i} ||x_q - x_r|| \) (35) |
|----------------------------|------------------------------------------------------------------|
| Mean distance of all pairs of points | \( \text{diam}(C_i) = \frac{1}{|C_i||C_i - 1|} \sum_{x_q, x_r \in C_i} ||x_q - x_r|| \) (36) |
| Mean distance from centroid| \( \text{diam}(C_i) = \frac{1}{|C_i|} \sum_{x_q \in C_i} \left\| x_q - \frac{1}{|C_i|} \sum_{x_r \in C_i} x_r \right\| \) (37) |

The distance between two classes of points is defined in terms of the distances between their constituting points. Similarly, the diameter of a class of points can be defined based on the distances between its points. However, various definitions of distance and diameter can be used. The definitions used in this thesis are described in Table 4.

The Dunn Index has been widely used for the evaluation of clustering algorithms [174] [175]. It is important to note that, although it is a clustering evaluation measure, it takes into account the relative positions of the data points in the space. This is in contrast to other clustering evaluation measures, such as the Rand index, the F-measure, or the Jaccard index, which consider that the result of clustering is the separation of the points into logical sets, without considering the relative distances between points in the data space. In logical clustering, the data points are assigned labels, denoting the cluster they have been assigned to. However, the Dunn Index does not use such labels, but the raw positions of the data in the space. This
fact renders the Dunn Index a suitable measure for class separation, even for the evaluation of dimensionality reduction methods.

5.3.2 Average Isoperimetric Quotient for visualization

According to the visualization Gestalt law of simplicity [4] [5], similar objects should be grouped in clusters of simple, compact shapes. In order to evaluate the regularity and simplicity of the shapes of the classes of objects, a shape compactness measure has been used, namely the Average Isoperimetric Quotient (AIQ). The Isoperimetric Quotient is a popular measure of the compactness of a shape [176], and is based on the Isoperimetric inequality, stating that the following holds for a closed curve:

$$4\pi A \leq P^2,$$

where \( P \) is the length of the curve and \( A \) the area that it encloses. The equality holds when the curve is a circle. The more compact a shape is, the more it approaches a circle and the closer the Isoperimetric Quotient is to 1. Hereby, the average of the Isoperimetric Quotients of the shapes of all classes in the visualization is used as a collective measure of compactness for all the classes.

Given a set \( X \) of \( N \) points \( x_1, x_2, \ldots, x_n, x_t \in X \) (hereby the points of the 2D visualizations) split into \( m \) classes \( C_1, C_2, \ldots, C_m, C_i \subset X, \bigcup_{i=1}^{m} C_i = X \), the Isoperimetric Quotient (IQ) of class \( C_i \) is defined as:

$$\text{IQ}_i = \frac{A_i}{A'_i},$$

where \( A_i \) is the area of the shape covered by the points of \( C_i \) and \( A'_i \) is the area of a circle having the same perimeter \( P_i \) as the shape of \( C_i \):

$$A'_i = \frac{P_i^2}{4\pi},$$

where \( P_i \) is the perimeter of the shape covered by the points of \( C_i \). As already mentioned, the IQ of a circle, which is considered as the most compact shape, is 1 and IQ is decreased for more complex shapes. The Average Isoperimetric Quotient is defined as the average over the IQs of all classes of points:

$$\text{AIQ} = \frac{1}{m} \sum_{i=1}^{m} \text{IQ}_i \quad (38)$$

101
5.3.3 Mean Recognition Rate for object classification

A common measure for the evaluation of the performance of an object classification result is the Mean Recognition Rate (MRR). After a set of testing data have been classified, the MRR measures the success of the classification, based on the true class labels of the data. The MRR is defined as:

\[ \text{MRR} = \frac{1}{T} \sum_{i=1}^{T} \frac{N_{\text{correct},i}}{N_{\text{total},i}} \]  

(39)

where \( N_{\text{correct},i} \) is the number of test data which have been correctly recognized and \( N_{\text{total},i} \) is the total number of test data. The closer the MRR is to 1, the more accurate the classification.

5.3.4 Rand index for clustering

The Rand Index is a common measure of the quality of a clustering of a set of data, based on their true class labels. Considering a set of \( N \) data objects, let \( C = \{c_1, c_2, \ldots, c_N\} \) be the ground truth class labels of the data and \( Q = \{q_1, q_2, \ldots, q_N\} \) be the cluster assignments of the data, as returned from the clustering algorithm. The Rand Index (RI) measure is defined as:

\[ \text{RI} = \frac{N_a + N_b}{N_a + N_b + N_c + N_d}, \]  

(40)

where

\[ N_a = |\{i, j \in \{1 \ldots N\} : c_i = c_j, q_i = q_j\}| \]
\[ N_b = |\{i, j \in \{1 \ldots N\} : c_i \neq c_j, q_i \neq q_j\}| \]
\[ N_c = |\{i, j \in \{1 \ldots N\} : c_i = c_j, q_i \neq q_j\}| \]
\[ N_d = |\{i, j \in \{1 \ldots N\} : c_i \neq c_j, q_i = q_j\}| \]

and \(|A|\) denotes the cardinality of set \( A \). In other words, \( N_a \) is the number of true positives, \( N_b \) the number of true negatives, \( N_c \) the number of false negatives and \( N_d \) the number of false positives. A clustering that splits the data in a manner that follows exactly the true class labels has a Rand Index of 1, which is the largest value that RI can take. Thus the higher the Rand Index is, the more accurate the clustering is with respect to the ground truth data.
5.4 Experimental evaluation of multi-objective dimensionality reduction

This section contains the results of the experimental evaluation of the multi-objective dimensionality reduction approach, described in Section 3. Three kinds of experiments have been conducted, to assess various aspects of the proposed method. The first experiment aims to illustrate the fact that using multi-objective optimization for dimensionality reduction can produce mappings which cannot be accomplished by scalarization methods that combine the modalities using weights. The second experiment is conducted in order to compare between using two different types of objective functions to be used as the multiple objectives. Finally, the third and largest set of experiments aim to evaluate the performance of the multi-objective dimensionality reduction method in realistic settings, using real-world datasets, in comparison to existing methods.

5.4.1 Experiments with artificial data

As already mentioned, the purpose of the first experiment is to show the superiority of the proposed technique compared to scalarization techniques, with respect to the variety of solutions that can be discovered. In order to conduct the experiment in a controlled environment, the artificial dataset described in Section 5.1.1 has been used. For this experiment, the minimum angle objective functions (Section 3.3) have been used and the output dimensionality has been set to 2. By solving the multi-objective optimization problem of equation (8), a set of Pareto-optimal solutions (mappings) has been calculated. The Pareto front of the optimal solutions is illustrated in Fig. 10. Due to the small size of the problem, a brute-force approach was adopted for the computation of the Pareto front, where all possible placements, with the points restricted on a grid, were considered and the minimum angle objective for each modality was calculated. The points of the Pareto front are depicted as blue points. Each point represents a specific mapping of the data on the two-dimensional space. For each mapping, the horizontal and vertical axes are the values of the minimum angle objective functions for the size and the color modalities, calculated according to equation (9). The two-dimensional placements corresponding to some representative points of the front are presented in Fig. 11. The edges of the MSTs calculated for
Figure 10: The points of the Pareto front (blue), representing solutions (mappings) of the multi-objective optimization problem. Some representative points are depicted in bold. The horizontal and vertical axes are the values of the minimum angle objective functions for the size and the color modalities. The red points represent placements calculated by minimizing a weighted sum of the multiple objective functions, for various weight combinations. The green points represent solutions calculated by an early fusion of the modalities.

The closer a point on the Pareto front is to the horizontal axis (i.e. small value for the color objective) the more suitable the corresponding mapping is for revealing the structure of the color MST (e.g. point C in Fig. 10 and the corresponding placement (C) of Fig. 11). Points that are in the middle of the Pareto front, such as point B, represent placements where both MSTs are simultaneously mapped in the best possible way.

In order to compare the proposed method with standard weighted-sum-based methods, mappings for the same dataset have also been calculated by minimizing a weighted sum, $J_{\text{sum}}$, of the multiple objective functions,
Figure 11: Two-dimensional mappings corresponding to point A, B and C in Fig. 10. Mappings corresponding to the left part of the Pareto front of 10 favor the size modality, while mappings corresponding to the right part of the front favor the color modality.

defined as:

\[
J_{\text{sum}}(p) = \sum_{m=1}^{M} w_m J_m(p)
\]  

(41)

\[
w_m \in [0, 1], m = 1 \ldots M
\]

\[
\sum_{m=1}^{M} w_m = 1
\]

where \( M = 2 \) in the artificial dataset used hereby. Several combinations of the weights \( w_1 \) and \( w_2 \) have been used, by discretizing the \([0, 1]\) range with a step of 0.05. For each weight combination, the objective function of equation (41) has been minimized and the optimal placement is projected on the plane. The resulting placements for the various weight combinations are depicted in red in Fig. 10.
As can be derived from Fig. 10, the mappings calculated with the weighted sum-based method are Pareto-optimal points of the multi-objective optimization problem, which is expected, as mentioned in [158]. However, not all Pareto-optimal placements are discovered by altering the weights of the weighted-sum method. As it becomes obvious from Fig. 10, the Pareto front is non-convex. The points that lie in the concave part of the Pareto front cannot be found by the weighted-sum method, as noted in Section 3.1 and Fig. 4. Thus, points such as point B in Fig. 10 are missed out, although they correspond to visually aesthetic placements, which is a significant advantage of the proposed approach over commonly used weighted-sum-based methods.

The proposed method has also been compared to a method which fuses the modalities at an earlier stage, similar to the method of [177]. Specifically, after calculating the pairwise distances among the data of the dataset, using the unimodal distance measures $d_m$ (refer to section 3.2), multimodal distances are calculated for all pairs, by summing the unimodal ones:

$$d(O_1, O_2) = \sum_{m=1}^{M} w_m d_m(O_1, O_2)$$  \hspace{1cm} (42)

$$w_m \in [0, 1], m = 1 \ldots M$$

$$\sum_{m=1}^{M} w_m = 1$$

Using these fused distances, a multimodal graph $G(O, E)$, $E \in O \times O$, and the corresponding minimum spanning tree $T(O, E')$, $E' \subseteq E$ are constructed, similar to the unimodal graphs and MSTs of section 3.2. By minimizing the minimum angle objective function for this fused MST, a final placement is calculated for the data. By altering the weights of the distance sum, different placements are produced. In order to map these placements on a Pareto diagram and compare them to the ones produced by the multi-objective and the weighted objectives methods, each final placement is evaluated in terms of the unimodal objectives of section 3.2.

The points of the Pareto diagram corresponding to the placements produced with the early fusion method are depicted in green in Fig. 10. All points lie to the upper right of the Pareto front, meaning that none is Pareto-optimal, with respect to the unimodal objective functions.
5.4.2 Comparison of different objective functions

Using the descriptors of Section 5.2, distances are calculated among the objects of the datasets and distance graphs and the corresponding MSTs are constructed for each modality, according to Section 3.2. The multiple objective functions of the multi-objective framework (Section 3.1) are defined using graph aesthetic-based measures of the unimodal MSTs.

As a further evaluation of the proposed framework, the two graph aesthetic measures of Section 3.3 have been used as the objective functions of the multi-objective framework (Section 3.1), again for a two-dimensional mapping. The results of using each measure are compared, using the image dataset of Section 5.1.2 to investigate their applicability.

Using these two kinds of objective functions, two sets of Pareto-optimal solutions are calculated. Comparing the two sets of solutions is not straightforward, due to the fact that there is no one-to-one correspondence between the solutions of one Pareto front to the solutions of the other. There is no direct mapping between the position of a point on a Pareto front and the corresponding trade-off among the multiple objectives.

As a guide for the comparison of the two cost functions, only the points of the Pareto front which correspond to the minimization of the weighted sum of the objectives, for various weights, are considered. If a specific weight combination is used, the solutions produced by the two cost functions can be compared.

In Fig. 12, an example of using the minimum angle (Fig. 12a) and the potential (Fig. 12b) objective functions for a case of two modalities is presented. The PHOG and GIST image descriptors, described in Section 5.2.1 are used. The coordinates of the points in the figures are screen coordinates, normalized to lie within a square of unary side. The mapping correspond to a weight combination where the weights of the two modalities are equal. From this qualitative comparison, it can be seen that the potential objective produces a more aesthetically pleasing placement of the data, where the data classes are more easily perceived and more uniform in shape, conforming to the dimensionality reduction effectiveness criteria. The superiority of the potential objective to the angle one is due to the fact that the minimum angle objective only considers the angle between the edges, allowing the points to overlap. However, this superiority is dependent on the dataset used, i.e. the potential objective is superior to the angle one in this specific dataset,
as well as in the other multimedia datasets used in the following sections. In other cases, the minimum angle objective function could be superior to the potential objective, as in the example of Section 5.4.1, where it was able to demonstrate a non-convex Pareto front.

Figure 12: Comparison of two-dimensional mappings of the image dataset, using the minimum angle and potential objective functions. The PHOG and GIST image descriptors are used as modalities. The colors denote the different image classes. The coordinates of the points are screen coordinates, normalized so as to lie within a square of unary side, while preserving the aspect ratio of the original points. (a) Mapping of the dataset using the minimum angle objective function. (b) Mapping of the dataset using the potential objective function.

A quantitative comparison of the two objective functions has also been performed, using the evaluation measures of Section 5.3 and considering three modalities. Various values for the weights have been used, covering the whole possible range, by discretizing the $[0, 1]$ range for each unimodal weight with a step value of 0.1 and satisfying the constraint $\sum_{m=1}^{M} w_m = 1$. As a representative diagram, Fig. 13 depicts the values of the Dunn index for the two objective functions at the various weight combinations, using the centroid distance as the class distance and the maximum distance between points as the class diameter.

For presentation purposes, only the SIFT, PHOG and GIST descriptors are used. The actual independent variables for the diagrams are just the
weights of the two of the three modalities (here SIFT and PHOG), as the third weight must have a value such that the sum of all weights is 1. If more than three modalities were used, the independent weights would be more, hence diagrams of higher dimensions would be needed.

It is obvious from Fig. 13 that the use of the potential objective leads to larger Dunn indices, which confirms that the potential objective results in mappings where the data classes are more clearly separated and uniform. For the experiments of the following sections, the potential objective function will be used.

5.4.3 Comparison with existing methods

In a third series of experiments, the proposed method, using the potential objective function, has been compared with two existing methods for multi-modal dimensionality reduction. The results of the experiments show that the proposed method outperforms existing ones, even when the Pareto front is convex. The two methods which are used for comparison are the following:

MKL-LPP [28] In [28], the Multiple Kernel Learning (MKL) dimensionality reduction framework was proposed. Multiple kernel matrices are formed
Figure 14: Comparison of MKL-LPP, MST-FD and the proposed method, using the image dataset. The SIFT, PHOG, GIST, GB and GB-DIST image descriptors of Section 5.2.1 are used as multiple modalities. The colors denote the different image classes. The coordinates of the points produced by all methods are screen coordinates, normalized so as to lie within a square of unary side, while preserving the aspect ratio of the original points. (a) Mapping using the MKL-LPP method of [28]. (b) Mapping using the MST-FD method of [177]. (c) Mapping using the proposed method.

from the data, by using several feature representations for them. The multiple kernels are fused in one kernel matrix via a weighted sum. Various existing unimodal dimensionality reduction methods, formulated as graph embedding methods, are then incorporated in the framework and make use of the fused kernel. Dimensionality reduction is formed as an optimization problem where the optimization variables are the sample coefficients as well as the weights of the kernel weighted sum. Hereby, Locality Preserving Projections (LPP) [50] is used as a dimensionality reduction method, which utilizes neighborhood information extracted from the data. LPP was used in [28] for the task of unsupervised low-dimensional data embedding. Using LPP in the MKL framework, the data are projected into a low-dimensional space. Afterwards, Multidimensional Scaling [34] is employed in order to reduce the dimensionality of the data to two dimensions and present the result graphically.

**MST-FD [177]** In [177], the Force-Directed Maximum Spanning Tree (MST-FD) technique is presented for mapping a set of multimodal data on the two-dimensional space. First, unimodal similarities are calculated...
among the data for each modality. Then a multimodal similarity measure is defined between two multimodal objects, as a weighted sum of the unimodal similarities between them. The resulting multimodal similarities are used to form a similarity graph among the data. The edges of the graph are weighted according to the similarities among the vertices. The maximum spanning tree of the complete similarity graph is then calculated. The tree connects vertices having large similarities, hence it is a representation of the structure of the dataset. The maximum spanning tree is then placed on a two-dimensional plane, using a force-directed graph placement algorithm [60].

Both the MKL-LPP and the MST-FD methods mentioned above produce a single solution as their result. This solution corresponds to a specific weight combination for the weighted sums employed. This weighted combination is determined either automatically from the data (MKL-LPP) or interactively, via user feedback (MST-FD). On the other hand, in the method proposed hereby, a set of solutions, i.e. the Pareto set, is calculated. In order to compare the multi-objective approach to MKL-LPP and MST-FD, there is a need to compare between single solutions. The solutions of MKL-LPP and MST-FD are characterized by the specific weights assigned to the modalities, so if the same weights are used for both methods, this is a baseline for comparing between them. In the multi-objective approach, no weights are considered, so the comparison is not so straightforward. However, as noted in Section 3.1 and in [158] and depicted in Fig. 4, the multi-objective approach contains a weighted sum-based scalarization approach as a sub-case. There are some points in the Pareto front that correspond to the solutions of an approach that considers a weighted sum of the objectives. If, moreover, the Pareto front is convex, the whole Pareto front corresponds to the solutions of the weighted sum approach, with varying weights. Hereby, in order to compare the multi-objective approach to MKL-LPP and MST-FD, a set of weights will be considered as the comparison base. The solutions of MKL-LPP and MST-FD corresponding to these weights will be considered, while for the multi-objective approach, the Pareto-optimal solution corresponding to the solution of a weighted sum-based scalarization method using the same weights will be considered. The following experimental results show that the proposed method outperforms the existing ones, even if only these solutions are considered, hence indicating the applicability of the proposed method.
also in cases of convex Pareto fronts.

In Fig. 14, example two-dimensional mappings of the image dataset of Section 5.1.2, using the two existing methods and the proposed one, are depicted. The SIFT, PHOG, GIST, GB and GB-DIST descriptors of Section 5.2.1 are used as modalities. These mappings correspond to equal weights for all modalities. Comparing the three mappings, the classes are better separated in the MST-FD and the proposed methods, rather than in MKL-LPP. Moreover, the shape of the classes of points in the proposed method is more uniform than in the other two methods. If the output mapping is used in e.g. an image retrieval task, this makes it easier for a user to find similar images to a given one, since, according to the simplicity law, it is expected that they are distributed uniformly around and near the given image. If this result is used for visualization, the available visualization space is also better utilized in the proposed method, since more area is covered by the points, making it easier to e.g. replace them with image thumbnails in a real application.

The quantitative evaluation of the two-dimensional mappings of Fig. 14 is presented in Table 5. For the evaluation of each mapping, the Dunn index, with all notions of class distance and class diameter, and the AIQ compactness measures of Section 5.3 are used.

The Dunn index for the proposed method is much larger than that of both the MKL-LPP and the MST-FD methods, in all cluster distance and diameter metrics. This suggests that the classes are separated more clearly in the proposed method. The AIQ measure is also larger for the proposed method, indicating that the image classes are simpler and more uniform in shape.

A more extensive comparison with the existing methods can be performed by considering many different weight combinations for the modalities. The weights are considered according to the setting of Section 5.4.2. The mappings corresponding to the various weight combinations, as produced by all methods, have been compared using the Dunn index, with the various notions of cluster distance and diameter, and the AIQ compactness measure. Some representative results are illustrated in Fig. 15. For presentation purposes, only the SIFT, PHOG and GIST descriptors are used, similar to Section 5.4.2.

As can be derived from Fig. 15, the multi-objective method outperforms
Table 5: Values of the evaluation measures for the two-dimensional mappings of Fig. 14.

<table>
<thead>
<tr>
<th>Evaluation measure</th>
<th>MKL-LPP</th>
<th>MST-FD</th>
<th>Multi-objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>DI (closest points, maximum)</td>
<td>0.0006</td>
<td>0.0152</td>
<td><strong>0.0364</strong></td>
</tr>
<tr>
<td>DI (furthest points, maximum)</td>
<td>0.1884</td>
<td>0.5203</td>
<td><strong>0.6612</strong></td>
</tr>
<tr>
<td>DI (centroids, maximum)</td>
<td>0.0170</td>
<td>0.0267</td>
<td><strong>0.1541</strong></td>
</tr>
<tr>
<td>DI (closest points, mean pairwise)</td>
<td>0.0018</td>
<td>0.0450</td>
<td><strong>0.0938</strong></td>
</tr>
<tr>
<td>DI (furthest points, mean pairwise)</td>
<td>0.5901</td>
<td>1.5424</td>
<td><strong>1.7065</strong></td>
</tr>
<tr>
<td>DI (centroids, mean pairwise)</td>
<td>0.0532</td>
<td>0.0792</td>
<td><strong>0.3978</strong></td>
</tr>
<tr>
<td>DI (closest points, mean centroid)</td>
<td>0.0025</td>
<td>0.0620</td>
<td><strong>0.1271</strong></td>
</tr>
<tr>
<td>DI (furthest points, mean centroid)</td>
<td>0.7978</td>
<td>2.1237</td>
<td><strong>2.3126</strong></td>
</tr>
<tr>
<td>DI (centroids, mean centroid)</td>
<td>0.0719</td>
<td>0.1091</td>
<td><strong>0.5391</strong></td>
</tr>
<tr>
<td>AIQ</td>
<td>0.3711</td>
<td>0.2895</td>
<td><strong>0.6592</strong></td>
</tr>
</tbody>
</table>

both the MKL-LPP and the MST-FD methods in all weight combinations, producing mappings with a larger Dunn index and AIQ, indicating that classes are more clearly separated and more uniform in shape.

The same figures also provide an illustration of the importance, or the appropriateness, of each modality for the problem of dimensionality reduction in this dataset. As can be observed from the diagrams of Fig. 15, the PHOG and GIST modalities are more appropriate, since the evaluation measures take large values even when only these two modalities are considered, while the measures decrease as more weight is assigned to the SIFT modality. This is not a characterization of the features in general, but with respect to the specific dataset.

Similar experiments have been conducted using the image-sound dataset of Section 5.1.3. Fig. 16 depicts an example two-dimensional mapping of the dataset, using the three examined methods. The SIFT and COLOR descriptors from Section 5.2.1 are used as the image modalities and the MFCC descriptor of Section 5.2.2 for the sound modality. The mappings of Fig. 16 correspond to equal weights for the modalities. Qualitatively, it can be observed that the proposed method outperforms the other two by simultaneously organizing the data according to their classes, forming point classes that are simple in form and utilizing more of the available area.
Figure 15: Comparison of the Dunn index (a-c) and the AIQ (d) for MKL-LPP, the MST-FD and the multi-objective method, for the image dataset and for various modality weight combinations. For the Dunn index, the cluster diameter is defined using the maximum distance between points, and the cluster distance is defined using (a) the closest points distance, (b) the furthest points distance and (c) the centroids distance.

For a quantitative comparison, the Dunn index and the AIQ compactness evaluation measures of Section 5.3 have been used. The values of the evaluation measures for the mappings of Fig. 16 are presented in Table 6. The values for the proposed method are larger than for the existing methods for most of the notions of class distance and diameter, used for the calculation of the Dunn Index, as well as for the AIQ measure, confirming that the multi-objective method manages to conform both to the proximity and simplicity laws class separation. As a further experiment, Fig. 17, similar to Fig. 15, depicts the values of the Dunn index and AIQ for the three methods and for various modality weights. Although in Fig. 17(b) and 17(c) the values of the Dunn index for some weight combinations are smaller for the proposed method, there is a 6.95% and 1.56% improvement in the average values of the Dunn index of the proposed method over the MST-FD method, using the furthest point distance (Fig. 17(b)) and the centroids distance (Fig. 17(c)), respectively. Similarly, there is a 37.92% and 25.38%
improvement, respectively, compared to the MKL-LPP method.

Figure 16: Comparison of MKL-LPP, MST-FD and the multi-objective method for the image-sound dataset. The SIFT and COLOR descriptors of Section 5.2.1 and the MFCC descriptors of Section 5.2.2 are used as the modalities. The colors denote the different image classes. The coordinates of the points produced by all methods are screen coordinates, normalized so as to lie within a square of unary side, while preserving the aspect ratio of the original points. (a) Mapping using the MKL-LPP method of [28]. (b) Mapping using the MST-FD method of [177]. (c) Visualization using the multi-objective method.

Table 6: Values of the evaluation measures for the two-dimensional mappings of Fig. 16

<table>
<thead>
<tr>
<th>Evaluation measure</th>
<th>MKL-LPP</th>
<th>MST-FD</th>
<th>Multi-objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>DI (closest points, maximum)</td>
<td>0.0025</td>
<td>0.0087</td>
<td><strong>0.0253</strong></td>
</tr>
<tr>
<td>DI (furthest points, maximum)</td>
<td>0.4939</td>
<td>0.6751</td>
<td><strong>0.6809</strong></td>
</tr>
<tr>
<td>DI (centroids, maximum)</td>
<td>0.0416</td>
<td><strong>0.0551</strong></td>
<td>0.0465</td>
</tr>
<tr>
<td>DI (closest points, mean pairwise)</td>
<td>0.0070</td>
<td>0.0190</td>
<td><strong>0.0597</strong></td>
</tr>
<tr>
<td>DI (furthest points, mean pairwise)</td>
<td>1.3608</td>
<td>1.4773</td>
<td><strong>1.6079</strong></td>
</tr>
<tr>
<td>DI (centroids, mean pairwise)</td>
<td>0.1146</td>
<td><strong>0.1206</strong></td>
<td>0.1097</td>
</tr>
<tr>
<td>DI (closest points, mean centroid)</td>
<td>0.0089</td>
<td>0.0240</td>
<td><strong>0.0844</strong></td>
</tr>
<tr>
<td>DI (furthest points, mean centroid)</td>
<td>1.7168</td>
<td>1.8678</td>
<td><strong>2.2736</strong></td>
</tr>
<tr>
<td>DI (centroids, mean centroid)</td>
<td>0.1445</td>
<td>0.1525</td>
<td><strong>0.1552</strong></td>
</tr>
<tr>
<td>AIQ</td>
<td>0.4216</td>
<td>0.1652</td>
<td><strong>0.5770</strong></td>
</tr>
</tbody>
</table>

In order to test the performance of the multi-objective method in a case
Figure 17: Comparison of the Dunn index (a-c) and AIQ (d) for MKL-LPP, the MST-FD and the multi-objective method, for the image-sound dataset and for various modality weight combinations. For the Dunn index, the cluster diameter is defined using the maximum distance between points, and the cluster distance is defined using (a) the closest points distance, (b) the furthest points distance and (c) the centroids distance.

of three diverse modalities, experiments with the image-text-video dataset of Section 5.1.4 have also been conducted. The SIFT descriptor has been used for the image modality, the term histogram for the text and the MMV descriptor for the video one. Indicative Dunn index and Average Isoperimetric Quotient diagrams of two-dimensional mappings created with the multi-objective and the existing methods are depicted in Fig. 18, for various modality weights. For the Dunn index, the diagram corresponding to using the maximum distance between points for the cluster diameter and the furthest distance between points as the cluster distance is presented. The multi-objective method outperforms both the MKL-LPP and the MST-FD methods in most weight combinations.

Finally, in order to test the effectiveness of the multi-objective approach in performing dimensionality reduction in large datasets, further experiments have been conducted, using the image-text dataset of Section 5.1.5, which consists of several thousands of objects. Fig. 19 illustrates the Dunn in-
Figure 18: Comparison of the Dunn index (a) and the Average Isoperimetric Quotient (b) of the multi-objective, the MKL-LPP and the MST-FD method, for the image-text-video dataset and for various modality weight combinations. For the Dunn index, the cluster diameter is defined using the maximum distance between points and the cluster distance is defined using the furthest distance between points.

Figure 18: Comparison of the Dunn index (a) and the Average Isoperimetric Quotient of two-dimensional mappings created with the proposed and the existing methods, for various modality weights. This figure is similar to Figs. 15, 17 and 18, however the diagrams are two-dimensional, since only two modalities are hereby considered. In Fig. 19, for the Dunn index, the diagram corresponding to using the mean distance between points for the cluster diameter and the furthest distance between points as the cluster distance is presented, although similar results are obtained when using the other notions of cluster distance and diameter of Section 5.3.1. Again it can be seen that the proposed method outperforms the MKL-LPP and the MST-FD methods in most weight combinations, both in terms of cluster separation and in terms of cluster compactness.

As an indicative comparison with the Multimodal Graph Embedding (MGE) method, presented in the next section, MGE achieves a Dunn index of 1.46 and an AIQ of 0.26, for a subset of the image-text dataset. A subset was used, due to the very large size of the whole image-text dataset, which made the application of MGE require very large memory resources. Furthermore, in MGE, using various weights for the modalities is not relevant, since MGE computes an optimal set of weights. The optimal weights which led to the above numbers were 0.14 for the image and 0.84 for the text modality. The above results indicate that the MGE can perform better than the multi-objective approach, with respect to class separation, but its performance is
Figure 19: Comparison of the Dunn index and the Average Isoperimetric Quotient of the multi-objective, the MKL-LPP and the MST-FD methods, for the image-text dataset and for various modality weight combinations. For the Dunn index, the cluster diameter is defined using the mean distance between points and the cluster distance is defined using the furthest distance between points.

worse when it comes to the cluster shape. Further comparisons between the proposed multi-objective and MGE approaches can be found in Section 5.5.3 below.

5.5 Experimental evaluation of Multimodal Graph Embedding

This section covers the experimental evaluation of the Multimodal Graph Embedding (MGE) dimensionality reduction method, described in Section 4. MGE has been tested with realistic multimedia datasets, in applications with various dimensionality reduction methods. The overall procedure for the experiments is the following: The input data are mapped to a low-dimensional space, using the MGE method, and then supervised and unsupervised learning algorithms are performed on the low-dimensional data. The performance
of the learning algorithms is measured with suitable evaluation metrics for each task, and is compared to using the single modalities and using the multiple kernel learning framework of [28]. For each of the following tasks, the dimensionality reduction method and the evaluation measures used are described and quantitative results are presented.

5.5.1 Evaluation in supervised object classification

The first application of the MGE method is for the task of object classification. The purpose is to map the multimodal data to a low-dimensional space, where the objects can be classified accurately.

**Dimensionality reduction method**  For this supervised learning task, the Local Discriminant Embedding (LDE) [51] dimensionality reduction method has been used. The LDE method was used within the MKL-DR framework with and without the use of the proposed modality weighting scheme, and the results of the MKL-DR method and MKL-DR enhanced with MGE were compared. The LDE method was selected for comparison with the MKL-DR framework of [28], as it was also used thereby. The image dataset of Section 5.1.2 has been split to a training and a testing subset. The training subset has been used to train the kernel LDE method and the testing subset has been used to test the performance of the compared methods, in an object recognition task.

For LDE, the elements of the unimodal affinity matrices $W_m$ and the constraint matrices $B_m$ are calculated as follows:

$$w_{m,ij} = \begin{cases} k_{m,ij}, & \text{if } c_i = c_j \text{ and } \\
 & \text{x}_i \text{ is among the } k \text{ nearest neighbors of } \text{x}_j, \\
 & \text{or } \text{x}_j \text{ is among the } k \text{ nearest neighbors of } \text{x}_i, \\
 & \text{according to } K_m \\
0, & \text{otherwise} \end{cases}$$
where $k_{m,ij}$ is the element of $K_m$ in the $i$-th row and $j$-th column, and $c_i$, $i = 1 \ldots N$ are class labels assigned to the input data. In LDE, the values of the $w_{m,ij}$ and $b_{m,ij}$ are commonly either binary, using a value of 1 in place of the $k_{m,ij}$, or real-valued, calculated from the distances between points, using a heat kernel. Hereby, the $k_{m,ij}$ values are used, as they are already produced from a heat kernel.

**Quantitative evaluation measure** Using the LDE affinity and constraint matrices derived from the training data, the MGE method is applied in order to calculate the optimal modality weights and then MKL-DR is performed to calculate the optimal coefficients $A$ and $\beta$ that map the training data on the low-dimensional space. Using these coefficients, the training data are mapped on low-dimensional points $Y$, using Eq. (25). In order to map a testing point on the same space, the vectors $k_m \in \mathbb{R}^N$, $m = 1 \ldots M$, is first calculated, whose elements are the kernel function values between the test point and each of the training points, using the same kernel function used among the training data, and considering each modality $m$ separately. Then, the low-dimensional mapping $y$ of the testing point is $y = A^T k_\beta$, $k_\beta = \sum_{m=1}^{M} \beta_m k_m$. After a testing point is mapped on the low-dimensional space, it is classified into one of the training classes using a nearest neighbor rule, by setting its class label equal to the most frequent class label appearing in its nearest training points.

For the quantification of the object recognition performance, the Mean Recognition Rate (MRR), described in Section 5.3.3, is used, similar to [28]. The recognition rate is averaged over a number of $T = 10$ random splits of the data, in order to eliminate any random effects of the specific training/testing splits.

**Quantitative results** The quantitative evaluation of the Multimodal Graph Embedding (MGE) framework has been conducted in comparison with the
performance of the single modalities and with the Multiple Kernel Learning (MKL) framework of [28], which uses a fixed, average, weight configuration for combining the multiple unimodal affinity matrices. Although a common approach for the evaluation of a multimodal learning method is to compare against a baseline using as features a mere concatenation of the multiple features of each object, such an approach is not applicable hereby, since some of the descriptors, e.g. SIFT, are not simple vectorial representations, but rather consist of sets of vectors, which are moreover different in size for different objects.

For each of the following experiments, for the performance of the single modalities, the unimodal LDE method has been used. For the multimodal cases, the MKL method has been applied twice, using the average affinity matrix, as in [28] and using the affinity matrix which results from the adaptive weighting of the proposed framework.

A series of three experiments have been conducted for each dataset, demonstrating the effect of various parameters on the final performance. The parameters examined are:

- the number of nearest neighbors used for the calculation of the unimodal affinity matrices (experiment #1),
- the number of dimensions of the output space, after the dimensionality reduction (experiment #2), and
- the ratio of the training data over all the data used (experiment #3).

The three experiments have been conducted using the image and image-text-video datasets of Sections 5.1.2 and 5.1.4, respectively. Table 7 summarizes the parameters used in the various experiments for each dataset.

The experimental results are depicted in Fig. 20, for the image and the image-text-video datasets. It is apparent from these figures that the performance of the multimodal methods is significantly higher than using each modality separately. Moreover, the use of the adaptive method of MGE for the determination of the optimal affinity matrix weights leads to higher recognition rates than with the use of the average affinity matrix of the MKL method. In all cases, MGE achieves about 3-5% increase in the recognition rate, compared to the MKL method. This increase in performance is kept
Table 7: Parameters used in the three experiments for the various datasets, in the task of supervised object recognition.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Image dataset</th>
<th>Image-text-video dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp. #1</td>
<td>Exp. #2</td>
</tr>
<tr>
<td>Dataset size</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>Number of modalities</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Number of classes</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Nearest neighbors</td>
<td>10-100</td>
<td>30</td>
</tr>
<tr>
<td>Output dimensions</td>
<td>50</td>
<td>10-100</td>
</tr>
<tr>
<td>Training ratio</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

relatively stable through all experiments and independently of the parameter being varied.

In case of the image dataset, when the number of nearest neighbors of the affinity matrices increases, the recognition rate for both methods seems to drop early and maintain a value little less than 0.8 for MGE and 0.77 for MKL, for 40-90 nearest neighbors, with a maximum of 0.822 and 0.792 for MGE and MKL, respectively. With respect to the number of output dimensions $P$, both methods seem to gain much in performance with $P$ increasing from 10 to 20 dimensions, and then they maintain a relatively stable value of 0.82 and 0.77 for MGE and MKL, respectively. Regarding the training ratio, both methods perform better when the number of training data increases, reaching a maximum of about 0.83 and 0.79 for MGE and MKL, respectively, at a training ratio of about 0.7. As a comparison to the state-of-the-art accuracies on the Caltech-101 benchmark dataset used hereby, the recent work of [30], which compares with state-of-the-art methods for object recognition, achieves an accuracy of about 73%, in the whole Caltech-101 dataset, when using about 50% of the images per class for training. The proposed method achieves about 80% accuracy for the same training image ratio. However, this is achieved for a subset of the whole dataset, the one used in [28], as described in Section 5.1.2, so the comparison is not straightforward. Nevertheless, these percentages indicate the potential of the proposed approach in object recognition tasks, compared to the state-of-the-art.

In case of the image-text-video dataset, the performance of the two methods is relatively fixed with respect to the number of nearest neighbors, with a value of about 0.78 for MGE and 0.74 for MKL. With respect to the number of output dimensions, the recognition rate seems to gradually decrease.
Figure 20: Comparison of the Mean Recognition Rate (MRR) for the separate modalities, the MKL and the MGE methods, with respect to (a) the number of nearest neighbors, $k$, used for the affinity matrices, (b) the number of the output dimensions, $P$, and (c) the ratio of training data over all data.

for both methods, while keeping their difference to about 4%. A decrease in performance is also apparent when the training ratio is varied, where the recognition rate falls from around 0.83 to 0.65 from training ratios varying from 0.2 to 0.8, while always the proposed method has an advantage of about 4%.
Table 8: Modality weights for the supervised classification task, calculated by MGE, for the two datasets.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Weight</th>
<th>Modality</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>0.514</td>
<td>SIFT</td>
<td>0.135</td>
</tr>
<tr>
<td>PHOG</td>
<td>0.025</td>
<td>TH</td>
<td>0.580</td>
</tr>
<tr>
<td>GIST</td>
<td>0.207</td>
<td>MMV</td>
<td>0.285</td>
</tr>
<tr>
<td>GB-DIST</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB</td>
<td>0.254</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The modality weights calculated by MGE for this task are presented in Table 8. These weights are representative ones, corresponding to $k = 30$ for the image dataset and $k = 60$ for the image-text-video dataset, while $P = 50$ for both datasets. The presented weights are average weights over 10 random splits of the training/testing data. It can be observed that the modalities with the highest weights are those with the highest accuracies, as depicted in Fig. 20. The proposed method generally assigns higher importance to relatively accurate modalities, although it may also make use of inferior modalities, if this leads to more consistent final graphs, as is the case of the image dataset, where the relatively inaccurate GB modality receives a relatively high weight.

5.5.2 Evaluation in unsupervised object clustering

As a second application of MGE, it has been used as a preprocessing dimensionality reduction step for data clustering. This is an unsupervised learning task, utilizing only the intrinsic characteristics of the data. In the next sections, the dimensionality reduction method and the evaluation metric used are described, followed by the experimental results.

**Dimensionality reduction method** The Locality Preserving Projections (LPP) [50] dimensionality reduction method is used for the application of MGE in this unsupervised setting, similar to [28]. In the LPP method, the affinity matrix is constructed using only intrinsic information of the data, i.e. the similarities among them. The elements $w_{m,ij}$ and $b_{m,ij}$ of the unimodal affinity matrices $W_m$ and the penalty matrices $B_m$, respectively, are
computed as follows:

\[
w_{m,ij} = \begin{cases} 
k_{m,ij}, & \text{x}_i \text{ is among the } k \text{ nearest neighbors of} \\
x_j, \text{ or } x_j \text{ is among the } k \text{ nearest} \\
\text{neighbors of } x_i, \text{ according to } K_m \\
0, & \text{otherwise}
\end{cases}
\]

\[
b_{m,ij} = \begin{cases} 
\sum_{n=1}^{N} w_{m,in}, & i = j \\
0, & \text{otherwise}
\end{cases}
\]

where \( k_{m,ij} \) is the element of \( K_m \) in the \( i \)-th row and \( j \)-th column, and \( c_i, i = 1 \ldots N \) are class labels assigned to the input data. Similar to the LDE method of Section 5.5.1, the \( k_{m,ij} \) values are used as heat kernel values. Hereby, matrix \( B_m \) is equal to the diagonal matrix \( D_m \).

**Quantitative evaluation measure** For the evaluation of the performance of the proposed framework, the low-dimensional data resulting from the dimensionality reduction method are clustered using the k-means clustering method, with the actual number of clusters as input, and then the Rand Index measure, described in Section 5.3.4 is used to quantify the clustering performance, considering the actual labels of the data as the ground truth.

**Quantitative results** Similar to the object recognition task of Section 5.5.1, the MKL framework of [28] has been used to map the data to a low dimensional space, while the use of the average weights of the unimodal affinity matrices of [28] has been compared to the use of the weights as calculated by MGE. A comparison with the unimodal cases is also provided.

Two experiments have been conducted for both the image and the image-text-video datasets, where the effect of the following two parameters on the clustering accuracy has been investigated:

- the number of nearest neighbors used for the calculation of the unimodal affinity matrices (experiment #1) and
- the number of dimensions of the output space, after the dimensionality reduction (experiment #2).
Table 9: Parameters used in the two experiments for the various datasets, in the task of unsupervised object clustering.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Image dataset</th>
<th></th>
<th>Image-text-video dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp. #1</td>
<td>Exp. #2</td>
<td>Exp. #1</td>
<td>Exp. #2</td>
</tr>
<tr>
<td>Dataset size</td>
<td>600</td>
<td>600</td>
<td>1388</td>
<td>1388</td>
</tr>
<tr>
<td>Number of modalities</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Number of classes</td>
<td>20</td>
<td>20</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Nearest neighbors</td>
<td>10-100</td>
<td>30</td>
<td>10-100</td>
<td>30</td>
</tr>
<tr>
<td>Output dimensions</td>
<td>30</td>
<td>10-100</td>
<td>50</td>
<td>10-100</td>
</tr>
</tbody>
</table>

Table 9 summarizes the parameters used in the various experiments for each dataset.

Fig. 21 depicts the experimental results, for the image and the image-text-video datasets. The performance of the multimodal methods is, in general, superior to using the single modalities, especially for the image-text-video dataset. However, in the image dataset, the performance of the GIST and PHOG modalities is better than the MKL method. This can be attributed to the fact that the MKL method uses an average of the unimodal kernel matrices, thus considering modalities with inferior performance, such as the GB-DIST and GB modalities, of equal importance. On the other hand, in the proposed MGE method, the modality weights are calculated adaptively, assigning higher weight to the modalities with the most consistent affinity matrices. Thus, the proposed method manages to compete with the most accurate modalities and to surpass them, in terms of clustering performance.

Considering the image dataset, when the number of nearest neighbors is varied, the proposed method manages to keep an increase of about 3% in performance, compared to MKL, apart from a number of over 80, where the MKL method reaches the performance of the MGE method. A similar increase in performance is apparent when the number of output dimensions varies, with both methods having a decreasing trend with increasing output dimensions. MGE achieves a maximum Rand index of close to 0.94 for 30 dimensions, compared to a maximum of 0.93 for the MKL method. Similar results and a similar increase in the performance of MGE compared to MKL are apparent also from the results of the image-text-video dataset.
Figure 21: Comparison of the Rand Index (RI) for the separate modalities, the MKL and the MGE methods, with respect to (a) the number of nearest neighbors, $k$, used for the affinity matrices, and (b) the number of the output dimensions, $P$.

The modality weights calculated by MGE for this task are presented in Table 10. These weights correspond to $k = 30$ for the image dataset and $k = 60$ for the image-text-video dataset, while $P = 50$ for both datasets. Similar to the supervised classification task, the calculated modality weights roughly correspond to the accuracies of the separate modalities, as depicted in Fig. 21.

As an further illustration of the effect of using the adaptively calculated weights instead of the uniform weights of MKL, Fig. 22 depicts the multi-modal affinity matrices for the two datasets, using the MKL and the MGE methods. The brighter a pixel is, the larger the affinity value in the corresponding edge. For illustration purposes, the data are ordered so that, in the ideal case, the data classes would appear as subsequent white squares on the main diagonal, with the rest of the image being black. In the image dataset, there are 20 classes of equal size, while in the image-text-video dataset there
Table 10: Modality weights for the unsupervised clustering task, calculated by MGE, for the two datasets.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Weight</th>
<th>Modality</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>0.587</td>
<td>SIFT</td>
<td>0.106</td>
</tr>
<tr>
<td>PHOG</td>
<td>0.150</td>
<td>TH</td>
<td>0.627</td>
</tr>
<tr>
<td>GIST</td>
<td>0.123</td>
<td>MMV</td>
<td>0.267</td>
</tr>
<tr>
<td>GB-DIST</td>
<td>0.075</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB</td>
<td>0.065</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

are 13 classes of various sizes. In the MKL method, the multimodal affinity matrix is constructed by considering equal weights for all modalities, while in the MGE method, the automatically calculated weights, presented in Table 10, are used. It is apparent that using the proposed method leads to a clearer depiction of the data classes, which are more clearly differentiated from each other and from the black background. This means that the multimodal affinity matrix of the proposed method is closer to the ground truth semantic class separation, thus leading to more accurate dimensionality reduction.

The application of the MGE approach for clustering has also been compared to the state-of-the-art clustering method of [27], which is a multimodal method for spectral clustering. This method, namely Affinity Aggregation for Spectral Clustering (AASC), also considers multiple affinity matrices and merges them with a weighted sum. However, the weights of the sum are learned together with the clustering coefficients, which leads to an iterative optimization problem. The AASC method has been experimentally evaluated in [27] using the Caltech-101 dataset, and especially a subset consisting of 20 classes, similar to the setting of this thesis. Since the Rand Index measure is not used in [27], the Normalized Mutual Information (NMI) measure [178] has been used instead. NMI is a measure of agreement between two clusterings and is used to measure how much the computed clustering matches the ground truth class separation. The NMI measure takes values in the $[0, 1]$ range, with a value of 1 meaning that the two clusterings are the same, while a value of 0 means that the two clusterings are totally irrelevant to each other. The larger the NMI value, the closer the clustering to the ground truth. For the 20-class Caltech-101 dataset, the AASC approach
Figure 22: The affinity matrices for the unsupervised clustering task, for the image and the image-text-video datasets, using the MKL and the MGE methods. For MKL, equal weights for all modalities are used, while for MGE, the automatically calculated weights, presented in Table 10, are used. Using MGE, the data classes are separated more clearly.

of [27] achieves a NMI value of 0.6458. The proposed MGE approach achieves a NMI value of 0.6597, with the number of nearest neighbors used being 30 and the output dimensionality being 30. It can be seen that the MGE approach achieves a slightly larger NMI value, indicating that the clustering is more accurate than in the AASC method. However, it should be noted that although the AASC method uses 5 modalities, as also is the setting hereby, it uses different types of images features for them than the MGE approach, so the comparison should be made with caution. Nevertheless, the results demonstrate the potential of the MGE approach for clustering, compared to the state-of-the-art.
Table 11: Parameters used in the two experiments for the various datasets, in the task of unsupervised object visualization.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Image dataset</th>
<th>Image-text-video dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp. #1</td>
<td>Exp. #2</td>
</tr>
<tr>
<td>Dataset size</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>Number of modalities</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Number of classes</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Nearest neighbors</td>
<td>10-100</td>
<td>30</td>
</tr>
<tr>
<td>Output dimensions</td>
<td>30</td>
<td>10-100</td>
</tr>
</tbody>
</table>

5.5.3 Evaluation in unsupervised object visualization

Finally, the proposed method has been tested in a data visualization application, in order to present the data as points on a two dimensional screen. Although the proposed method could be used to directly map the data on a two-dimensional space, the two-step procedure followed in [28] has been used, for easier comparison: First, the multimodal data are mapped as points on a low-dimensional space of dimension higher than 2. Then these points are further mapped on the two-dimensional space, using the common Multidimensional Scaling (MDS) [34] method, in order to be visualized on the screen. The following sections describe the dimensionality reduction method used, the evaluation metrics utilized for the assessment of the visualization performance and the experimental results.

Dimensionality reduction method

The Locality Preserving Projections (LPP) [50] method is also used here, as in the clustering task above, as the dimensionality reduction method. The reader is referred to Section 5.5.2 for more details.

Quantitative evaluation measure

In order to compare different visualizations, a measure is needed which quantifies the suitability of a particular data placement. Hereby suitability is defined in terms of how well the visualization conforms to the proximity Gestalt law of visual perception, stating that objects placed close to each other are perceived as belonging to the same class [4].

Hereby, the Dunn index [173], described in Section 5.3.1, has been used
Figure 23: Comparison of the Dunn Index (DI) for the MKL, the multi-objective and the MGE methods, for the image dataset, with respect to (a) the number of nearest neighbors, $k$, used for the affinity matrices, and (b) the number of the output dimensions, $P$. Results are presented for two notions of cluster distance, while for the cluster diameter, the mean distance from the centroid is used.

as an objective evaluation metric for measuring how much objects of the same class are placed close to each other, while objects of different classes are separated.

**Quantitative results** This task is similar to the unsupervised data clustering task of Section 5.5.2, so a similar experimental setting is considered here, by investigating the effect of the number of nearest neighbors and the number of output dimensions on the visualization evaluation metrics, while comparing the MGE method to the MKL method of [28] and the unimodal cases. A comparison with the multi-objective dimensionality reduction method, presented in Section 3 is also provided. The multi-objective method results in many Pareto-optimal solutions, from which one needs to be selected. The selection of one solution is not a trivial task and the design
Figure 24: Comparison of the Dunn Index (DI) for the MKL, the multi-objective and the MGE methods, for the image-text-video dataset, with respect to (a) the number of nearest neighbors, $k$, used for the affinity matrices, and (b) the number of the output dimensions, $P$. Results are presented for two notions of cluster distance, while for the cluster diameter, the mean distance from the centroid is used.

of a strategy for the automatic selection of one solution is out of the scope of this thesis. Exhaustively searching for the one solution that achieves the highest performance when using five modalities, as in this experiment, is also not feasible. Instead, the Pareto solution corresponding to equal weights for all modalities is used. This is a representative solution, considering all the modalities and achieving relatively clear visual results empirically, with respect to class separation. The multi-objective approach does not depend on the number of nearest neighbors, $k$, while it has been tested on two dimensions, thus, it is depicted as a straight line in the subsequent diagrams.

Table 11 summarizes the parameters used in the various visualization experiments for each dataset used. The results of these experiments are presented in Figs. 23 and 24. For the notions of cluster distance and diameter
used, MGE generally leads to superior performance, compared both to MKL and the multi-objective method, indicating that the real data classes are more clearly visible in the visualizations, when the affinity matrix weights are determined through the proposed adaptive method. Similar results are also obtained for other notions of cluster diameter.

As a qualitative evaluation of the proposed method, sample visualizations of the image and the image-text-video datasets are produced and are compared to visualizations produced using the MKL method, as well as the multi-objective method. The resulting visualizations are presented in Fig. 25. For the MKL and MGE methods, the number of nearest neighbors used is $k = 5$, for the image dataset, and $k = 60$, for the image-text-video dataset. The number of output dimensions, before the application of Multi-dimensional Scaling, is, in both datasets, $P = 30$. In Fig. 25, the multimodal data are represented as points on the two dimensional plane, where the colors of the points indicate their true labels. It can be observed that the true classes are more clearly separated in the visualizations resulting from the application of MGE, than with the other methods.

Figure 25: Example visualizations of the image and the image-text-video datasets using the MKL, the multi-objective and the MGE methods.
5.6 Summary

In this section, experiments were presented which evaluate the effectiveness of the multi-objective and the Multimodal Graph Embedding dimensionality reduction approaches. With respect to the multi-objective method, optimization is performed simultaneously for all objectives, which leads to the calculation of low-dimensional mappings where the structure of each separate modality is best preserved. Furthermore, the multi-objective method can discover efficient solutions which a weighted sum-based method cannot find, i.e. those that lie on the non-convex region of the Pareto front. The comparison of the approach to the state-of-the-art shows the applicability of the proposed method for performing dimensionality reduction on real datasets. The proposed approach outperforms the state-of-the-art even in cases where the Pareto front is convex and only solutions corresponding to a weighted sum of the objectives are kept.

The performance of Multimodal Graph Embedding has been tested in a variety of experimental applications and datasets. MGE has been compared to using each modality separately, as well as to the Multiple Kernel Learning (MKL) framework [28], which considers an average weighting scheme for the calculation of the multimodal affinity matrix. It has also been compared to the multi-objective dimensionality reduction method of Section 3. Experiments have been performed where the resulting low-dimensional data are used for supervised object recognition and unsupervised object clustering and visualization, with multimedia datasets of various modalities. The experimental results verify the improved performance of MGE over the other methods, indicating that the modality weights can be effectively determined automatically from the multimodal data and used, within the MGE framework, for dimensionality reduction.
6 Applications of multi-objective dimensionality reduction and multimodal graph embedding

The application area that the proposed multi-objective dimensionality reduction and multimodal graph embedding techniques, described in Sections 3 and 4, were initially intended for was dimensionality reduction of multimedia objects for the purpose of multimedia retrieval and exploration of multimedia databases. However, the proposed techniques are quite generic in nature and can handle any data that can be represented as multimodal objects consisting of modalities in the form of multiple feature vectors. In this section, the techniques will be used in two different application areas, namely accessibility-based search engines and mobile network security. This is to illustrate the diverse fields that the proposed methods can be utilized in.

6.1 Application of multi-objective dimensionality reduction in accessibility-enhanced search engines

In this section, the multi-objective dimensionality reduction method presented in Section 3 will be used in the context of accessibility-based multimedia search engines. The use of search engines for web pages and multimedia is among the most common activities of users in the Web. The majority of search engines consider only the user query as their input and provide a set of results ordered according to their relevance to the query. However, apart from the query, other factors concerning characteristics of the users, can be used as a context, in order to provide results that are personalized and adapted to the specific user. So far, personalization has been addressed by using the search history of the user or of other users [179] [180] [181] [182], or the user’s location [183] [184] [185], in order to provide results that are better fitted to the specific user.

A significant target group for web page design and development is people with disabilities. Designing and evaluating web pages for the disabled is becoming an increasingly important topic. However, most attention so far has been given to developing guidelines for the graphical presentation of web pages and for providing alternative representations for elements, such as images, which are not accessible by all people. Apart from many individual studies that have been conducted and carried out over time [186] [187] [188],
the most significant accessibility related guidelines concerning web-pages, which can also be extended to multimedia data are included and described in Guidelines of Web Content Accessibility (WCAG). However the attention of these guidelines is focused on web pages and on their accessibility as a whole, in terms of their functionality and the easiness with which a user can control, operate or access their content. Few works have addressed the problem of evaluating the accessibility of individual images, in terms of their visual content. In [189], the authors discover areas within images which are inaccessible for people with color-blindness. This is accomplished by simulating the perception of the image by the user and calculating edge differences between the original and the simulated images. In [190] and [191], the content of the images is processed in order to enhance the images, so as to be better accessible for people with color-blindness and decreased contrast sensitivity, respectively. Modifying the functionality of a web page or service, for instance of a search engine, so as to be adapted to people with disabilities, has not been given much focus yet.

In this respect, the application described in this section deals with utilizing multimodal dimensionality reduction in order to re-rank the results returned by an image search engine, so as to promote images that are easier for a visually impaired user to perceive. The goal is to provide an ordering of the resulting images that is personalized to a specific user with specific visual impairments. The impaired user is considered as having a multitude of vision impairments. The image reranking is accomplished by first extracting multiple accessibility scores from the images, denoting their accessibility with respect to a number of supported vision impairments, such as cataract and glaucoma. These accessibility scores are in correspondence to the relevance scores extracted by conventional search engines. However, the procedure followed for their extraction is totally different than calculating relevance scores and, furthermore, multiple accessibility scores are extracted, instead of a single relevance score. Then, the multi-objective dimensionality reduction method is utilized, with appropriately defined objective functions, in order to provide multiple re-rankings of the images. From these re-rankings, one is selected, according to the specific user impairments. The re-ranking of a set of images can be considered as mapping the images on a one-dimensional space, thus the problem can be considered as an instance of multi-objective dimensionality reduction with the target space being of dimensionality equal
to one.

The procedure followed for the multi-objective accessibility-based re-ranking of a set of images is presented in detail in the following sub-sections. The extraction of accessibility-related features is described in Section 6.1.1. The objective functions used for the multi-objective dimensionality reduction method are detailed in Section 6.1.2, while the strategy for the selection of the final ranking from the Pareto-optimal ones is described in Section 6.1.3. Finally, Section 6.1.4 contains the experimental evaluation of the accessibility-based reranking method, which verifies the applicability of multi-objective dimensionality reduction in this application area.

### 6.1.1 Accessibility feature extraction

Accessibility-based reranking of the search results relies on extracting accessibility scores from the images, in analogy to the relevance score of them, as calculated by a standard search engine. For each of the supported vision impairments, an accessibility score \( a_m \in [0, 1] \), \( m = 1 \ldots M \), is extracted from an image, evaluating how accessible this image is for an average person having the respective impairment. The accessibility score of an image takes values in the \([0, 1]\) range, where 0 means that the image is not accessible, while 1 means that it is accessible.

The overall procedure of extracting the accessibility score \( a_m \) of an image \( r \), for impairment \( y_m \) is the following. The impairment \( y_m \) is modeled as a filter \( f_m \), which distorts the original image, according to vision impairment \( y_m \). The filtered image, \( r_m \), is a simulation of how an average impaired user having disability \( y_m \) would perceive the original image \( r \). Next, a comparison is performed between the original and the filtered image, using a distortion measure \( g_m \), in order to quantify the distortion imposed by the impairment filter. The amount of distortion, \( a_m \), is used as an indicator of how accessible this image is for persons having the \( y_m \) impairment. If the image is distorted at a high degree, then much information of the original image is lost, hence the image is not accessible. The amount of distortion can be used as a measure of accessibility. Hereby, \( a_m \) will be normalized to the \([0, 1]\) range, with 0 meaning that the image is not accessible and 1 meaning that it is totally accessible. A graphical overview of the procedure is presented in Figure 26.
Formally, a vision filter $f_m$ is defined as a function

$$f_m : \mathcal{I} \rightarrow \mathcal{I},$$

where $\mathcal{I}$ is the space of images, which takes an image $r$ as its input and produces another image $r_m$, which is the simulation of how a person having impairment $y_m$ would perceive the original image. Depending on the characteristics of impairment $y_m$, the filter $f_m$ may act on the space domain, e.g. by modifying the color of each pixel, on the frequency domain, e.g. by blurring the image with a low-pass filter, or on both domains.

The distortion function $g_m$ is defined as

$$g_m : \mathcal{I} \times \mathcal{I} \rightarrow [0, 1],$$

i.e. it takes two images as its input and produces a number in the range $[0, 1]$, measuring the distortion of the second input image, compared to the first. Various functions can be used as distortion functions, for instance information loss, difference of color histograms or difference of detected edges. Different distortion functions may be suitable for different impairments.

Three impairments will be considered hereby, namely cataract, glaucoma and protanopia (red-green color blindness), in order to illustrate the capabilities of the multi-objective method. However, the approach is easily extensible to any number of impairments, without any increment in complexity being introduced. In the following sections, the filters used for these two impairments will be described, followed by the distortion measures used for them. For the implementation and application of the filters, an average impaired user is considered.
Cataract  Cataract is the deterioration of vision due to the clouding of the lens. The effect of cataract can be analyzed in a series of simpler effects, including the following:

- decreased sensitivity in low contrast changes
- decreased visual acuity (blurriness)
- increased sensitivity to glare sources (bright areas in the image)
- perception of bright clouds (like "cataracts") in the visual field
- yellowing of the image

In [192], filters for various vision impairments are defined. The implemented cataract filter includes all the above effects as separate sub-filters. Herein, the implementation of [192] is used for the glare sensitivity, clouding and yellowing effects. The application of these filters depends on a parameter $x \in [0,1]$, modeling the amount of disability of the user. Hereby, an average impaired user is considered, thus a value of $x = 0.5$ is used.

For the contrast sensitivity and visual acuity effects, a unified approach, different from [192], is followed, so as to produce more realistic results. The approach is based on the Contrast Sensitivity Function (CSF) of the human eye. Contrast is generally defined as the luminance difference between two points in an image, normalized by the average image luminance. There is a threshold in the contrast values that the human eye can see. Luminance differences smaller than this threshold cannot be perceived. However, this threshold varies with the spatial frequency of the luminance source. This variation of the contrast threshold, or usually its inverse, the contrast sensitivity, is described by the Contrast Sensitivity Function (CSF).

The CSF can be formulated as a band-pass filter [193] [194] [191] with a peak at middle frequencies. Adopting the formulation of [191], the contrast sensitivity function for an impaired user, considering middle to high frequencies, can be modeled by the following exponential function:

$$\text{CSF}(u) = (1 - L_c)e^{-0.166u/(1-L_d)},$$  \hspace{1cm} (43)  

where $u$ is the magnitude of the spatial frequency in degrees per visual angle and $L_c \in [0,1]$ and $L_d \in [0,1]$ are parameters modeling the impairment of the user. The $L_c$ parameter models the contrast sensitivity of the user.
Larger values of $L_c$ denote a lower contrast sensitivity, meaning that the user cannot distinguish small differences in intensity. This is modeled by scaling down the magnitude of the CSF in Eq. (43), as $L_c$ grows. The $L_d$ parameter models the visual acuity of the user. Larger values of $L_d$ denote a poorer visual acuity, meaning that the user cannot see details of high spatial frequency. This is modeled by shrinking the CSF function of Eq. (43) along the frequency axis, as $L_d$ grows. For an average user, the values used for $L_c$ and $L_d$ are $L_c = 0.5$ and $L_d = 0.5$.

This type of analysis, with the use of CSFs, incorporates both contrast sensitivity and visual acuity, and is thus closer to the actual user perception. The application of the contrast sensitivity and visual acuity filter can be performed by multiplying the CSF function of Eq. (43) to the magnitude of the frequency spectrum of the image. The complete cataract filter consists of the application of the glare, clouding and yellowing filters, $f_{gcy}$ of [192], followed by the contrast sensitivity and visual acuity filter $f_{csf}$, modeled by the CSF. The application of the cataract filter to an sample input image is as illustrated in Figure 27a.

**Glaucoma**

Glaucoma is the damage of the optic nerve, usually related to increased fluid pressure in the eyes. The effect of glaucoma on the human vision is the perception of a dark area around the center of vision. The size and intensity of the dark area depends on the severity of the impairment. The implementation of [192] is used herein for the glaucoma filter. As in the cataract filter, the amount of disability is modeled by a parameter $x \in [0, 1]$, where a value of 0 means that the person does not have the impairment at all, while a value of 1 means that the person has the impairment in the largest amount. Fig. 27b illustrates the effect of applying the glaucoma filter on an example image.

**Protanopia**

Color blindness is the inability to see adequately or at all a region of the visible light spectrum, either the long wavelengths (protanopia), the middle wavelengths (deuteranopia), or the short wavelengths (tritanopia). In [192], filters for all these types of color-blindness are defined and used. The same filters are also used in this research. The filters are implementations of the color-blindness simulators described in [195].

The application of the color blindness filters depends on a parameter
Figure 27: Application of the filters for the supported impairments on an example image. (a) Application of the cataract filter. The cataract filter consists of the glare, clouding and yellowing filter of [192], followed by the contrast sensitivity and visual acuity filter based on the CSF function. (b) Application of the glaucoma filter. (c) Application of the protanopia filter.

\[ x \in [0, 1] \], modeling the amount of disability of the user. A value of \( x = 0 \) denotes that the user does not have color blindness, while a value of \( x = 1 \) that the user has the impairment (protanopia) in the highest degree. Values in the range \((0, 1)\) model mild impairments, where the user can partly see the corresponding range of the spectrum (protanomaly). For an average impaired user, a value of \( x = 0.5 \) is used.

Without loss of generality, only protanopia will be considered hereby; however, other color-related impairments, such as deuteranopia and tritanopia, can be implemented similarly. In Figure 27c, the application of the protanopia filter on a sample image is depicted.

**Distortion functions**  After the calculation of the filtered image, for either of the three impairments, a distortion function is used to compare the original and the filtered image. The distortion function measures the information loss caused by the filtering procedure. The output of this function is a number in the range \([0, 1]\), which constitutes the accessibility score of the image. Small
distortion means better accessibility of the original image, while distortion close to 1 means that the original image is perceived highly distorted, thus it is not accessible.

The distortion function used for all impairments is the sum of three distortion measures, measuring differences in the color histogram, the detected edges and the pixel-by-pixel color values of the images.

For the difference of histograms, the images are first transformed to the $Luv$ color space and then only the luminance ($L$) channel is used. The luminance histograms $h$ and $h'$ of the original image $r$ and the filtered image $r'$, respectively, are then constructed, quantizing the luminance values in $b$ bins. Let $h_i \in [0,1]$, $i = 1 \ldots b$, be the value of the histogram in the $i$th bin. The histogram values are normalized so that $\sum_{i=1}^{b} h_i = 1$. Similarly for $h'$. Then, the histogram difference distortion function, $g_{\text{histogram}}$, is defined as the Euclidean distance between $h$ and $h'$:

$$g_{\text{histogram}}(r, r') = \sqrt{\sum_{i=1}^{b} (h_i - h'_i)^2}$$

For the implementation of this function, the histograms have been considered to consist of 64 bins, i.e. $b = 64$.

For the difference in the edges detected in the original image $r$ and the filtered one $r'$, the images are again transformed to the $Luv$ color space and only the luminance channel is considered. Then, the Sobel edge detection operator is applied to both of them, producing images $r_s$ and $r'_s$, respectively. Images $r_s$ and $r'_s$ are grayscale images, where the intensity of each pixel, normalized to the $[0,1]$, indicates if there are edges at this position in the original and filtered images. The closer the intensity is to 1, the sharper an edge exists in the respective initial image.

The amount of edges in the original and filtered images can be approximated by summing the intensity values of the images produced after edge detection. Let $r_{s,i,j}$ be the intensity of the pixel in the $i$th row and the $j$th column of image $r_s$, and similarly for $r'$. Let also $s$ and $s'$ be the amount of edges in images $r$ and $r'$, respectively. Then

$$s = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} r_{s,i,j},$$

$$s' = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} r'_{s,i,j}.$$
where $H$ and $W$ is the height and the width of the images, respectively.

The edge-related distortion function, $g_{\text{edge}}$, can then be defined as the normalized difference between $s$ and $s'$:

$$g_{\text{edge}}(r, r') = \frac{s - s'}{s}. \quad (44)$$

The third distortion function measures the pixel-by-pixel difference in color between the original and the filtered images. Similarly to the previous functions, the distortion function $g_{\text{pixel}}$ takes two images $r$ and $r'$ as its input. Let $r_{i,j,R}$, $r_{i,j,G}$ and $r_{i,j,B}$ be the $R$, $G$ and $B$ color components of the pixel in the $i$th row and $j$th column of image $r$, and similarly for $r'$. Let also $d_{i,j}^2(r, r')$ be the squared color difference between the pixels in the $(i, j)$ position in the $r$ and $r'$ images:

$$d_{i,j}^2(r, r') = (r_{i,j,R} - r'_{i,j,R})^2 + (r_{i,j,G} - r'_{i,j,G})^2 + (r_{i,j,B} - r'_{i,j,B})^2$$

The distortion function is defined as follows:

$$g_{\text{pixel}}(r, r') = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} c_{t,i,j}(r, r') d_{i,j}^2(r, r'), \quad (45)$$

where $H$ and $W$ are the height and the width of the image, respectively, in pixels,

$$B = \sum_{i=1}^{H} \sum_{j=1}^{W} c_{t,i,j}(r, r')$$

is a normalization constant, and

$$c_{t,i,j}(r, r') = \begin{cases} 1, & \text{if } d_{i,j}^2(r, r') > t \\ 0, & \text{otherwise} \end{cases}$$

is a parameter introduced to keep only the differences that are larger than a threshold value $t$. This thresholding has been introduced in order to ignore small differences in color that are due to noise introduced by the filter.

Finally, the total distortion function, $g$ is defined as the weighted sum of $g_{\text{histogram}}$, $g_{\text{edge}}$ and $g_{\text{pixel}}$:

$$g(r, r') = w_{\text{histogram}} g_{\text{histogram}}(r, r') + w_{\text{edge}} g_{\text{edge}}(r, r') + w_{\text{pixel}} g_{\text{pixel}}(r, r'), \quad (46)$$

$$w_{\text{histogram}}, w_{\text{edge}}, w_{\text{pixel}} \in [0, 1], \quad w_{\text{histogram}} + w_{\text{edge}} + w_{\text{pixel}} = 1.$$
final distortion function. For the implementation presented in Section 6.1.4, the weights have been experimentally set to \( w_{\text{histogram}} = 0.2 \), \( w_{\text{edge}} = 0.45 \) and \( w_{\text{pixel}} = 0.35 \). These specific values were selected so that the resulting image accessibility scores of a set of images match the ground truth scores, determined as presented below in Section 6.1.4, as much as possible.

### 6.1.2 Objective functions used in multi-objective dimensionality reduction

Using the accessibility score extraction procedure described in Section 6.1.1, accessibility scores for each image of the search results can be calculated for the supported vision impairments. Using these scores, the images of the result set can be ordered according to their accessibility for each of the impairments. In a system supporting only one type of impairment, this ordering would be sufficient. However, the system described hereby supports users having more than one vision impairments at the same time. In this case, ordering the results according to a single impairment is not adequate.

Thus, multimodal techniques can be applied to handle this problem, hereby the proposed multi-objective dimensionality reduction technique described in Section 3. The problem of linear ranking can be considered as mapping the images on a one-dimensional space, thus it can be considered as an instance of multi-objective dimensionality reduction, with the target space being of dimensionality one. Although the same force-directed objective functions used in Section 3 can also be used hereby, reduced to one dimension, the fact that the output dimensionality is one allow the use of different objective functions, targetted specifically at the problem of re-ranking.

The discounted cumulative gain (DCG) is used for the definition of the objective functions. The DCG is commonly used for the evaluation of the effectiveness of the rankings produced by search engines. Considering that a set of \( N \) results is ordered, so that result \( r_i \) is in position \( i \), with 1 being the top most position, the DCG is calculated as follows:

\[
DCG = \text{rel}(r_1) + \sum_{i=2}^{N} \frac{\text{rel}(r_i)}{\log_2 i},
\]

where \( \text{rel}(r_i) \) is the relevance score of the result \( r_i \). The larger the cumulative gain, the more promoted (i.e. placed in higher positions) are results that are relevant to the query.
The DCG is used hereby as an objective function for evaluating the orderings of the results, only that the relevance score is replaced with the accessibility scores for the various supported impairments. If \( p = (r_1, r_2, \ldots, r_N) \) is a variable representing a specific ordering of the \( N \) results returned by the search engine, then the three objective functions used in the multi-objective setting are the following:

\[
J_{\text{cataract}}(p) = 1 - \frac{1}{N} \left( a_{1,\text{cataract}} + \sum_{i=2}^{N} \frac{a_{i,\text{cataract}}}{\log_2 i} \right),
\]

\[
J_{\text{glaucoma}}(p) = 1 - \frac{1}{N} \left( a_{1,\text{glaucoma}} + \sum_{i=2}^{N} \frac{a_{i,\text{glaucoma}}}{\log_2 i} \right),
\]

\[
J_{\text{protanopia}}(p) = 1 - \frac{1}{N} \left( a_{1,\text{protanopia}} + \sum_{i=2}^{N} \frac{a_{i,\text{protanopia}}}{\log_2 i} \right),
\]

where \( a_{i,\text{cataract}}, a_{i,\text{glaucoma}} \) and \( a_{i,\text{protanopia}} \) are the specific accessibility scores for cataract, glaucoma and protanopia, respectively, for result \( r_i \) in the ordering \( p \). The DCG has been normalized by the number of results \( N \) and subtracted from 1, in order for the optimal ordering to be calculated by minimizing the objective functions, instead of maximizing them.

Using the above three objective functions, all possible rankings of the results can be evaluated for each of the three impairments and the Pareto-optimal ones can be calculated using multi-objective optimization techniques. Hereby, the SPEA2 genetic algorithm [156] is used for the calculation of the Pareto front. After the calculation of the Pareto-optimal rankings, one of them needs to be selected, in order to be presented to the user.

### 6.1.3 Ranking selection using the impairment profile

For the selection of one of the Pareto-optimal solutions calculated with the methodology of the previous sections, information about the specific amount of disability of the user in each of the supported impairments needs to be considered. This information is encoded in a vector of values \( x \), called the **user impairment profile**:

\[
x = (x_1, x_2, \ldots, x_M), \quad x_m \in [0, 1], \quad m = 1 \ldots M.
\]

Each value \( x_m \) of the impairment profile is a measure of the disability of the user in each of the \( M \) supported impairments. The measure of disability
is in a scale from 0 to 1, where 0 means that the user has the respective impairment at the highest degree, while 0 means that the user does not have the impairment.

For the three hereby described impairments, cataract, glaucoma and protanopia, the user profile contains three values:

\[ x = (x_{\text{cataract}}, x_{\text{glaucoma}}, x_{\text{protanopia}}). \]

Using the values of the impairment profile as coordinates, the user impairment profile \( x \) can be positioned in the same space as the Pareto-optimal rankings. This allows the selection of one of the Pareto-optimal rankings, one that is closer to the user impairment profile. The Chebyshev distance between the profile and the points of the Pareto front is used for this purpose. The Chebyshev distance is commonly used in achievement function-based scalarization methods in multi-objective optimization [196], where the distance of the solutions to a reference point is utilized. This resembles the hereby profile-based selection approach, thus the Chebyshev distance is used hereby as well.

The Chebyshev distance between two vectors \( x = (x_1, x_2, \ldots, x_M) \) and \( y = (y_1, y_2, \ldots, y_M) \) is defined as:

\[ d_{CH}(x, y) = \max_m(|x_m - y_m|). \]

Let \( J(p) = (J_{\text{cataract}}(p), J_{\text{protanopia}}(p)) \). The selected ranking, \( p_{\text{opt}} \) is calculated as follows:

\[ p_{\text{opt}} = \arg \min_p d_{CH}(x, J(p)). \]  \hspace{1cm} (51)

After the ranking that is closest to the user profile has been selected, it can finally be presented to the visually impaired user.

### 6.1.4 Experimental evaluation

The experimental evaluation of the accessibility-based multi-objective reranking method has been performed by utilizing both artificial and real impaired users, using two different image datasets. Experimentation with artificial users allows a more controlled environment and a qualitative assessment of the method, while comparing with the perception of real users allows to assess it in real-world scenarios and quantify the results.
Qualitative evaluation with artificial users As a first experiment for
the evaluation of the accessibility-based reranking method, a dataset of 14820
images of Italian monuments, collected from Flickr, as part of the CUbRIK
project [197], was used. Each image is associated with textual information,
in the form of a title and tags, which can be used by a text-based search
engine for image retrieval. A Solr-based search engine was used for image
search and retrieval. The 10 top results are considered for accessibility-based
reranking. For this use case, cataract and protanopia have been used as the
supported impairments.

In Figure 28, a set of rerankings of the results of an example query
are presented, for three artificial users: one having cataract (b), one hav-
ing protanopia (c) and one having both cataract and protanopia (d), with
\( x_{\text{cataract}} = 0.5 \) and \( x_{\text{protanopia}} = 0.3 \). As a query, the word “palace” is sub-
mitted.

The first row (a) shows the the original ranking of the results, as returned
by the text-based search engine, ordering the results according to their rel-
relevance to the query. Below each image, the accessibility scores extracted
from it for cataract and protanopia, using the methodology of Section 6.1.1,
are presented.

For each user, Figure 28 depicts various rankings of the results, with the
ordering going from left (top results) to right. The first row in each user
is the accessibility-based ranking of the results, computed using the multi-
objective re-ranking method, with the accessibility scores of the images and
the values of the user impairment profile. For demonstration purposes, the
second row contains a simulation of how the user perceives the results of
the first row. As the list goes from left to right, the results are harder to
perceive. This means that the ranking indeed promotes results which are
easier to see by a vision-impaired user.

Evaluation with real users As a second round of experiments, user stud-
ies have been performed with real visually-impaired users, in order to collect
ground truth data and evaluate the accessibility score extraction procedure
and the multi-objective ranking. For this experiment, all three impairments,
namely cataract, glaucoma and protanopia, were considered.

For the collection of the ground truth data, a web-based tool has been
implemented, through the use of which, impaired users are able to assess
the visibility of various sets of images and to submit appropriate rankings of them. In each user session, the user is presented with a set of 10 images, in random order. The images are randomly picked from a set of 100 fashion-
related images, taken from the Fashion dataset of the CUbRIK project [197]. As a first task, the user is requested to put the images in order, from the one which is most easy for them to see to the one which is most difficult to see. Whether an image is easy or difficult to see is left to the users perception, without the researcher providing any clues about items that may exist in the images, but are not seen by the user.

Once the user submits the ordering of the images, the second phase of the experiment follows. In this phase, the users are requested to check whether they see or not a number of visual characteristics appearing in the images, such as specific objects or colors. The visual characteristics for each image have been gathered from the manual annotation of the images by a number of users with full vision. The images for which the users are requested to check the visibility of the characteristics are the same as the ones which the users ordered in the previous phase of the experiment, and are presented to the users one after the other, in the same order as they had put them.

When the user completes the second phase of the experiment for all 10 images, one user session ends. Each user was requested to participate to five of the above two-phase sessions, each with another random set of 10 images. The images in each session are results of five fashion-related textual queries, namely “hat”, “jeans”, “shirts”, “shoes”, “skirt”.

In the experiments, 15 visually-impaired users have participated. The users were patients of the ophthalmological clinic of AHEPA hospital in Thessaloniki, Greece and of the Social Insurance Institute of Neapoli, Thessaloniki. The number of users is relatively small due to the difficulty in finding visually-impaired users for the purposes of collecting ground truth data. Of these users, 10 were women and 5 were men, while their ages ranged from 62 to 83 years old. Most patients suffered from glaucoma, in some cases along with cataract and protanopia. Two of the patients suffered only from cataract and one only from protanopia. A user impairment profile has been created for each user, corresponding to his/her impairments and their severity.

Each patient was requested to use the ground truth collection tool for five sessions. Each session corresponded to a sample query submitted to the CUbRIK search engine. In particular, text queries, such as “jeans” and “shoes” were submitted to the search engine. The top ten resulting images were used as the images presented to the user in one session of the
ground truth collection tool, in the order returned by the search engine, i.e. by decreasing relevance score. This ordering is hereby referred to as the “relevance ranking”.

For each image $i$ presented in the user sessions, a ground truth accessibility score $a_{i,gt}$ was calculated, based on the visual characteristics that the user checked as visible. The accessibility score was calculated as the ratio of the visible characteristics over the total number of visual characteristics existing for each image:

$$a_{i,gt} = \frac{n_{\text{visible},i}}{n_{\text{total},i}}$$  \hspace{1cm} (52)

The images of a session were provided as input to the multi-objective ranking method, and a set of Pareto-optimal rankings was calculated, considering all the supported impairments (cataract, glaucoma and protanopia). In order to select one of the rankings, a user profile corresponding to the impairments of each user and their amounts was created and used with the Chebysev distance-based selection strategy. The resulting ranking, hereby denoted as “multi-objective ranking”, was compared to the relevance-based ranking, in order to assess which is closer to the user perception. The user perception has been encoded in the ground truth accessibility scores of the images, so these scores are used in the comparison, as described below.

In order to compare the two rankings, the commonly used Discounted Cumulative Gain (DCG) metric has been used. Given a ranked list of $N$ search engine results, each at position $i$ and with a relevance score $\text{rel}_i$, DCG is a means to quantify whether the top results in the list have high relevance scores or not. The traditional DCG is calculated as in Eq. (47). The larger the DCG, the more accurate the ranking of the results, according to the relevance scores. Hereby, instead of the relevance scores $\text{rel}_i$, the ground truth accessibility scores have been used:

$$\text{DCG} = a_{1,gt} + \sum_{i=2}^{N} \frac{a_{i,gt}}{\log_2 i}$$  \hspace{1cm} (53)

For each session, the DCG evaluation metric has been used to compare the relevance ranking to the automatic ranking. The results for an example session are presented in Table 12. The query used for this example is “shirts” and they correspond to a user with a large amount of glaucoma and a small cataract. The relevance and automatic rankings are presented. The score columns contain the ground truth accessibility scores of the images. The
desired promotion of results with high accessibility scores is apparent both from the result positions in Table 12 and from the comparison of the DCG value, which is larger for the automatic ranking.

Table 12: Example of the relevance-based and multi-objective rankings for a user session.

<table>
<thead>
<tr>
<th>position</th>
<th>relevance</th>
<th>multi-objective</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>image ID</td>
<td>score</td>
</tr>
<tr>
<td>1</td>
<td>97</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>130</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>101</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>121</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>148</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>99</td>
<td>0.50</td>
</tr>
<tr>
<td>7</td>
<td>140</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>142</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>152</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>146</td>
<td>1.00</td>
</tr>
</tbody>
</table>

DCG 3.13 3.66

An illustration of the above rankings is presented in Fig. 29. The use of the accessibility filtering promotes images which are easier for a person with glaucoma and cataract to see, i.e. images having higher contrast and more vivid colors. For comparison, the bottom part of Fig. 29 contains the ground truth ranking of the images, according to the user annotations.

An average DCG value of 3.62 has been measured for the relevance ranking, averaged over all sessions. Using the automatic ranking, calculated by the accessibility filtering pipeline, an average DCG value of 4.09 has been measured, being by 0.47, or 12.57%, larger than the DCG of the relevance-based ranking. This verifies that the use of the Pareto-based ranking procedure, using the automatically extracted accessibility scores, for the various impairments, leads to rankings which are closer to the perception of the impaired users.
Figure 29: Illustration of the rankings of Table 12. (a) The results of the search engine, using “shirts” as the query, ranked according to their relevance to the query. (b) The images are automatically ranked using the multi-objective method, for a user with glaucoma and cataract. Images with vivid colors and sharp edges, such as image 1, which are easier for the user to see, have been promoted. (c) The images are ranked according to their ground truth accessibility scores for this user. Images with vivid colors and sharp edges have been put at the top positions.

6.2 Application of multi-objective and multimodal graph embedding dimensionality reduction in mobile network security

In this section, the proposed multi-objective and the multimodal graph embedding methods for dimensionality reduction, presented in Sections 3 and
4 will be used in another application area, this time in the context of mobile network security. Specifically, they will be used for visualizing, in two dimensions, behavioural aspects of the users of a mobile network, with the purpose of detecting unusual activity.

In an effort to identify and prevent malware infections and threats in mobile phone networks, data visualization can be of significant assistance to the network administrator. There are numerous network parameters to be visualized and numerous means to visualize them [198]. Recent attempts, such as the NEMESYS project [199], have indeed focused on utilizing data analytics and visual analytics techniques for anomaly detection and network security [200] [201]. Most existing network visualization methods, developed for security purposes, visualize communication patterns between network components, or specific characteristics of software activity, such as port usage [198].

The diversity of the malware types and signatures, as well as the multitude of manners in which they act towards specific users or towards the network, render the problem of infection identification as a very challenging one [202]. Behaviour-based approaches can be promising [203] [204], since they are based on features describing the behaviour of the malicious applications, such as the amount of traffic or the number of connected users to a host, and how it differs from normal traffic [205]. Behavioural aspects are used as the central characteristics in a number of techniques, most employing graphs, affecting the positioning of the visualized elements [206]. Behavioural features can be extracted from data which are available to the network administrator, such as Call Detail Records (CDRs). The idea is to extract behavioural features from the raw data and then to calculate distances or similarities among users, hosts etc., thus constructing similarity graphs which can then be visualized on a screen, and assist the analyst in identifying clusters of malicious behaviour. This approach has been followed, for instance in works such as [207], [208] and [209].

Most existing works can visualize only a single attribute of behaviour. However, multiple diverse features can be extracted from the raw data, describing various behavioural aspects. For instance, CDR records contain information about the caller and the recipient, the duration of the call etc. The combination of these multiple features via multimodal fusion techniques for visualization can provide deeper insights in the function of the network
than using each feature independently. Multimodal dimensionality reduction techniques, although having been used in the literature in application fields such as search engines and multimedia processing, have not yet been used in the context of mobile network security.

In this respect, the multimodal dimensionality reduction methods proposed in this thesis can be used in order to handle the multiple behavioural modalities extracted from the raw data and produce two-dimensional mappings of behaviours, in which the separation between normal and abnormal behaviour can be visually detected with ease. Two Denial-of-Service attack use cases will be considered as example applications. After a description of the utilized datasets, in Section 6.2.1, and the behavioural descriptors extracted from them and used as modalities, in Section 6.2.2, the two-dimensional mappings of the datasets, produced by the multi-objective and the multimodal graph embedding dimensionality reduction methods, will be presented in Sections 6.2.3 and 6.2.4, in order to verify the applicability of the proposed methods in this application field.

6.2.1 Mobile network datasets used

The proposed methods are evaluated in two Denial-of-Service attack use cases. In the first scenario, a number of mobile phones is considered to be compromised by a malware that automatically and arbitrarily sends large numbers of SMS messages to the numbers of the phone user’s contact list. In the second scenario, a number of mobile phones is considered to be compromised by malware that performs attacks using the Radio Resource Control (RRC) communication protocol. Using these datasets, the goal of two-dimensional dimensionality reduction is to visually group users with similar behaviour and separate the users involved in the attacks from the users operating normally. The datasets used for these two scenarios are presented in the following sub-sections.

SMS flood attack dataset  The data for the SMS flood attack scenario have been generated using GEDIS Studio [210], which is an online tool for generating CDR data. The generated data correspond to the scenario of [211]. Without loss of generality, a small network consisting of one cell tower is considered. A number of 4800 users, connected to the cell tower, have been simulated. According to [211], the normal SMS traffic rate for one
network sector is 0.7 messages/sec. For 4800 users, this corresponds to 12.6 messages/day/user. For demonstration purposes, users were divided into two normal usage groups, one with a large rate of 17.79 messages/day/user (2000 users) and one with a small rate of 8.89 messages/day/user (2800 users), with a total average of the original 12.6 messages/day/user. A period of 7 days has been simulated. A 10% ratio of the first group of users and a 5% ratio of the second group of users constitute the mobile devices which have been infected by the SMS flooding malware, during this period. In the last hour of the 7th day, the infected devices cause an SMS flood, by increasing their rate of SMS traffic by 8 times their normal rates, according to the setting of [211]. The final behavioural groups are described in Table 13.

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Number of users</th>
<th>SMS sending rate (messages per day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1800</td>
<td>17.79</td>
</tr>
<tr>
<td>2</td>
<td>2660</td>
<td>8.89</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>17.79 on the first 6 days 142.3 on the 7th day</td>
</tr>
<tr>
<td>4</td>
<td>140</td>
<td>8.89 on the first 6 days 71.12 on the 7th day</td>
</tr>
</tbody>
</table>

**RRC attack dataset** In the RRC attack scenario, a number of mobile phones is considered to be compromised by malware which exploits the RRC communication protocol, which manages the bandwidth in a mobile network. A Denial-of-Service attack is caused by sending an increased amount of RRC packets. The dataset has been produced by the mobile network simulator developed in the context of the NEMESYS project [199] and is described in [212]. There are 200 phones, 100 of which are involved in the attack: 50 of them are less aggressive, sending periodic packets whenever the user is inactive, and the other 50 are very aggressive, triggering a DCH attack whenever a demotion is detected. The behavioural groups of this dataset are described in Table 14.
Table 14: User groups of the RRC attack dataset.

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Number of users</th>
<th>Type of behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>Less aggressive, sending periodic packets when user is inactive.</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>More aggressive, triggering DCH attack when demotion is detected.</td>
</tr>
</tbody>
</table>

Table 15: Record attribute types.

<table>
<thead>
<tr>
<th>$k$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td>The type of a recorded event, e.g. “VOICE CALL” or “SMS”.</td>
</tr>
<tr>
<td>hour</td>
<td>The hour that an event, e.g. a call or an SMS message, occurred.</td>
</tr>
<tr>
<td>from</td>
<td>The ID of the user from whom the event originated, e.g. the user that started a call or sent an SMS message.</td>
</tr>
<tr>
<td>to</td>
<td>The ID of the user to which an event is directed, e.g. the recipient of a call or SMS message.</td>
</tr>
<tr>
<td>length</td>
<td>The length of a transmitted packet (e.g. an RRC protocol packet) in bytes.</td>
</tr>
</tbody>
</table>

6.2.2 Extracted behavioural descriptors

The raw CDR data have been used to construct histograms of various attributes, which are used as behavioural descriptors of the mobile network users. The raw data are considered as a set $\mathcal{R}$ of records $R \in \mathcal{R}$. As an example, in case of Call Detail Records, a record $R$ could be a phone call or the transmission of an SMS message. Each record $R$ is itself a set of attributes, $r_k$: $R = \{r_k, \ k \in \mathcal{A}\}$, where $\mathcal{A}$ is the set of all attribute types existing in the raw data. Each attribute is a specific piece of information regarding a record, such as, in case of a phone call, the ID of the caller, or the duration of the call. The attribute types that are used in the descriptor definitions below are presented in Table 15. For instance, a record could consist of the following attributes: $r_{\text{type}} = \text{“SMS”}$, $r_{\text{from}} = 0001001431$, $r_{\text{to}} = 0001001532$, $r_{\text{hour}} = 16$. 

156
A histogram $h$ is considered as a vector $h \in \mathbb{R}^D$:

$$h = (h_1, h_2, \ldots, h_D). \quad (54)$$

A histogram is constructed based on the values for a specific attribute. The histogram is split into $D$ equal-sized bins, covering the range of the possible values for the associated attribute. The value of the histogram for the $i$-th bin, $h_i$, is defined as:

$$h_i = |\{R \in \mathcal{R} \cap \mathcal{C} : r_k \in \text{bin}_i\}|, \quad (55)$$

where $|A|$ denotes the cardinality of set $A$, $\mathcal{C}$ is a set of records satisfying specific constraints that may be needed for the construction of the histogram, such as keeping only those SMS messages that are sent towards premium numbers, $k$ is the associated attribute type and bin$_i$ denotes the set of values in the range of the $r_k$ attribute that constitute the $i$-th bin.

The histogram attribute and the constraint set for a specific descriptor are selected so that the histogram captures a specific aspect of user behaviour. For instance, a histogram of the times within the day that a user sends SMS messages is different between normal usage, where the distribution is generally random, and abnormal usage, where large peaks may be observed at specific times. For the attack scenarios considered hereby, the following three behavioural descriptors are extracted from the datasets of Section 6.2.1:

- **Time Histogram Descriptor (THD),** which concerns the times that users send SMS messages,

- **Correspondents Histogram Descriptor (CHD),** which concerns the correspondents of the SMS messages,

- **Length Histogram Descriptor (LHD),** which concerns the RRC packet lengths exchanged between users.

The three descriptors and details for their construction are described in Table 16. The form of each descriptor differs between normal and abnormal behaviour, as depicted in Fig. 30, so that each has a potential to distinguish normal from abnormal behaviour.

The extraction of all descriptors depends on the selection of a time period within which the network activity is considered. This time period can be
Table 16: Behavioural descriptors used in the network security application.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Histogram Descriptor</td>
<td>( D = 24 )</td>
<td>A histogram of the hours of the day at which a user sends SMS messages.</td>
</tr>
<tr>
<td>( \text{bin}_i = { R \in \mathcal{R}</td>
<td>r_{\text{hour}} = i } )</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{C} = { R \in \mathcal{R}</td>
<td>r_{\text{type}} = \text{“SMS” AND } r_{\text{from}} = u } )</td>
<td></td>
</tr>
<tr>
<td>Correspondent Histogram Descriptor (CHD)</td>
<td>( D ) is equal to the number of contacts of each user. ( \text{bin}_i = { R \in \mathcal{R}</td>
<td>r_{\text{to}} = c_i } ) ( \mathcal{C} = { R \in \mathcal{R}</td>
</tr>
<tr>
<td>Length Histogram Descriptor (LHD)</td>
<td>( D = B/T )</td>
<td>A histogram of the packet lengths that have been transmitted from the user for the time period of interest.</td>
</tr>
<tr>
<td>( \text{bin}_i = { R \in \mathcal{R}</td>
<td>r_{\text{length}} \in { T \text{i}, \ldots, T(i+1) - 1 } } )</td>
<td>( B ) is the maximum packet length, in bytes. ( T ) is the size of each bin, in bytes.</td>
</tr>
<tr>
<td>( \mathcal{C} = { R \in \mathcal{R}</td>
<td>r_{\text{type}} = \text{“PACKET” AND } r_{\text{from}} = u } )</td>
<td></td>
</tr>
</tbody>
</table>

pre-defined, or it can be interactively selected by the mobile network operator. Thus, the operator, by selecting different time periods, can control the descriptors used and have an overview of the evolution of the attack through time.

6.2.3 Experimental results for the SMS flood attack scenario

The mobile phone users of the network have been considered as the multimodal objects. The CDR records generated by GEDIS Studio have been used to extract multiple descriptors for the users, describing various aspects of user behaviour. These descriptors have been used as the multiple modalities of the multi-objective and the multimodal graph embedding dimensionality reduction methods, with the target dimensionality being 2, in order for the users to be visualized as points on the two-dimensional screen. For this scenario, the THD and CHD descriptors of Section 6.2.2 have been used.

Multi-objective dimensionality reduction

Fig. 31 depicts the data corresponding to day 7 of the GEDIS dataset, which is the day of the attack, visualized using the multi-objective dimensionality reduction method. The colors represent the different ground truth user groups, assigned for demonstration purposes. The blue color represents
Figure 30: Examples of the behavioural descriptors used in the network security application. (a) Time Histogram Descriptor (THD), (b) Correspondent Histogram Descriptor (CHD), (c) Length Histogram Descriptor (LHD). The top row depicts examples of normal behaviour, while the bottom row examples of abnormal behaviour.

Users with high usage rate, while the red color users with low usage rate. The fainter blue and red colors represent those subsets of the two user groups which are infected by the SMS-flooding malware.

Figs. 31(a) and (b) depict the mappings produced using the two user descriptors separately. The ground truth user groups are not clearly separated in either of these figures. Fig. 31(c) illustrates the combination of both descriptors, with a multi-objective mapping corresponding to a solution from the center of the Pareto front. In this image, the four groups are visually apparent. There are two large circular clusters corresponding to the two user groups of normal usage, i.e. groups 1 and 2. A small cluster of points is separated from each large circular cluster, corresponding to the infected users, as confirmed by the ground truth colors. This separation indicates that these users exhibit very different behaviour than the rest of the users, allowing the analyst to detect the SMS flooding attack.

Fig. 32 depicts the multi-objective mappings of each day contained in the dataset, separately. In the first six days, two circular clusters are formed,
Figure 31: Two-dimensional mappings of the SMS flood attack dataset using the multi-objective dimensionality reduction method, for the 7th day of the simulation (day of the attack). (a) Mapping using the THD descriptor. (b) Mapping using the CHD descriptor. (c) Mapping using both descriptors, corresponding to a solution from the center of the Pareto front.

denoted by the dashed circle, representing the two groups of users with high and small SMS rates. The separation is confirmed by the ground truth class colors. In the 7th day, a number of users are suddenly separated from the circular groups, indicating unusual behaviour. As indicated by the ground truth colors, the users with the unusual behaviour are those which have been infected by the SMS-flooding malware. Thus, the mobile network operator, by selecting different time periods can have an overview of the evolution of the attack.

Moreover, the operator can interactively select different Pareto-optimal solutions of the multi-objective method, in order to focus on different behavioural aspects. The mappings of Fig. 33 are produced using the data of day 7, i.e. the same data as in Fig. 31, but using various Pareto optimal solutions, favoring one or the other modality. When a solution favors the THD descriptor, the infected users tend to be more separated from the larger clusters and form a common cluster of infected users between them. On the other hand, when the solution favors the CHD descriptor, the infected clusters tend to be closer to the normal clusters and separated from each other. This means that the THD descriptor is more suitable to discriminate the infected users from the rest of the users, while the CHD descriptor
Figure 32: Multi-objective mappings for each of the 7 days of the period covered by the SMS flood attack dataset. In the first six days, two separate circular clusters are identified, representing the two clusters of normal usage. This pattern is repeated in all the first six days, denoted by the dashed circles. In the 7th day, the infected users are separated from the circular pattern.

It should be noted that a single solution of the Pareto front could also be selected automatically, as in the accessibility-related application of Section 6.1. This would require the existence of a profile that encodes the preferences of the operator with respect to the various aspects covered by the multiple descriptors. For instance, the operator could assign more importance to the time-related modality, in order to detect anomalies, if he/she knew that, at a certain occasion, the mobile phone users are expected to communicate with diverse numbers of correspondents, but not at specific times of day. In other words, the differences with respect to time are more important than the differences with respect to the number of correspondents. Specifying a preference profile in this mobile network-related use case is not as straightforward as in the accessibility case, where the profile was determined by the user’s impairments. In the mobile-related case, the preferences may differ depending on the occasion, so that a set of profiles may be more appropriate, from which the operator can select, according to the occasion. The specification of such a set of profiles goes beyond the purpose of this thesis, although it could be an interesting subject for further research.
Figure 33: Two-dimensional mappings of the SMS flood attack dataset for the 7th day of the simulation, using the multi-objective dimensionality reduction method. Various Pareto-optimal solutions are presented, ranging from ones favoring the THD modality, towards the left side, to ones favoring the CHD modality, towards the right side.

Multimodal Graph Embedding dimensionality reduction

The projection of the users of the SMS flood attack dataset on the two-dimensional plane, produced by the MGE method is illustrated in Fig. 34. For this projection, data from the 7th day of the simulation have been used, which is the day of the attack. The colors represent the different ground truth user groups, as described for the multi-objective method above. Mappings using the THD and CHD descriptors separately are presented in Figs. 34a and b. As is apparent, neither of the descriptor types can distinguish the four ground truth groups of data of this day. Fig. 34c illustrates the combination of both modalities, as produced by the MGE method. In this figure, the infected user groups, indicated by the fainter blue and red colors, are separated from the rest, and four clusters are formed in total, corresponding to the four ground truth user groups of this day.

The mobile network operator can select the time period of interest, in order to have an overview of the time evolution of the attack. Fig. 35 depicts the results of the MGE method, for each day of the simulation period separately. In the first six days, a common pattern of the data is observed, splitting them roughly into two clusters, corresponding to the two clusters of normal phone usage. In the 7th day, a different pattern is observed, indi-
Figure 34: Two-dimensional mapping of the SMS flood attack dataset, using the MGE method, for the 7th day of the simulation (day of the attack). The dark blue and red points represent two groups of normal users, while the faint blue and red colors represent users which exhibit abnormal behaviour. (a) Mapping using the THD descriptor. The anomalous users are overall separated from the rest of the users, but the two groups of anomalous users are not separated from each other. (b) Mapping using the CHD descriptor. The anomalous users are not separated from the other users. Few points are visible due to overlapping. (c) Mapping using both descriptors. All four data clusters are separated from each other.

6.2.4 Experimental results for the RRC attack scenario

In the second use case scenario, the RRC dataset has been used in order to apply the proposed dimensionality reduction methods to RRC attacks. Each record of the RRC dataset consists of the ID of the sender, the ID of the correspondent server, the time the message was sent and the length of the message. The THD and LHD descriptors, described in Section 6.2.2 are used as the modalities, in this scenario.

**Multi-objective dimensionality reduction**

Two-dimensional mappings produced by the multi-objective method using the CDR records of the RRC attack dataset are presented in Fig. 36. Figs. 36(a) and 36(b) depict the results when using each modality sepa-
Figure 35: Mappings produced by the MGE method for each of the 7 days of the SMS flood attack dataset. The colors represent the ground truth user classes. Blue: normal users with high usage rate, red: normal users with low usage rate, faint blue and red colors: the subsets of the two user groups which are affected by the SMS-flooding malware. In the first six days, a common pattern of two separate clusters is observed, representing the two clusters of normal usage. In the 7th day, the infected users are separated from the normal ones, forming separate clusters.

As is obvious from the first two sub-figures, using each descriptor type separately produces mappings which are not clear enough to distinguish all the three classes of users. When using the THD descriptor, apart from a concentration of normal users in the lower left part, none of the classes is visually separated from the others. Using the LHD descriptor manages to distinguish the more aggressive users from the rest. However, the less aggressive users are not clearly separated from the normal ones and the concentration of the normal users in one part of the visualization is even less clear than in the THD descriptor. Fig. 36(c) depicts the visualization produced when using a solution that is in the center of the Pareto front. In this case, all three user groups are separated from each other, each occupying a separate part of the visualization.

Similarly to the SMS flood scenario above, the operator can interactively select different solutions of the Pareto front, in order to focus on one of the descriptors or the other. The mappings for various Pareto-optimal solutions
Figure 36: Mappings of the RRC attack dataset using the multi-objective
dimensionality reduction method. Blue points represent normal users, yellow
points represent less aggressive compromised users and red points represent
more aggressive compromised users. (a) Mapping using the THD descriptor.
(b) Mapping using the LHD descriptor. (c) Mapping using both descriptors,
from the center of the Pareto front.

are depicted in Fig. 37. As is apparent from this figure, focusing on the
THD descriptor separates the normal from the compromised users, but not
the two classes of compromised users from each other. On the other hand,
focusing on the LHD descriptor, the more aggressive compromised users
are separated from the rest, but the least aggressive users are not clearly
separated from the normal users. Considering solutions from the center of
the Pareto front, i.e. corresponding to similar weights for both modalities,
manages to separate all three classes. The operator can select among these
different configurations, in order to focus on various aspects of the dataset.

Multimodal Graph Embedding dimensionality reduction

The two-dimensional mappings of the RRC attack dataset produced by
the MGE method are presented in Figure 38. Figures 38(a) and (b) depict
the mappings produced when the LHD and THD descriptors, respectively,
are used independently. The colors represent the ground truth classes, as
described for the multi-objective method, above. It can be observed that
neither of the descriptors is capable of separating all three classes of usage.
The LHD descriptor manages to separate the more aggressive users (red)
from the rest, but does not clearly distinguish the normal (blue) from the less aggressive ones (yellow). On the other hand, the THD descriptor manages to better separate the normal users from the infected ones, but does not clearly separate the two classes of infected users. Using both descriptors in combination, by performing the MGE procedure, and adaptively calculating the modality weights, produces the mapping depicted in Figure 38(c). In this figure, all three classes of usage are separated from each other. The MGE method manages to keep the advantages of both descriptors, in order to produce a clearer outcome. When visually presented to the operator, this projection provides more information and better clues for the existence of various behaviours in the data. The weights calculated for this multimodal result are 0.43 for the LHD and 0.57 for the THD modalities, which denotes a balanced trade-off between the two modalities.

6.3 Summary

In this section, applications of the proposed multi-objective and multimodal graph embedding dimensionality reduction methods, in diverse application areas, other than multimedia classification and clustering, were presented. Specifically, two applications were considered: an accessibility-enhanced image search engine and a mobile network data visualization application, for security purposes, i.e. detecting unusual behaviour. The proposed multimodal dimensionality reduction methods make no assumption about the nature of the multimodal data, apart from them being in vectorial form. Hence, they can be applied in a wide variety of applications.
Figure 38: Two-dimensional mappings of the RRC attack dataset, using the MGE method. Blue: normal users, yellow: less aggressive infected users, red: more aggressive infected users. (a) Mapping using the LHD descriptor. (b) Mapping using the THD descriptor. (c) Mapping using both descriptors, with automatic weights: LHD: 0.43, THD: 0.57.

In the accessibility-enhanced image search engine application, the goal is to re-rank the results of a standard image search engine, in order to promote to the first positions images that are easy to see for a user having a multitude of vision impairments. For this purpose, the images were analyzed so as to extract accessibility scores from them, denoting how much accessible an image is for a person having each of the supported impairments. Three types of vision impairments were considered, namely cataract, glaucoma and protanopia. The procedure for the extraction of the accessibility score is the same for all types of impairments. First, a vision impairment filter is applied to the image, in order to produce an image that illustrates how an average person having each of the impairments sees the input image. Then, the input and the filtered images are compared, and the distortion imposed by the filtered is measured. The inverse of the amount of distortion is used as the accessibility score of the image. The extracted scores were then used as the multiple features in the multi-objective dimensionality reduction method, in order to map the images on a one-dimensional space. The ordering of the images in this space constitutes the final ranking. The Discounted Cumulative Gain has been used to define the multiple objective functions for the multi-objective optimization. Further, in order to select one of the resulting Pareto-optimal solutions, the solution that is closer to the user impairment profile, i.e. the amount of disability in each of the supported impairments,
of the specific user is selected. Experimental evaluation with both artificial and real users illustrates the applicability of the multi-objective method in this application field.

In the mobile network security application, the goal is to visualize mobile network data, in order to visually distinguish users having abnormal behaviour from users with normal behaviour, so as to detect attacks of malware installed in the users’ phones. Two use cases were considered, one being an SMS flood attack scenario, and the other being an RRC attack scenario. In both cases, the input data were Call Detail Records collected by the mobile network operator. The raw data were used to extract multiple histogram-based behavioural descriptors, which are vectorial representations encoding various aspects of user behaviours. Three types of descriptors were defined, one encoding the times within the day that users send SMS messages, one encoding the distribution of correspondents towards whom the users send SMS messages and one encoding the distribution of lengths of RRC packets transmitted by the users. Using the multiple extracted descriptors as multiple modalities, the multi-objective and the multimodal graph embedding dimensionality reduction methods were performed in order to present the resulting two-dimensional visualizations to the mobile network operator. The results of the experimental application of the proposed methods to the datasets of the two use cases are visualizations where the users that have been infected by the malware are visually distinguished from the non-infected ones.
7 Conclusions

The subject of this thesis was the introduction, development and evaluation of novel techniques for multimodal dimensionality reduction. Two multimodal dimensionality reduction methods were proposed, one formulated as a multi-objective optimization problem and the other as a multimodal extension of the existing unimodal graph embedding dimensionality reduction framework. A summary of the thesis is included below, in Section 7.1, followed by a critical discussion of the outcomes, in Section 7.2, and directions for future research, in Section 7.3.

7.1 Summary of the thesis

The thesis started with an introduction to the problem of multimodal dimensionality reduction and to the approaches introduced in the thesis. A plethora of unimodal dimensionality reduction methods exist, which only handle data of a single modality. The success of multimodal fusion techniques in many research areas suggest a similar success in the field of dimensionality reduction, when multiple modalities are available for the data, which is verified by significant works in the related literature. The fact that these works do not utilize multi-objective optimization techniques, which have been successfully used in the fields of economics and engineering, along with the fact that existing generic unimodal dimensionality reduction frameworks have not been extended to multiple modalities, form the motivation for the approaches proposed hereby.

Next, a review of the state-of-the-art in the fields of unimodal dimensionality reduction, multimodal fusion and multimodal dimensionality reduction followed. A number of unimodal dimensionality reduction techniques were investigated, including global methods, local methods and methods relying on neural networks. Several of these methods can be formulated in a general graph embedding framework. Literature related to multimodal fusion was surveyed, covering representative works from a variety of application fields. The survey was specialized to fusing modalities for the purpose of dimensionality reduction or similar problems, examining several existing approaches. The literature survey assisted in clarifying those areas that have a potential for further research, thus motivating the hereby proposed methods.

The first proposed method formulates the problem of multimodal dimen-
sionality reduction as a multi-objective optimization problem. It is based on the fact that most unimodal dimensionality reduction techniques are formulated as an optimization problem, with an appropriately defined objective function to be minimized. The existence of multiple modalities naturally leads to the consideration of multiple such objective functions, which usually cannot be minimized simultaneously, thus requiring the use of multi-objective optimization techniques. Such techniques result in multiple solutions, instead of one, representing various trade-offs among the modalities, some of which could not be discovered by conventional methods, such as weighted sum-based ones. For the definition of the objective functions used in this thesis, multiple distance graphs were considered among the data, one for each modality, and their minimum spanning trees. The objective functions were based on force-directed placements of these trees, so that minimizing each objective would lead to a placement of the data that reveals the structure of the corresponding unimodal tree most clearly. The multi-objective optimization leads to placements where the structure of all trees is as apparent as possible, thus providing multimodal low-dimensional embeddings.

The second proposed method follows a different point of view. Since many unimodal dimensionality reduction techniques can be formulated as instances of a generic graph embedding dimensionality reduction framework, this method focuses on extending this framework to multiple modalities, which would immediately extend all unimodal methods that are its instances to multiple modalities as well. The graph embedding framework relies on the definition of proper affinity and penalty matrices, encoding which pairs of objects should be placed close to each other and away from each other, respectively, in the final low-dimensional mapping. The proposed multimodal graph embedding method fuses the multiple affinity and penalty matrices existing when multiple modalities are available, and provides the fused matrices to the graph embedding framework. The fusion is implemented as a weighted sum of the matrices, with the weights being learned via an optimization procedure, where an introduced notion of graph inconsistency is minimized. Graph consistency is defined as the amount at which the resulting graph forms closed triangles, since triads of objects that are connected only partially are not a consistent configuration.

The experimental evaluation of the proposed methods illustrated their
potential as solutions to the multimodal dimensionality reduction problem. Both methods were evaluated in the context of dimensionality reduction for multimedia, i.e. for mapping multimedia objects, described as multimodal objects consisting of multiple descriptors, either of the same type, such as multiple image descriptors, or of different types, such as image and sound descriptors. The target applications of the resulting low-dimensional embeddings were visualization, classification and clustering. In experiments on several multimedia datasets, both methods achieved a superior performance over existing state-of-the-art methods. The multi-objective method achieved clearer visualization results, as well as discovering informative data placements which cannot be discovered by weighted sum-based methods. The multimodal graph embedding method was applied to visualization, classification and clustering, illustrating its applicability to several machine learning tasks, and its superior performance over state-of-the-art.

The applicability of the two proposed methods was further investigated with their use in two diverse application fields, apart from multimedia exploration. The first application examined was the development of accessibility-aware image search engines, designed for the visually impaired users having a multitude of vision disabilities, in order to promote, in their result lists, those images that are most easily perceived by impaired users. The multi-objective dimensionality reduction technique was used for this application, using multiple accessibility scores extracted from the images as modalities and selecting one of the resulting optimal trade-off rankings based on the specific vision impairments of the user. Experimental evaluation with artificial and real users proved that the multi-objective rankings followed the user visual perception. The second application was data visualization for the purpose of mobile network security. Mobile network data contain several different attributes, such as caller and recipient IDs, date-times, communication types, etc. Such attributes were used to extract multiple behavioural descriptors for the users of a mobile network, which encode their behaviour with respect to different aspects, such as the times in a day that they communicate, or the distribution of their correspondents. These descriptors were used as multiple modalities that describe the users, in order to map the users as points on low-dimensional spaces, using both the multi-objective and the multimodal graph embedding dimensionality reduction methods. The data visualizations produced could effectively separate the abnormally behaving
users from the normal ones, achieving clearer results than using each modality separately.

7.2 Critical discussion

The most significant achievements of the current thesis are the following:

- The consideration of the multimodal dimensionality reduction problem as a multi-objective optimization one, which is generic and contains existing multimodal fusion techniques as its sub-cases.

- The use of the multi-objective method to discover low-dimensional mappings of data that reveal the structure of the underlying modalities more comprehensively than existing multimodal fusion techniques based on scalarization.

- The combination of the multi-objective method with force-directed objective functions to compete with existing multimodal dimensionality reduction methods.

- The introduction of a graph consistency measure, which evaluates the amount of consistency of a graph, in terms of the connectivity of triads of vertices in triangles.

- The use of the above graph consistency measure as an objective function in a maximization problem, in order to calculate the optimal weights for a weighted sum of graphs, so that the resulting graph achieves maximum consistency.

- The use of the weighted sum of multiple graph affinity matrices, with the weights automatically calculated based on graph consistency, in order to extend the unimodal graph embedding dimensionality reduction framework to multiple modalities.

- The application of the multi-objective dimensionality reduction method for the development of a novel accessibility-aware reranking procedure for image search engines, designed for visually impaired users.

- The application of the multi-objective and the multimodal graph embedding dimensionality reduction methods for the novel visualization of multiple behavioural mobile network data, for security purposes.
The multi-objective dimensionality reduction method introduces basically a new perspective to the problem of multimodal dimensionality reduction. One of its most significant characteristics, which was one of the main motivations for its formulation, is the fact that it does not require any kind of modality merging for its most part. The calculation of the Pareto-optimal solutions is based only on the notion of dominance between different solutions. The notion of dominance is independent of any assumption to the relative importance between the modalities. In this manner, the final set of Pareto-optimal solutions is the best set of solutions that one can arrive at, without merging the modalities. At this point, a further strength of the multi-objective approach emerges: the set of Pareto-optimal solutions is a superset of the solutions that would be computed using conventional multimodal fusion methods considering various relative importances between the modalities. In this respect, multi-objective dimensionality reduction can be considered as a more generic method, containing many existing multimodal fusion methods, such as weighted sum-based, co-training based, etc., as its sub-cases. Hopefully, such a perspective can open the way for taking advantage of existing or new multi-objective optimization and scalarization techniques, possibly already applied to different application fields, to address the problem of multimodal dimensionality reduction and assist in designing new dimensionality reduction algorithms. Of course, the resulting set of Pareto-optimal solutions is, in many times, only the first part of the problem. Although a set of solutions, instead of one, is often desirable, in order to allow the exploration of different modality trade-offs, it is also often desired that a single solution be provided as the result. Thus, the multitude of the solutions is both an advantage and a disadvantage of the multi-objective perspective, depending on the specific application.

With regard to the multimodal graph embedding dimensionality reduction method, its most significant advantage is the extension of the unimodal graph embedding dimensionality reduction method by merging the multiple available affinity matrices. This is advantageous in view that the existing literature mostly focuses on merging kernel matrices, or affinities in specific applications. Merging the affinity matrices of the generic graph embedding framework is equivalent to extending any unimodal dimensionality reduction method formulated as its instance to handle multiple modalities in a uniform and straightforward manner. In the graph embedding framework,
the affinity matrices are the “rulers” that guide the dimensionality reduction process to produce low-dimensional mappings having specific desired properties with respect to the relationships among the data. The concept behind the multimodal graph embedding framework is to use the available modalities in order to construct “rulers” that are better suited for dimensionality reduction. In other words, apart from the constraints imposed to the dimensionality reduction method by the available affinity matrices, the multimodal graph embedding method imposes constraints to the construction of the multimodal affinity matrix, using the introduced notion of graph consistency, in order to merge the discriminative abilities of all modalities and be able to produce more accurate low-dimensional embeddings. The multimodal graph embedding framework can be applied with many existing dimensionality reduction techniques, taking advantage of the universality of the underlying graph embedding framework, for several machine learning tasks as the targets. However, the applicability and positive performance of the multimodal graph embedding framework lies, as all graph embedding-based dimensionality reduction methods, on the assumption that the high-dimensional data lie on a low dimensional manifold that can be approached by considering neighborhood graphs constructed from the data. If this assumption does not hold, the method will still try to approach the data with a low-dimensional manifold, leading to unpredictable results. Another assumption which restricts the application of the multimodal graph embedding method is that the modalities are linearly related, so that considering a weighted sum as their combination is meaningful, an assumption that does not always hold.

7.3 Future work

The directions for future research regarding the methods proposed in the current thesis are summarized in the following list:

- The development of strategies for the selection of one of the Pareto-optimal solutions, in multi-objective dimensionality reduction.

- The examination of various objective functions evaluating the low-dimensional mapping of the data, to be used as the multiple objectives in multi-objective dimensionality reduction.

- The investigation of non-linear methods for merging the multiple affinity matrices in multimodal graph embedding dimensionality reduction.
- The examination of different measures, apart from graph consistency, to be used for the determination of the optimal modality merging, in multimodal graph embedding dimensionality reduction.

In more detail, as mentioned in the discussion section above, in many applications of the multi-objective dimensionality reduction method, a single solution may be desired, which needs to be selected from the Pareto-optimal ones. This is the part, after multi-objective optimization, that the relative importance among the modalities needs to be considered, either in the form of user-defined preferences, or in the form of automatic selection. Already, in the accessibility-aware image ranking application, described in Section 6.1, the final solution was selected as the one closest to the point representing the user impairment profile. In this respect, the user profile constitutes a set of “preferences” imposed by the user. Designing strategies for the automatic selection of one of the Pareto-optimal solutions is an interesting and important topic for further investigation. Suggestions for possible strategies include selecting the one closest to the ideal point of simultaneously minimal objectives or using concepts of risk management, such as the Sharpe ratio [213]. They could also be inspired from recent propositions for Pareto front selection, such as the Nash-based criteria used in [214].

The objective functions used as the multiple objectives for the multi-objective dimensionality reduction method were based mostly on visualization principles, i.e. revealing the structure of the unimodal minimum spanning trees. The generic nature of the core concept behind the multi-objective method, allows the use of any kind of objective functions to be used. Thus, objectives specifically designed for specific applications could be used, e.g. ones that use training data and attempt to maximize the discrimination between data belonging in different classes. Such objective functions could fine-tune the multi-objective method to specific machine learning applications, such as classification, in order to provide more accurate results.

Regarding the Multimodal Graph Embedding dimensionality reduction method, one important direction for future work is the use of non-linear methods for the combination of the multiple affinity matrices available. Although the weighted sum-based combination used hereby has been successful in comparison to existing methods, and although such linear combinations have successfully been used by several existing multimodal fusion techniques, the linear combination assumes a linear relationship among the modalities,
which may not exist, as also mentioned in the discussion section above. The modalities may be better described as having a non-linear relationship among them, or being correlated with each other. In such cases, non-linear or correlation-based combination methods need to be considered, instead of linear ones. Recent methods for exploring the often complex relations among the unimodal features, such as using deep learning as in [215], can also provide inspiration. Designing more sophisticated combination methods that capture the modality relations more accurately is a very interesting and promising issue for further research.

Finally, regarding the fourth of the above points, the Multimodal Graph Embedding method, as presented in this thesis, utilizes the introduced graph consistency measure in order to calculate the optimal combination of the multiple affinity matrices. The graph consistency measure is based on the intuition that triads of vertices which contain two edges should contain the third edge as well, in order for the configuration to be consistent. This measure was developed with the aim to construct a final affinity matrix that most closely resembles an ideal matrix that connects classes of points in more compact and interconnected regions of the graph. The proposed measure considers triads of points. However, higher-order cliques of vertices could be considered in order to measure the amount of consistency, which may lead to more discriminative final graphs. This is an interesting direction for further investigation. Other measures that could describe the consistency of a graph could also be used, for instance ones inspired by optimal graph construction methods such as the consensus k-NN method of [31] or the random forest-based method of [25].
References


[91] Apostolos Axenopoulos, Stavroura Manolopoulou, and Petros Daras. Multimodal search and retrieval using manifold learning and query


[21] Hanghang Tong, Jingrui He, Mingjing Li, Changshui Zhang, and Wei-Ying Ma. Graph based multi-modality learning. In *Proceedings of


[160] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach


