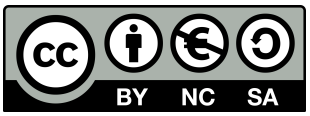
Covariance-based Descriptors for Pattern Recognition in Multiple Feature Spaces

Pol Cirujeda Santolaria

TESI DOCTORAL UPF / ANY 2015

DIRECTOR DE LA TESI
Dr. Xavier Binefa Valls
Department of Information and Communication Technologies



"Begin at the beginning," the King said, very gravely,
"and go on till you come to the end: then stop."

— LEWIS CARROLL, *Alice in Wonderland*

Gràcies Mama

Acknowledgements/Agraïments

Acknowledgements. That less-science, more-human-understandable part of a thesis. I'd like to think that everyone who I must thank already knows so –what a polite formula to cover my back in case I forget someone! But certainly, I have to mention some of you for your special support during these years.

First of all I would like to thank Xavier Binefa for the opportunity of following this PhD at UPF, and the freedom offered during these years. It has been really nice to share statistics course syllabus updates and state-of-the-art osteopathy discussions. I hope this thesis can return some payback in exchange for your support, guidance and patience.

I would like to thank all my colleagues of the CMTech lab: Brais, Luis, Marc, Oriol, Ciro, Fede, Adrià, Dima, all the intern students... and specially Xavi Mateo for the nice work together and future cheese-making plans. From each one of you I have learned something. Thanks.

A very special consideration goes also to the MedGift group members at HES-SO in Sierre, Switzerland. Henning, Adrien: thanks for a very generous opportunity and warming alpine reception. Yashin, Thomas, Oscar, Ale, Stefano, Visara, Alba, Roger, Ranveer, Imanol, Sebas, Michael Barry, Morgane and the rest of the crew: thanks for that wonderful time between mountains, raclettes, apéros, snowboarding, and excellent research. Santé!

It would have been a tough time if I would not have met a nice bunch of true friends at UPF. No exaggerations. Vane, Jana, Magda, Jonathan, Laura Sanchez, Ray, Trang, Simon... thanks for making of me a good researcher, besides the pure academic sense. Of course, I would like to thank all the wonderful staff that keeps the engine going: Montse, Lydia, Santa, Judith, Joana, Bea.

Als meus amics: Ricard, Alex (Arr!), Miki, Alba, Cris, Jordi Cremades, Jordi Sala. Per cuidar-me en general i per la muntanya, els cines, els sopars, els viatges... en particular. Moltes i moltes gràcies, no es poden tenir millors amics.

A la meva petita família. La meva mare, la meva àvia: Isa, Lolín... no puc expressar amb paraules el que m'heu donat. Espero algun dia saber-ho retornar. Us estimo.

I per descomptat a la Laura, que m'ha hagut de compartir amb aquest doctorat... pels ànims, el recolzament i els petons: infinites gràcies. Buongiorno principessa.

Finalment... un doctorat s'aguanta millor amb un gos. Byte, gràcies *chucho*!

Abstract

Data representation is of primary interest in any research field. In computer vision it is usual to work with features, to play with them, to construct useful representations so a computer can be later told how to recognize specific patterns of a desired model. So what if we could observe and understand these features not by their static content, but from a raised space where they had an extra value? For a computer, a person in an image can be a set of color pixels. For a good pattern recognition algorithm, should not it be better if persons were a description of "things happening together within a region"? For instance, would not a description of a given amount of color combined with a given amount of contour in a particular position be more flexible and discriminative at the same time for detecting persons in images?

This dissertation explores the use of covariance-based descriptors in order to translate feature observations within regions of interest to a descriptor space using the feature covariance matrices as discriminative signatures. This space constitutes the particular manifold of symmetric positive definite matrices, with its own metric and analytical considerations, in which we can develop several machine learning algorithms for pattern recognition. Regardless of the feature domain, whether they are 2D image visual cues, 3D unstructured point cloud shape features, gesture and motion measurements from depth image sequences, or 3D tissue information in medical images, the covariance descriptor space acts as a unifying step in the task of keeping a common framework for several applications.

In order to proof the validity and scope of this methodology, this thesis presents the foundations of the covariance-based descriptor framework –construction, differential geometry theory, and manifold-aware machine learning techniques– and places them within four concrete application cases: 2D image classification via Riemannian manifold boosting; 3D unstructured point clouds scene description and registration via a game theoretic cost minimization approach; human gesture classification from depth map sequences via a sparse representation manifold based classifier; and medical imaging analysis for tissue classification from computerized tomographies. All the presented approaches share a same baseline but with specifically tailored features and accompanying machine learning techniques.

Resum

La representació de les dades és d'especial interès en qualsevol àrea de recerca. A la visió per computador estem acostumats a treballar i jugar amb les característiques de les dades, a construir representacions útils que més endavant serveixin per ensenyar a un ordinador a reconèixer patrons específics en un determinat model. Però, i si poguéssim observar i entendre aquestes característiques no pel seu contingut sinó des d'un espai més elevat en el que tinguessin un altre valor? Per un ordinador, una persona en una imatge pot ser un conjunt de píxels de colors. Per a un bon algorisme de reconeixement de patrons, no seria millor si les persones fossin una descripció de "coses passant alhora dins d'una regió"? Per exemple, no seria més flexible, i alhora discriminativa, una descripció de la quantitat de color combinada amb una quantitat de contorn determinat i en una posició particular per tal d'indicar la presència d'una persona en una imatge?

En aquesta tesi s'explora l'ús de descriptors basats en la covariància per tal de traslladar la observació de característiques dins de regions d'interès a un determinat espai descriptiu que utilitzi les matrius de covariància de les característiques com a signatures discriminatives de les dades. Aquest espai constitueix la varietat de les matrius simètriques definides positives, amb la seva pròpia mètrica i consideracions analítiques, en la que podem desenvolupar diferents mètodes de *machine learning* per al reconeixement de patrons. Sigui quin sigui el domini de les característiques, ja siguin observacions visuals en imatges 2D, característiques de forma en núvols de punts 3D, gestos i moviment en seqüències d'imatges de profunditat, o informació de densitat en imatges mèdiques en 3D, l'espai del descriptor de covariància actua com un pas d'unificació en el repte de mantenir un marc de treball comú per a diverses aplicacions.

Amb l'objectiu de provar la validesa d'aquesta metodologia, aquesta tesi presenta els fonaments del descriptor basat en covariància –la seva construcció, la teoria necessària en geometria diferencial, i diverses tècniques de *machine learning* que tinguin en compte la seva varietat– i els situa en quatre casos concrets d'aplicació: classificació d'imatges 2D via la tècnica de *boosting*; descripció i registració d'escenes en núvols de punts 3D amb un procés de minimització basat en teoria de jocs; classificació de gestos en seqüències d'imatge de profunditat amb un classificador *sparse*; i anàlisi d'imatge mèdica per a classificació de teixit en imatges de tomografia computeritzada. Totes aquestes aplicacions comparteixen el mateix marc de treball però amb diferents observacions adequades a les dades i aplicant diferents algorismes de *machine learning* en funció de la natura del problema a resoldre.

Contents

List of Figures

List of Tables

Glossary

ASL American Sign Language
CSHOT Unique Colour Signatures of Histograms
CT Computerized Tomography
ESS Evolutionary Stable Strategy
FPFH Fast Point Feature Histograms
GGO Ground-Glass Opacity
$\mathbf{GL(2)}$ General linear group of 2D affine transformations
HOG Histogram of Oriented Gradients
ICP Iterative Closest Point
LBP Local Binary Patterns
RANSAC Random Sample Consensus
RGB Red Green Blue (color space)
SIFT Scale-Invariant Feature Transform
SKIG Sheffield Kinect Gesture dataset
$\mathbf{SO(3)}$ Special orthogonal group of 3D rotations
STIP Spatio-Temporal Interest Points
SURF Speeded Up Robust Features
SVD Singular Value Decomposition
SVM Support Vector Machines
$\mathbf{Sym}_{\mathbf{d}}^{+}$ Symmetric positive definite matrices manifold
$\mathbf{T}_{\mathbf{Y}}$ Tangent space at manifold point Y

List of Publications

This dissertation has led to the following publications:

Journal articles:

- **POL CIRUJEDA**, YASHIN DICENTE CID, XAVIER MATEO, XAVIER BINEFA. *A 3D Scene Registration Method via Covariance Descriptors and an Evolutionary Stable Strategy Game Theory solver*. In International Journal of Computer Vision, 2015.

- In preparation: **POL CIRUJEDA**, YASHIN DICENTE CID, HENNING MÜLLER, DANIEL RUBIN, TODD A. AGUILERA, BILLY W. LOO JR., MAXIMILIAN DIEHN, XAVIER BINEFA, ADRIEN DEPEURSINGE. *A Riesz-Covariance Texture Model for Prediction of Adenocarcinoma Recurrence in Lung CT*. To be submitted to Transactions on Medical Imaging.

Conference contributions:

- **POL CIRUJEDA**, HENNING MÜLLER, DANIEL RUBIN, TODD A. AGUILERA, BILLY W. LOO JR., MAXIMILIAN DIEHN, XAVIER BINEFA, ADRIEN DEPEURSINGE. *3D Riesz-Wavelet Based Covariance Descriptors for Texture Classification of Lung Nodule Tissue in CT*. In International Conference of the IEEE Engineering in Medicine and Biology Society, 2015.

- **POL CIRUJEDA**, XAVIER BINEFA. *Medical Image Classification via 2D Color Feature Based Covariance Descriptors*. In Working Notes of CLEF 2015 (Cross Language Evaluation Forum), CEUR Workshop Proceedings, 2015.

- OSCAR ALFONSO JIMÉNEZ DEL TORO, **POL CIRUJEDA**, YASHIN DICENTE CID, HENNING MÜLLER. *RadLex Terms and Local Texture Features for Multimodal Medical Case Retrieval*. In Multimodal Retrieval in the Medical Domain, 2015. Lecture Notes in Computer Science, Springer.

- **POL CIRUJEDA**, XAVIER BINEFA. *4DCov: a Nested Covariance Descriptor of Spatio-temporal Features for Gesture Recognition in Depth Sequences*. In International Conference on 3D Vision, 2014.

- **POL CIRUJEDA**, XAVIER MATEO, YASHIN DICENTE CID, XAVIER BINEFA. *MCOV: a Covariance Descriptor for Fusion of Texture and Shape Features in 3D Point Clouds*. In International Conference on 3D Vision, 2014.

- **POL CIRUJEDA**, XAVIER BINEFA. *Augmented Poselets for Human Body Pose Inference by a Probabilistic Graphical Model*. In SIGMM ACM Multimedia, 2012.

- ORIOL MARTINEZ, **POL CIRUJEDA**, LUIS FERRAZ, XAVIER BINEFA. *Dissociating Rigid and Articulated Motion for Hand Tracking*. In International Conference on Pattern Recognition, 2012.

Chapter 1

Introduction

> "Which of my photographs is my favorite? The one I'm going to take tomorrow."
>
> — IMOGEN CUNNINGHAM

AFTER AN ACCURATE REPRESENTATION, finding patterns in data is the fundamental first stone in many research areas. As a computer vision researcher, when one applies this to the solution of problems from visual inputs, research becomes even more thrilling as one deals with the utopia of "making computers see, understand and interact with our world". Tasks as recognising objects in images for classification and retrieval, understanding 3D scenes for context navigation in robotics, providing real-time feedback to motion gestures for human-computer interaction or analysing medical images for computer aided diagnose or tele-medicine are examples of research outcomes pursued in this dissertation.

Improving existing methodologies on a field as computer vision is challenging due to an active research community, although the growing number of open tasks also implies that inspiration can be taken from many great contributions presented in top conferences and journal articles month after month. The confluence between feature extraction, machine learning, pattern recognition, new computational approaches and the application of all together to the solution of many endless problems encourage infinite creativity. Nevertheless, one has to choose some focus, specially for a dissertation, so the motivation of this thesis resides in the choice of a particular feature extraction and description method, its implication to machine learning, and a selection of problems to solve in order to present some contributions which can hopefully raise the benefits of novel research in a crowded area.

Therefore, the main goal is to provide the formulation and demonstrate the properties of a framework based on the statistical notion of covariance of features for data description and classification. As a very basic introduction, the aim is to

model entities from different natures and their observable features: for instance, color in images, curvature or density in 3D surfaces or volumes, or motion patterns in image sequences. Different objects, such as persons in images, planar or round surfaces in 3D point clouds, or different sign language gestures, will have different sets of observable features which can be considered as samples of statistical distributions. The presented framework provides a compact representation of these feature distributions by means of their covariance matrix, which is the baseline for extracting models on different feature spaces but sharing a common methodology core. The straightforward implication of this representation space is that it has a particular non-Euclidean geometric distribution, as descriptors lay on the Riemannian manifold of symmetric positive definite matrices. This supposes using particular metrics and analytical properties which can be exploited for defining novel machine learning algorithms for classification and pattern recognition on the provided descriptor space.

1.1 Motivation and Objectives

What translates features to a common space is the statistical notion of covariance, which, used as a descriptive unit can provide a signature of how feature observations change together for given entities inside a region of interest. For instance, for defining color images, certainly one of the most basic feature varieties would be the complete RGB color space. A given image can contain all sort of sparse values within this color space. But in the attempt of modelling a particular image class space, for instance skin tissue captured in light microscopy images, these class samples could be better defined by particular sets with concrete distributions of red, green and blue color samples. These distributions will have characteristic statistics such as a particular mean and variance for each color feature, and joint pairwise feature covariances. Therefore, the class model might be characterized in a better space: the covariance of features. If we could imagine an abstract space of all these covariances, for instance for image classes of different nature (skin tissue, muscles, bones...) these covariance representations would appear clustered according to these differences as the distribution of colors should be particularly discriminative for each class. This can be considered as one of the strengths of this dissertation. Covariance notion, understood as a descriptor signature, can translate any feature space whether it is color or a very complex d-dimensional signal, to a common compact manipulable space. The goal is to move features to a place where it should be possible to extract a "dictionary" of the world of classes that should be modelled. As this can result a little abstract, this dissertation explores feature selection in several problems, all with the aim of demonstrating the flexibility of this framework: from 2D color images to 3D point clouds and dense

medical images, and 4D depth image sequences.

Features may change, but the descriptor space is homogenizing: this new variety can be analysed with different techniques according to different problems which are used as motivation and validation contexts. From part-based classification methods to dictionary learning and algorithms for descriptor matching and outlier rejection, it will be shown that this descriptor space is not only valid for data representation: due to a characteristic Riemannian geometry, it is also a convenient space for developing machine learning applications. So along a progressive motivation this dissertation presents, in a conceptually chronological order, the research carried on along different feature space applications covering the following questions:

- In object classification and retrieval from color images, can flexible and variable objects be defined via covariance-based descriptors? Is their conception suitable for modelling classes with high intra-class variability? In a particular application such as human body detection by classification, can a natural part-based method be provided so it fragments the classification solution in several part-oriented descriptor models? Furthermore, the feasibility of the proposed descriptor framework will be tested in a variable dataset of medical images from different acquisition device origins.

- 3D scene understanding is a straightforward jump regarding 2D images: pixels, instead of being organized on a plane, are just unstructured collections of points with 3D coordinates and color information. Can the already introduced framework be leveraged for recognizing and matching three-dimensional scene views fusing shape and texture? Several questions may arise: the descriptor might include view-invariant information for local surface description, and a holistic matching procedure must include global scene information for taking into account challenging conditions in scenes, such as pattern repetitions in different areas or symmetries. The goal is to provide a scene registration methodology for discarding outlier matches, registering 3D views, and perform object retrieval. Besides the obvious outcome of a descriptor for 3D point clouds, this has also the desired goal of demonstrating how extra layer algorithms can be integrated with the descriptor space exploiting its analytical properties.

- After working with 3D information, can the same framework be extended to 4D, understanding that as a collection of depth image frames which, along time, constitute human gesture sequences? This is not as straightforward jump as from 2D to 3D, but with some design strategies covariance-based descriptors for spatio-temporal gesture definition can be formulated. Different cadences and motion patterns have to be taken into account. In any

case, this is a challenging research line receiving recent attention thanks to the appearance of easy access depth acquisition devices and with major application impact as it can be used for real-time hand sign language translation or natural human-computer interaction.

- Finally, as a complementary research outcome, the presented method has had the opportunity of being tested in dense 3D medical images as computerized tomographies (CT). This application context is similar to 3D scene understanding, but data is now structured and represents dense volumes instead of surfaces. Using tissue texture-based features, the covariance-based descriptor framework can provide a straightforward 3D region descriptor for region classification in CTs. This has a major impact in medical image processing as can be tested in areas as nodule classification, unsupervised tissue segmentation or nodule recurrence models, with a close collaboration with clinicians expertise and their feedback.

1.2 Thesis Organization and Contributions

This thesis does not only provide a mandatory theoretic framework, but also a set of practical, algorithmic solutions to the presented problems. The use of four particular contexts will serve for dividing the present dissertation in respective chapters, each of them introducing the respective problem presentation, state of the art analysis for each one of the applications, the adapted framework formulation and feature space, the contributed machine learning algorithms and a subsequent experimental evaluation and discussion.

This dissertation is organized as follows:

- Chapter 2 presents covariance-based descriptors for 2D color images. The use of the manifold of covariance matrices in computer vision is introduced, with a review of their application as descriptors, their Riemannian geometry implications and a practical reference of the common operators and notions used in further chapters, and an overview of related methods with a background in machine learning. In the first part of the chapter, an existing part-based classification of human bodies in 2D images is reviewed as a first introduction to the complete framework. This includes a technique with manifold atlas-based learning via a boosting adaptation. The second part of the chapter presents a whole-image color-based covariance descriptor for medical image retrieval and a manifold-regularized sparse representation classifier. The outcomes of this research have been presented in the paper:

POL CIRUJEDA, XAVIER BINEFA. *Medical Image Classification via 2D color feature based Covariance Descriptors*. In Working Notes of CLEF 2015 (Cross Language Evaluation Forum), CEUR Workshop Proceedings, 2015.

- Chapter 3 presents a 3D covariance-based descriptor formulation for fusion of shape and texture features in unstructured point clouds. This descriptor is conceived for 3D scene understanding, with applications to pairwise view registration and object retrieval, which require a scene-aware methodology for point pairing and outlier rejection. The chapter also introduces a game theory based approach for adding geometric constraints as an additional layer to descriptor extraction and matching. This research line has yielded to the following publications:

  POL CIRUJEDA, YASHIN DICENTE CID, XAVIER MATEO, XAVIER BINEFA. *A 3D Scene Registration Method via Covariance Descriptors and an Evolutionary Stable Strategy Game Theory Solver*. In International Journal of Computer Vision, 2015.

  POL CIRUJEDA, XAVIER MATEO, YASHIN DICENTE CID, XAVIER BINEFA. *MCOV: a Covariance Descriptor for Fusion of Texture and Shape Features in 3D Point Clouds*. In International Conference on 3D Vision, 2014.

- Chapter 4 presents a more abstract covariance-descriptor extraction in sequences of depth image maps. The goal is to encode spatio-temporal information in action volumes, and classify the resulting descriptor space by means of the sparse representation method already introduced in chapter 2. The publication associated to this contribution is:

  POL CIRUJEDA, XAVIER BINEFA. *4DCov: a Nested Covariance Descriptor of Spatio-Temporal Features for Gesture Recognition in Depth Sequences*. In International Conference on 3D Vision, 2014.

- Chapter 5 provides a final application context of covariance-based descriptors in 3D computerized tomography medical images. Dense texture features are used in order to leverage the framework presented along this dissertation as a three-dimensional descriptor for lung tissue classification and as a similarity measure for organ region matching in multi-modal medical case based retrieval. A bag-of-covariances method is used in order to classify different nodule tissue regions in patient lungs, in collaboration with expert clinicians feedback. The contributions in this area have been gathered in these publications:

  POL CIRUJEDA, HENNING MÜLLER, DANIEL RUBIN, TODD A. AGUILERA, BILLY W. LOO JR., MAXIMILIAN DIEHN, XAVIER BINEFA, ADRIEN

DEPEURSINGE. *3D Riesz-wavelet Based Covariance Descriptors for Texture Classification of Lung Nodule Tissue in CT*. In International Conference of the IEEE Engineering in Medicine and Biology Society, 2015.

OSCAR ALFONSO JIMÉNEZ DEL TORO, POL CIRUJEDA, YASHIN DICENTE CID, HENNING MÜLLER. *RadLex Terms and Local Texture Features for Multimodal Medical Case Retrieval.* In Multimodal Retrieval in the Medical Domain, 2015. Lecture Notes in Computer Science, Springer.

This research is being currently used for temporal modelling of patient cancer recurrence and for unsupervised tissue segmentation in computer-aided diagnose, and the following publication is in ongoing preparation:

POL CIRUJEDA, YASHIN DICENTE CID, HENNING MÜLLER, DANIEL RUBIN, TODD A. AGUILERA, BILLY W. LOO JR., MAXIMILIAN DIEHN, XAVIER BINEFA, ADRIEN DEPEURSINGE. *A Riesz-Covariance Texture Model for Prediction of Adenocarcinoma Recurrence in Lung CT*. To be submitted to Transactions on Medical Imaging.

- Finally, chapter 6 presents the conclusions of the dissertation, providing a summary of its outcomes as well as a discussion on possible future continuations of the presented research lines.

Chapter 2

Manifolds in Computer Vision. 2D Image Retrieval and Classification

"I don't trust words. I trust pictures."

— GILLES PERESS

OBJECT DETECTION AND CLASSIFICATION in color images has been a recurrent topic in computer vision research along decades. There are particular object classes, such as human subjects, that pose an added challenge to classification methods due to their inherent variability and non-rigidity conditions. Other pattern recognition tasks, such as medical image-based case retrieval are affected also by considerable intra-class variations due to an often non-complete learning dictionary of class samples. These two applications, human body detection and medical image retrieval, will serve as context problems for the introduction of our covariance-based descriptor framework. We will present similar feature selection functions for each application, and the solution to both tasks by different machine learning methods tailored for each one of the application natures: boosting and a sparse representation based classifier.

2.1 Introduction

Pattern recognition, object detection and image classification in general have been long studied in the computer vision literature for a long time. For fairly rigid objects there exist state-of-the-art methods with consolidated techniques which include keypoint extraction, feature selection and sparse point descriptors. These can be classified, clustered or matched by standard machine learning techniques in order to establish the presence of the desired template. Complexity increases

as this query entity can suffer variabilities due to its non-rigidity or heterogeneity: then, the modelling approach should take into account this flexible nature.

This chapter introduces the benefits of covariance-based descriptors on 2D images, for two particular applications: human body image classification and medical image retrieval. Covariance-based descriptors were first introduced in [Tuzel et al., 2006] as flexible discriminative signatures for object recognition and texture classification. The main intuition about the statistical definition of feature variations under a region of interest is that covariance measures, together with their loss of structural information, can be salient enough for the classification of heterogeneous class samples. The context of application found in the two commented problems provides a good introduction to our framework, as it enables to test the fundamental benefits of covariance descriptors (robustness to noise, and invariance to rigid transformations due to the loss of structural information) in the classification of such variable entities that are human subjects and medical images from different sources. The aim of this chapter, besides the own introduction of the framework, is to show its versatility and integration with complementary machine learning techniques chosen accordingly to the different application problems.

This chapter is organized as follows: section 2.2 reviews some of the typical approaches for 2D region description, and starts the introduction of manifold-based approaches in computer vision. Section 2.3 introduces the basic framework of covariance-based descriptors for 2D images which will be common in further chapters of this dissertation (changing the features accordingly to different domains of application). Sections 2.4 and 2.5, respectively, present the integration of this framework with an adapted boosting approach for human body detection in 2D images and a sparse representation classification method for medical image retrieval. Finally section 2.6 presents the conclusions observed from both applications and its relation to further chapters.

2.2 Related Work

Typically, pattern recognition for object classification in 2D images has been tackled by a standard paradigm of feature extraction and data description which is further fed to classification methods: some standard techniques include boosting [Schapire, 2003; Viola and Jones, 2001], support vector machines (SVM) [Cortes and Vapnik, 1995], or neural networks [Haykin, 1998; Krizhevsky et al., 2012]. But classification methods would not be useful if data was not correctly represented on a first instance, providing a discriminative and easily separable representation space.

An overview to some of the state-of-the-art region representation methods in the image processing area reflects how low-level features are extracted in image

areas and encoded in compact representations. It is the case, for instance, of image descriptors such as Histograms of Oriented Gradients (HOG) [Dalal and Triggs, 2005], which evaluate local histograms of gradient orientations quantized on a dense grid over the image, and characterize local appearance and shape of objects by its edge information. Local Binary Patterns (LBP) [Ojala et al., 2002] or Geometric Blur [Berg and Malik, 2001; Berg et al., 2005] are other examples of defining an image region by observing neighbourhood points of a given pixel, by a codification of its compared intensity values or by a spatially varying kernel filtering procedure, respectively. Other well-known descriptor methods include SIFT (Scale-Invariant Feature Transform) [Lowe, 2004], or SURF (Speeded Up Robust Features) [Bay et al., 2008] which, besides oriented gradient information based descriptors, also provide interest point extraction methods and integrated template matching and registration algorithms.

Covariance-based descriptors for color image area classification were first presented in [Tuzel et al., 2006]. This first approach introduced the concept of covariance matrices as a feature representation procedure, and demonstrated its feasibility for object detection and texture classification in color images. This was rapidly extended to more image classification applications [Porikli et al., 2006; Porikli and Kocak, 2006]; placed into the context of more complex systems for object detection [Tuzel et al., 2007, 2008b]; or used by other authors [Yao and Odobez, 2008; Kluckner et al., 2009]. This supposed the peak rise of a novel methodology that could be easily extended in the future. Descriptors had been understood as procedures for data codification so far, but the representation of image features by their covariance matrix gave rise to a more meaningful characterization: from then on, descriptors were not only codifying data, they started to be entities with a geometric significance on an abstract data space. Features were being translated from the space of their own static values (color spaces, gradient magnitudes and orientations, pixel-level curvatures...) to a higher space where particular joint aggregations of these values make sense to a given model. Furthermore, being symmetric positive definite matrices, these descriptors would lie on a particular non Euclidean space: the manifold of symmetric positive definite matrices Sym_d^+.

The particular case of covariance matrices is of course one of the important cores of this dissertation, but the computer vision community has paid attention to matrix manifolds in general in the recent years. This is also an inspiration for this thesis, as it will be exploring the application of machine learning techniques with particular Riemannian geometry in mind. Dealing with manifold distributed data supposes a trade-off between using particular non Euclidean metrics at the expense of having a well defined space that introduces less error in data modelling, comparison or estimation [Lee, 2003; Absil et al., 2009]. These matrix manifolds can model 2D affine transformations, 3D rotations, or d-dimensional data feature

covariances. The surveys conducted in [Lui, 2012; Li et al., 2014] establish a comprehensive overview on a complete set of particular manifolds and several applications to face and action recognition, tracking and data clustering. Manifolds such as $GL(2)$ (the general linear group of all possible 2D affine transformations) or $SO(3)$ (the special orthogonal group of 3D rotations) can be used for defining the space of allowed transformations of an object along different stages of a sequence. For instance, [Tuzel et al., 2008a] or [Kwon and Park, 2009] use the particular group of affine transformations for tracking applications based on manifold regression. [Belkin et al., 2006; Elgammal and Lee, 2004] have explored several approaches for manifold regularization, keeping the positive definiteness of the descriptors. [Sivalingam et al., 2009; Pitelis et al., 2013; Huang et al., 2014] have presented metric learning and atlas based modelling techniques for the supervised and semi-supervised learning of manifold distributed data keeping into account their Riemannian geometry. In the same sense [Tenenbaum et al., 2000; Roweis and Saul, 2000; Xiong et al., 2007; Hamm and Lee, 2008; Jayasumana et al., 2013; Harandi et al., 2014a] have presented Riemannian geometry variants of well-known algorithms for data alignment and dimensionality reduction in Sym_d^+, as well as sparse coding representations [Harandi et al., 2012; Cherian and Sra, 2014].

2.3 Covariance Descriptors for 2D Images

Raw pixel values, like color or gradients, had been historically used as the most basic image features for detection and classification tasks. Despite of that, the pursuit of robustness against shape variations was a constant desired goal that can not be achieved by features themselves. Furthermore, in the area of computer vision, it is usual to manage significant amounts of imagery data. Thus, it should be desirable to find a "perfect" representation which should provide the most efficient trade-off between high representation capability, robustness against unexpected noisy inputs -occlusions, pose changes, illumination conditions, etc- and low computational cost, both in terms of memory and calculation requirements.

Covariance-based descriptors were firstly introduced in [Tuzel et al., 2006] for object and texture classification. Their conception idea is to define an image region not by a set of associated features themselves, but by a representation of these feature joint variations. Feature vectors within a region can be considered as samples of an d-dimensional distribution, which will characterize that particular region. The first step for the computation of the descriptor is, therefore, the choice of some base features at a pixel level. In order to formally define this 2D covariance-based descriptor, a feature selection function $\Phi_{2D}(I)$ for a given image

I is defined as:

$$\Phi_{2D}(I) = \left\{ \phi^{2D}_{x,y}, \forall x, y \in I \right\}, \tag{2.1}$$

which provides a set of feature vectors $\phi^{2D}_{x,y}$ for each one of the pixel coordinates $\{x, y\}$ inside all the image I. For a first task of 2D image description the following 8-dimensional feature vector can be designed, as shown in figure 2.1:

$$\begin{aligned} \phi^{2D}_{x,y} = \Big[& x, y, |I^x|_{x,y}, |I^y|_{x,y}, \sqrt{(I^x)^2_{x,y} + (I^y)^2_{x,y}}, \ ... \\ & ... \ |I^{xx}|_{x,y}, |I^{yy}|_{x,y}, \arctan \frac{|I^x|_{x,y}}{|I^y|_{x,y}} \Big] \end{aligned} \tag{2.2}$$

These feature vectors include low-level visual cues such as image gradient, its magnitude and curvature, and spatial coordinates. This introduces the descriptor formulation for a generalist 2D image descriptor without color information at the moment. This will be used in the following section for human body detection in images, since contour information is considered more important than pixel color values in order to classify persons by their shape rather than by their appearance. Designing feature selection functions is class and problem-dependant, as analysed in [Cargill et al., 2009] where different sets of features involving color spaces, gradient magnitudes or second order derivatives were reviewed for image classification applications. Despite of that, one of the main advantages of covariance-based descriptor is that their construction is independent from this feature selection stage. For any given region $R \subset I$, containing a set $\{\phi^{2D}_i\}_{i=1..S}$ of S d-dimensional feature vectors, the following $d \times d$ covariance matrix of features will provide a characteristic signature for R:

$$Cov_R = \frac{1}{S-1} \sum_{i=1}^{S} \left(\phi^{2D}_i - \mu\right) \left(\phi^{2D}_i - \mu\right)^T, \tag{2.3}$$

where μ is the mean of the points $\{\phi^{2D}_i\}_{i=1..S} \in R$.

Estimating covariance matrices from data features as a form of descriptor provides many benefits both in its meaning and in computational implications. In a first place, the discriminative value comes from providing a signature for the distribution of features within the region of interest, as their joint covariance values. This notation loses all the information about feature structure, therefore this provides invariance to rotation and spatial transformations. In terms of computation and dimensionality, this descriptor does not involve any major operation and translates the $S \times d$ dimensional observations to a $d \times d$ matrix.

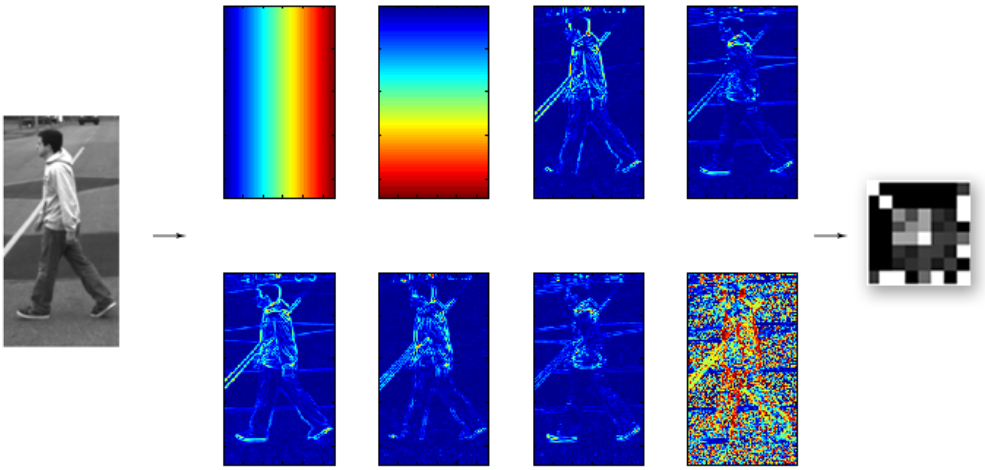


Figure 2.1: Original region of interest, 8-dimensional feature layers (from equation 2.2) for all region coordinates, and obtained covariance descriptor.

The final consideration about covariance matrices, as symmetric positive definite matrices, is that they do not lay on a Euclidean space, but on a Riemannian manifold. Far from supposing a drawback, this is one of the pillars of this dissertation.

2.3.1 Riemannian Geometry

Covariance descriptors have the form of $d \times d$ covariance matrices which, besides providing a compact and flexible representation, causes them to lie in the Riemannian manifold of symmetric positive definite matrices Sym_d^+. This space is an open convex cone so that $X + tY \in Sym_d^+$ for $t > 0$ given $X, Y \in Sym_d^+$. This has a major impact on considering covariance matrices of data features as descriptive units, as their spatial variety is geometrically meaningful: samples of classes sharing similar feature characteristics will remain under close areas in this descriptor space. According to [Arsigny et al., 2006], the Riemannian manifold can be approximated in close neighborhoods by the Euclidean metric in its tangent space, T_Y, where the symmetric matrix Y is a reference projection point in the manifold. T_Y is formed by a vector space of $d \times d$ symmetric matrices, and the mapping of a manifold element X to its tangent space $x \in T_Y$ is made by the point-dependent tangent mapping operation:

$$x = \log_Y(X) = Y^{\frac{1}{2}} \log\left(Y^{-\frac{1}{2}} X Y^{-\frac{1}{2}}\right) Y^{\frac{1}{2}}. \tag{2.4}$$

In an analogous manner, the exponential mapping of a point $x \in T_Y$ returns its

original point representation X in the Sym_d^+ manifold:

$$X = \exp_Y(x) = Y^{\frac{1}{2}} \exp\left(Y^{-\frac{1}{2}} X Y^{-\frac{1}{2}}\right) Y^{\frac{1}{2}}. \tag{2.5}$$

Derivatives of a point of the manifold $X \in Sym_d^+$ lie on the vector space T_X. This notion can be used for decomposing the connected Riemannian manifold in different planes of an atlas, with particular applications in machine learning [Sivalingam et al., 2009; Huang et al., 2014]. See figure 2.2 for a schema of these operations.

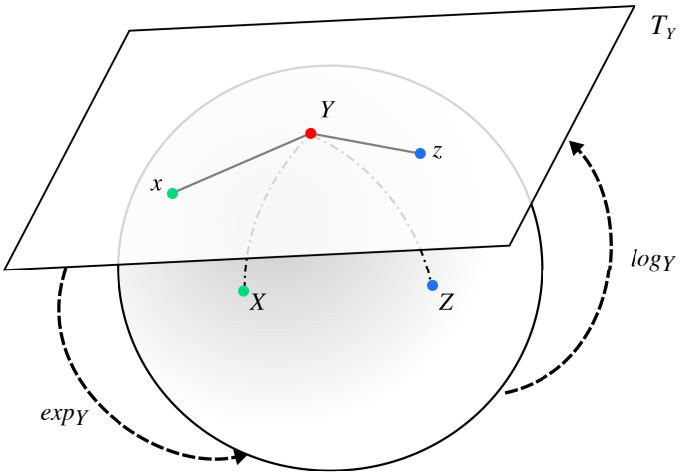


Figure 2.2: Mapping of points in Sym_d^+ manifold to the tangent space T_Y.

In certain cases such as covariance descriptor similarity computation, which is equivalent to computing point distances in a single tangent plane, the projection point can be established to a common reference point such as the Identity matrix, and therefore the tangent mapping operators become:

$$\log(X) = U \log(D) U', \tag{2.6}$$

$$\exp(y) = U \exp(D) U', \tag{2.7}$$

where U and D are the elements of the single value decomposition (SVD) of $X \in Sym_d^+$.

The Riemannian metric providing the distance between two points $X, Y \in Sym_d^+$ is given by the length of the geodesic curve connecting those points on the manifold [Förstner and Moonen, 1999; Pennec et al., 2006]. Provided that this would be equivalent to the norm of the tangent vector from X to Y in the tangent

space, the geodesic distance is defined as $d(X, \exp_X(y)) = \|y\|_X$. Using equation 2.4, an invariant Riemannian metric on Sym_d^+ is defined as:

$$\delta(X,Y) = \sqrt{Trace\left(\log\left(X^{-\frac{1}{2}}YX^{-\frac{1}{2}}\right)^2\right)}, \tag{2.8}$$

or more simply $\delta(X,Y) = \sqrt{\sum_{i=1}^{d}\log(\lambda_i)^2}$, where λ_i are the positive eigenvalues of $X^{-\frac{1}{2}}YX^{-\frac{1}{2}}$.

Other metric distances as Bregman divergences have been proposed in last years research [Cherian et al., 2013; Harandi et al., 2014b]. These divergences are focused to higher dimensionality covariance matrices via the use of Sym_d^+ specific kernels. In applications as covariance-based descriptors in low-level feature spaces it might not be mandatory to assess the use of those divergences, nevertheless their implementation could be left as a possible future extension.

In order to minimize the error regarding a projection point, the mean of a set of covariance matrices is a good estimator. Due to the convexity of Sym_d^+ manifold, the mean of a set of points $X_{i=1..N}$ on a Riemannian manifold has to be approximated in order to satisfy:

$$\mu = \underset{X' \in Sym_d^+}{\operatorname{argmin}} \sum_{i=1}^{N} \delta^2\left(X_i, X'\right) \tag{2.9}$$

[Karcher, 1977; Moakher, 2005] propose several gradient descent procedures for the computation of the mean iteratively. In [Pennec et al., 2006] the so-called *Log-Euclidean weighted mean* using the tangent space re-projection for a finite set of points in Sym_d^+, $X_1, ..., X_n$ is presented as:

$$\mathbb{E}_{LE}(X_1, ..., X_N) = exp\left(\frac{1}{\sum w_i}\sum_{i=1}^{N} w_i log(X_i)\right) \tag{2.10}$$

Finally, for a minimal representation in the tangent space T_Y it is possible to exploit the property of the projected symmetric matrices of containing only $d(d+1)/2$ independent coefficients, in their upper or lower triangular parts. This yields to the following vectorization operation in order to obtain a linear orthonormal space for these independent coefficients:

$$\hat{x} = vect(x) = \left[x_{1,1}\ \sqrt{2}x_{1,2}\ \sqrt{2}x_{1,3}\ ...\ x_{2,2}\ \sqrt{2}x_{2,3}\ ...\ x_{d,d}\right] \tag{2.11}$$

where x is the mapping of $X \in Sym_d^+$ to the tangent space, resulting from equation 2.4. The vectorization of a $d \times d$ covariance matrix is a minimal representation of all its $d\,(d+1)/2$ independent coefficients, which are found in the upper or lower triangular part of the matrix. As the off-diagonal entries would be counted twice in a norm computation, they are scaled down in this operation by the $\sqrt{2}$ coefficients. The obtained vector $\hat{x}$ will lie in the Euclidean space $\mathbb{R}^m$, where $m = d(d+1)/2$.

2.4 Human Body Detection

Human bodies are "objects" naturally constituted by parts with high variability, which have traditionally posed a great challenge in computer vision applications as classification and tracking. Even if supervised part-based learning approaches for human body part inference and pose estimation have been presented [Felzenszwalb et al., 2010; Yang and Ramanan, 2013], the method presented in this section is aimed at full human body detection in 2D images with a focus on the discriminative capabilities of the introduced covariance-based descriptors. Using the previously defined Riemannian geometry operators, this section reviews a part-based classifier paradigm in the form of manifold atlas-based learning, as an introduction to the development of algorithms with manifold designed constraints.

Covariance-based descriptors capture the amount of feature variability relative to a 2D image region, losing any structural information. This should offer tolerance to intra-class noise such as person pose or scale transformations. The geometric distribution of the descriptor plays also a major role in the methodology presented in this section. As different descriptor samples may constitute a sparse topology on the Riemannian manifold –due to the implicit object variation–, a single projection to the tangent plane would not be accurate. Therefore the integration of these conditions in a boosting framework is introduced, which provides a strong classification criterion as the result of a collection of weak classifiers. In this case, the weak classifiers will be represented by separation planes learnt by regression in the tangent space relative to a subset of covariance descriptors. This provides an atlas based model which, iteratively, learns the complete manifold distribution and naturally contributes to a part-aware classification methodology.

This section is a recap of the approach presented in [Tuzel et al., 2008b] and is intended to provide a proof of concept about the suitability of covariance-based descriptors for certain applications, along with an example of the integration of standard machine learning techniques to the particular manifold topology of the descriptor space.

2.4.1 Riemannian Manifold Boosting

The simplest classifier in a $\mathbb{R}^n$ linearly separable Euclidean space can be taken as an example: a point and a directional vector can define a discriminative function that divides the data space into two class regions. Equivalently, in a n-dimensional differentiable manifold, a point and a geodesic vector could be used as the definitional elements for expressing a separation curve on the manifold. Thanks to the exponential mapping function, this curve can be found in its projection as a vector into the tangent space relative to a reference point. But, as reviewed in previous sections, the projection to a tangent space is only valid for small neighbourhoods of the Riemannian manifold, therefore there does not exist a unique mapping that preserves the distances of the points for all the elements on the manifold.

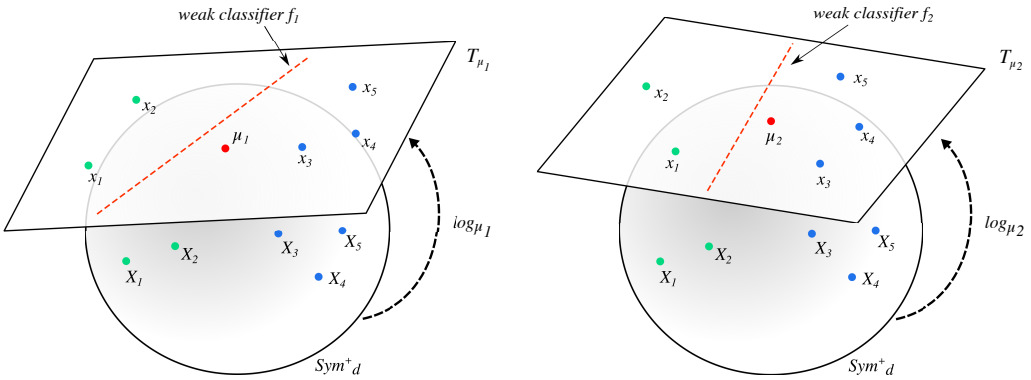


Figure 2.3: Example of two iterations of the boosting procedure for learning weak classifiers for a given set of samples on the manifold. The tangent space with respect to each iteration mean provides the best weak classifier obtained by regression on the tangent space. Then the mean is relocated thanks to the obtained function, and provides a new chart of the manifold atlas for the next step projection. This procedure is repeated until a minimum projection error threshold is achieved or a maximum number of weak classifiers are obtained, keeping the set of projection means and weak classifiers as the manifold classification atlas.

A possible approach for accurately learning a classification function on top of the complex distribution of points on the manifold is to provide an iterative approach for training several classification functions in tangent spaces that act as charts of an atlas of the manifold. These classifiers will be aggregated on a single classification criterion in a boosting approach [Schapire, 2003]. Boosting is considered as a general meta-algorithm for performing supervised learning by a set of simple classification functions, so-called weak learners, which work together in

benefit of a single strong learner. This approach serves two purposes: preserving the structure of the manifold and learning the classifiers on the associated tangent spaces by standard machine learning techniques.

For obtaining the minimum error of a set of points $X_i \in Sym_d^+$ with respect to the tangent space projection on a point $Y \in Sym_d^+$, we intuitively want to minimize the following error expression along the different charts of the manifold atlas:

$$\epsilon_Y = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\delta(X_i, X_j) - \|vect(\log_Y(X_i)) - vect(\log_Y(X_j))\|_2\right)^2 \qquad (2.12)$$

which evaluates the sum of squared pairwise differences between distances of points in the manifold (equation 2.8) and their homologous distances on the tangent space projection with respect to Y (using the vectorization and tangent mapping operations defined in equations 2.11 and 2.4 respectively).

Since the mean of a set of points of the manifold (equation 2.10) is precisely the point at the minimum distance of all the samples in the set, it can be used as the projection point of each atlas chart. In order to classify all the points of the manifold with the minimum error, the mean will be refined iteratively in the following boosting framework as shown in figure 2.3.

LogitBoost [Friedman et al., 2000] is a particular Boosting algorithm which casts the iterative learning of weak classifiers into a statistical framework. Weak learners are approximated by minimizing the negative binomial log-likelihood of the supervised learning data. That is, considering a binary classification problem with the labels $y_i \in \{0, 1\}$, ("no-person", "person") the probability of a sample X of being in class "1" is given by:

$$p(X) = \frac{e^{F(X)}}{e^{F(X)} + e^{-F(X)}}, \qquad (2.13)$$

where $F(X)$ is defined as the strong aggregation of L weak learners $\{f_l(X)\}_{l=1..L}$:

$$F(X) = \frac{1}{2} \sum_{l=1}^{L} f_l(X) \qquad (2.14)$$

The set of weak classifiers $\{f_l(X)\}_{l=1..L}$ can be learnt by fitting a weighted least squares regression function of the training samples (in its tangent space projected vectorized form, $x_i = vect(log_{\mu_l}(X)) \in \mathbb{R}^m$). μ_l is the projection point of the current manifold chart, which can be relocated iteratively thanks to the response values z_i and weights w_i of the current weak classifier evaluation and the supervised samples likelihood (equation 2.13):

$$z_i = \frac{y_i - p(X_i)}{p(X_i)(1 - p(X_i))}, \tag{2.15}$$

$$w_i = p(X_i)(1 - p(X_i)) \tag{2.16}$$

As a summary, with this boosting procedure we are learning the functions $f_l(x) : \mathbb{R}^m \mapsto \mathbb{R}$ by regression in each tangent space and the projection mean associated to each weak classifier, $\mu_l \in Sym_d^+$. The full procedure is presented in algorithm 2.1. Finally, the classification decision of an unknown sample can be obtained by the aggregation of the learnt set of weak classifiers as depicted also in figure 2.4:

$$class(X) = sign\left[\frac{1}{2}\sum_{l=1}^{L} f_l(X)\right] \tag{2.17}$$

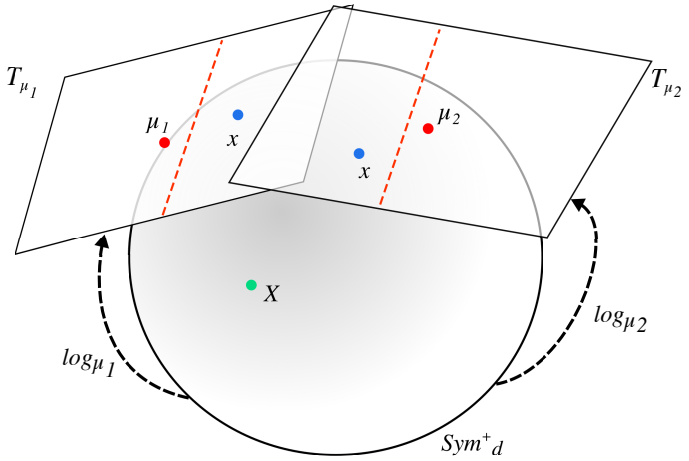


Figure 2.4: Classification evaluation of a sample $X \in Sym_d^+$ by its projection to the atlas of tangent spaces T_{μ_l} and the evaluation of the respective weak classification functions. The final classification decision is the aggregation of the sign of X with respect to these functions.

2.4.2 Cascade of Riemannian LogitBoosts for Part-based Classification

The introduced Riemannian LogitBoost approach is a feasible method for classification of a set of covariance-based descriptors but human bodies, being part-based objects, can not be ideally defined by a single descriptor model covering

Algorithm 2.1: One-level Riemannian LogitBoost learning algorithm

input : Training set $\{(X_i, y_i)\}_{i=1..N}$, $X_i \in Sym_d^+$, $y_i \in \{0, 1\}$
L max. number of weak classifiers
output: Strong classifier function as a set $\{F_l\} = \{\mu_l, f_l\}_{l=1..L}$

Initialize
$F(X_i) = 0$;
$p(X_i) = \frac{1}{2}$;
$w_i = 1/N$;

for $l \leftarrow 1$ **to** L **do**
- Compute response of samples $\forall i$ at the current iteration:
$z_i = (y_i - p(X_i))/w_i$;

- Compute weighted mean of the points through eq. 2.10:
$\mu_l = exp\left(\frac{1}{\sum w_i} \sum_{i=1}^{N} w_i log(X_i)\right)$
- Map each training point to the tangent space at μ_l:
for $i \leftarrow 1$ **to** *# of training samples* **do**
$x_i = vect_(log_{\mu_l}(X_i))$;
end

- Fit the f_l function by by *weighted least squares regression* of samples x_i to response values z_i using weights w_i.

- Update strong classifier, sample class probabilities and weights for all samples X_i:
$F(X_i) = F(X_i) + \frac{1}{2} f_l(x_i)$;
$p(X_i) = \frac{e^{F(X_i)}}{e^{-F(X_i)} + e^{F(X_i)}}$
$w_i = p\left(X_i\right)\left(1 - p\left(X_i\right)\right)$;

- Store in the output set: $\{F_l\} = \{\mu_l, f_l\}_{l=1..L}$;
end

the complete human body variability. Therefore, a part-based classification is the natural extension to deal with this situation. Provided that covariance-based descriptors are fast to compute, [Tuzel et al., 2008b] developed a cascade scheme of part-oriented LogitBoost classifiers, where region descriptors are obtained in a greedy manner. A rejection cascade organization is a simple way of applying a set of sequential classification methods to an unknown sample in order to improve accuracy or, as it is our case, also to introduce specific part-based classification information.

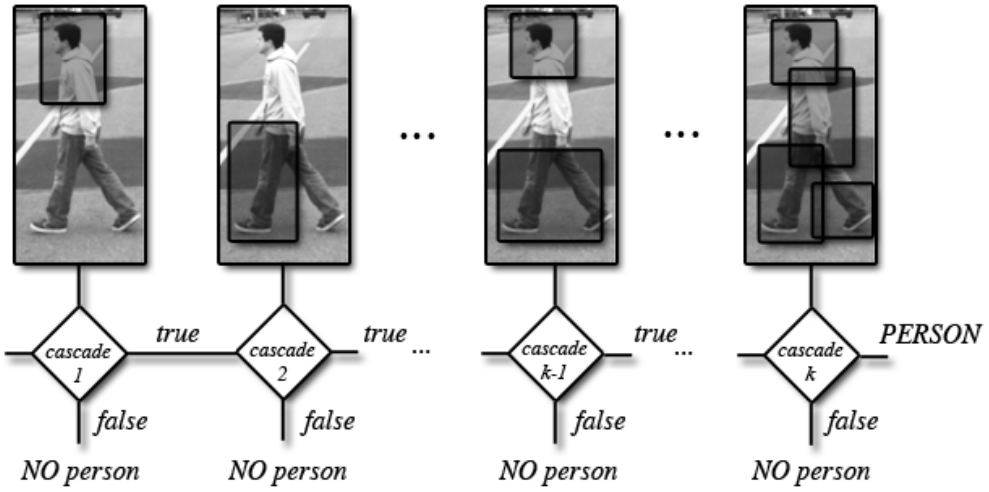


Figure 2.5: Scheme of Cascaded Riemannian LogitBoost classification.

Besides this straight part-based classification motivation, a cascade scheme provides the main enhancements of allowing the learning of specifically part-oriented classifiers, with their own optimal manifold atlas, via an independent error minimization and a variable number of weak learners tailored to each cascade stage. A rejection cascade classification acts as follows: provided a set of $N = N_p + N_n$ positive and negative learning samples, respectively, each cascade level k is trained in order to classify the set of $\{X_i^-\}_{i=1...N_n}$ negative examples (as depicted in figure 2.5). Samples correctly classified are removed from the learning set, therefore the cascade focuses on correctly training the remaining elements at level $k+1$, improving accuracy. In order to adapt the number of weak classifiers at each cascade level, a margin constraint is imposed. If one wants to determine that a cascade level k is optimized to detect at least the 99 percent of the available positive examples N_p^k, and to reject at least the 35 percent of the negative

examples N_n^k, the margin constraint is defined as:

$$margin_k = p_k(X_p) - p_k(X_n); \tag{2.18}$$

where X_p will be the positive sample having the $(0.99N_p)^{th}$ largest probability over the positive samples, and X_n is the negative sample having the $(0.35N_n)^{th}$ smallest probability over the negative samples. Weak classifiers at cascade level k are being added while this constraint is under a threshold, e.g: $margin < 0.2$. When this is no more satisfied, the cascade level k is considered as completely trained and a decision boundary for that stage is stored, $threshold_k = F_k(Rn)$.

Computationally, the final learnt model is based on the set of K LogitBoost classifiers, where each level k also stores the relative coordinates of used part subregions (which are selected in a greedy manner, provided that images are aligned and normalized):

$$\{F_k\} = \{(\mu_{k,l}, f_{k,l}, r_{k,l})\} \tag{2.19}$$

where $r_{k,l} = [x, y, x', y']$, the part subregion coordinates. The complete methodology is provided in algorithm 2.2.

Finally, the classification of a given sample X is obtained by the evaluation of the following expression for each level k:

$$\text{class}_k(X) = \text{sign}\left[\sum_{l=1}^{L_k} f_{k,l}(\text{vect}(\log_{\mu_{k,l}}(X_{r_{k,l}}))) - threshold_k\right] \tag{2.20}$$

2.4.3 Discussion

The methodology introduced in this section has been implemented in MATLAB with the main purpose of testing the value of an atlas-based manifold learning algorithm, and gaining expertise on the development of machine learning algorithms introducing Riemannian geometry constraints in its formulation. A complete qualitative and quantitative experimental evaluation of this method is provided in [Tuzel et al., 2008b] in terms of image classification and detection accuracy. Tuzel *et al.* include an exhaustive experimental set-up on top of the *INRIA* pedestrian dataset [Dalal and Triggs, 2005] and the *DaimlerChrysler* dataset [Munder and Gavrila, 2006], which demonstrates how this new family of descriptors can outperform standard descriptors, such as linear classifiers using Histogram of Oriented Gradient features. Using the *INRIA* dataset we assessed the accuracy performance of the cascade boosting classifier, reproducing the experimental results of Tuzel *et al*.

Algorithm 2.2: Cascade Riemannian LogitBoost learning algorithm

input : Training set $\{(X_i, y_i)\}_{i=1..N}$, $X_i \in Sym_d^+$, $y_i \in \{0, 1\}$
K = max. number of cascade levels;
margin = 0.2;
R = 200; (max. number of part subregion window candidates)

output: K cascade model $\{F_k\} = \{(\mu_{k,l}, f_{k,l}, r_{k,l})\}$

for $k \leftarrow 1$ **to** K **do**
  Prepare set of negative learning examples: classify $\{X_i^-\}_{i=1...N_n}$ with $(k-1)$ classifier cascade levels and remove correctly classified samples.
  Initialize
  $F_k(X_i) = 0; p_k(X_i) = \frac{1}{2}; w_i = 1/N;$
  $X_p = (0.99N_p)$-th X^+;
  $X_n = (0.35N_n)$-th X^-;

  while $p_k(X_p) - p_k(X_n) < margin$ **do**
    - Compute response of samples $\forall i$ at the current iteration:
    $z_i = (y_i - p(X_i))/w_i;$
    - Compute random subregion covariance descriptors:
    $\{X_{subregion_{k,l}}\}_{t=1...R}$
    for $subregion \leftarrow 1$ **to** R **do**
      - Perform weak learning steps as in algorithm 2.1:
      - Compute $\mu_{k,l}$, map subregion points to the tangent space, and fit the $f_{k,l}$ function by by *weighted least squares regression* of samples $\{X_{subregion_{k,l}}\}$ to response values z_i and weights w_i.
    end
    - Choose best subregion classifier $f_{k,l}$ s.t. minimizes:
    $l(y, p(x)) = -\sum_{i=1}^{N} [y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))]$
    - Update strong classifier, sample class probabilities and weights for all samples X_i:
    $F(X_i) = F(X_i) + \frac{1}{2} f_l(x_i);$
    $p(X_i) = \frac{e^{F(X_i)}}{e^{-F(X_i)} + e^{F(X_i)}}$
    $w_i = p(X_i)(1 - p(X_i));$
    - Sort samples by descending $p(X_i)$. Update X_p and X_n
  end
  - Store in the output set: $\{F_k\} = \{(\mu_{k,l}, f_{k,l}, r_{k,l})\}$ and threshold$_k = F_k(Rn)$.
end

Nevertheless, the main interest of this section is focused on assessing the impact of dividing the manifold in charts of an atlas, so using the same experimental conditions as Tuzel *et al.* (splitting the *INRIA* dataset into a training set of 2416 positive samples and 1218 negative samples, using a set of 1132 positive samples and 453 samples for testing, and rescaling the sample images to 64×128 pixels for descriptor computation) it was intended to test the effect of a growing number of cascade stages on the overall accuracy level of the classifier. Figures 2.6 and 2.7 present, respectively, the results of overall accuracy $(TP+TN/TP+FP+FN+TN)$, precision $(TP/TP+FP)$ and specificity $(TN/TN+FP)$ of the classification method according to a different number of evaluated cascade stages for classifying the testing set. As expected, these levels increase as long as cascade levels are increased as well. One of the benefits of the cascade mechanism is that it preserves a high level of specificity, as by design each classifier level is trained with the constraint of performing good negative sample rejection. Additionally, an overall performance convergence in terms of accuracy can be observed for a given number of used cascade stages, which indicates that a further division of the manifold into subsequent atlas charts would not be necessary.

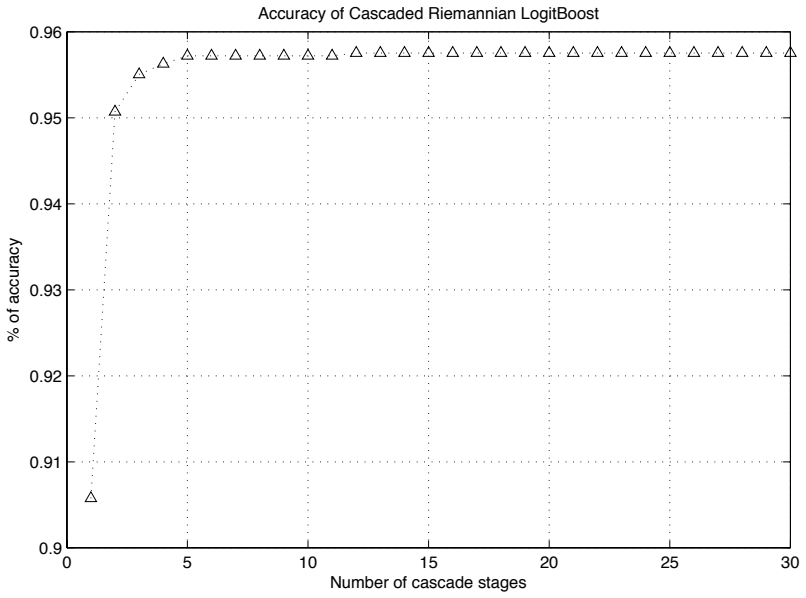


Figure 2.6: Accuracy of the classifier vs. number of stages used for rejection.

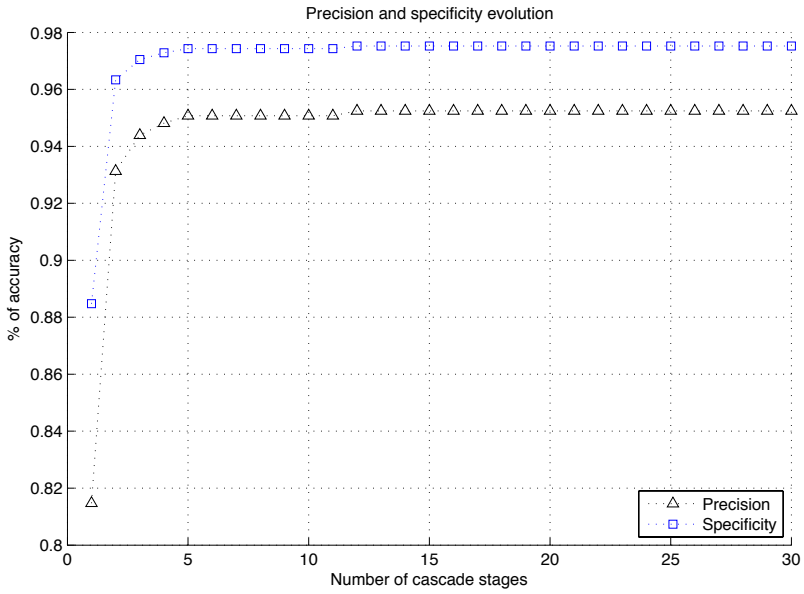


Figure 2.7: Precision and specificity values vs. number of cascade stages.

2.5 Medical Image Retrieval

The challenge in medical image classification and image-based retrieval comes from several handicaps: low availability of training samples, subtle changes between different image sources, differences on data origin, or low-level feature variabilities [Müller et al., 2004]. The ImageCLEF image retrieval challenge [Villegas et al., 2015] provides a benchmark to test the impact of different image classification and feature selection methods, specially those using visual and/or textual information in medical image classification [García Seco de Herrera et al., 2015]. As previously reviewed, in the computer vision area research many pattern recognition methods have been developed for image classification. Most of them include the development of content and feature selection functions, or the usage of keypoint extractors and associated descriptors which can be later categorized by supervised classification methods (support vector machines, boosting, neural networks, etc). This chapter section is based on the submission of a covariance-based descriptor image classificatio method to the medical subfigure classification task of ImageCLEF, which provides 30 different classes including diagnose images (radiology, visible light photography, microscopy, etc.) and also generic biomedical illustrations. More details on the challenge task can be found in [García Seco de Herrera et al., 2015].

The presented approach adapts the already introduced framework with color-aware feature vectors, considers covariance-based descriptors as discriminative signatures for whole images, and formulates a sparse representation based classification approach for learning a dictionary of medical image classes on the descriptor manifold. Special motivation comes by the demanding conditions found in the different images of the medical classification subtask. The evaluation results are presented and discussion of this methodology arises from the direct output on the challenge participation. We are particularly interested in seeing if this proposed description, using purely visual information, is discriminative enough with respect to methods from other participants which are based on textual information rather than only visual features. This would assert the feasibility of the presented methodology and proof that its performance can be on par with other methods which use also complementary textual features for complex image retrieval.

2.5.1 Covariance-based Descriptor for Medical Images

An inspection of the provided data of this medical image retrieval challenge makes evident that class separation from purely visual cues is not a trivial task. Different image sources might share visual features, or suffer from a lack of discriminative salient cues (see figure 2.8), and images often have different sizes. This yields to the intuition of what should be taken into account, and how the covariance-based

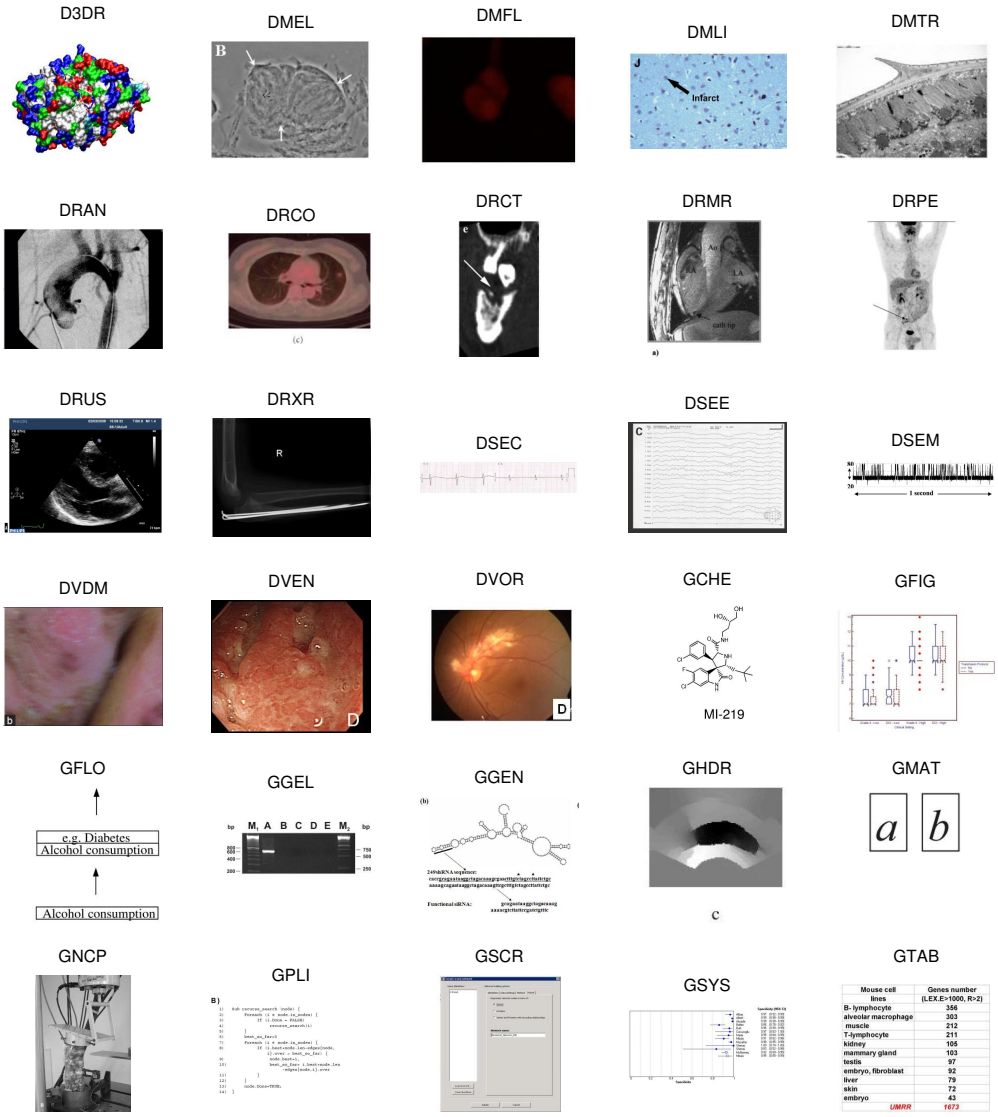


Figure 2.8: Example of different samples of the different 30 classes present on the ImageCLEF medical classification task. More details on class hierarchy and terminology can be found in [García Seco de Herrera et al., 2015]

descriptor framework suits this application. There are several information cues that are equally important: not only texture patterns, but also color, sparsity, structure features... And even more important than the features themselves is the fact that the modelling must take into account all the feature interactions together. For instance, a diagram figure in a medical publication can be in grayscale just as an electron microscopy image, but structural features in a diagram contain pure lines

or geometrical shapes which are not present on a biological tissue captured by the microscope. At the same time, different microscopy devices might capture similar natural tissue patterns, but a visible light microscope can capture a different range of color spectra than a transmission microscope. Therefore, in an analogy with a natural visual perceptual system, the goal is to model the space of different visual cues and their joint relationships, just as the notion that is embodied by the proposed covariance-based descriptor framework.

An ideal image representation must encode all images in a common compact, size invariant notation regardless of the different image sizes. In order to formally define this 2D color feature based covariance descriptors, the following feature selection function $\Phi_{2D}(I)$ for a given image I is denoted as:

$$\Phi_{2D}(I) = \{\phi\_med_{x,y} \forall x, y \in I\}, \tag{2.21}$$

which provides a set of feature vectors $\phi\_med_{x,y}$ for each one of the pixel coordinates $\{x, y\}$ inside all the image I. These 11-dimensional feature vectors are expressed as:

$$\begin{aligned} \phi\_med_{x,y} = \Big[& x, y, R_{x,y}, G_{x,y}, B_{x,y}, |I^x|_{x,y}, |I^y|_{x,y}, \ ... \\ & ... \ |I^{xx}|_{x,y}, |I^{yy}|_{x,y}, \sqrt{(I^x)^2_{x,y} + (I^y)^2_{x,y}}, \arctan \frac{|I^x|_{x,y}}{|I^y|_{x,y}} \Big], \end{aligned} \tag{2.22}$$

which are similar to the features used in equation 2.2 but now include RGB color values in addition to pixel spatial coordinates, first and second order image intensity derivatives and their magnitude and pixel curvature. These cues provide information about the color distribution of a given image class, as well as their texture patterns and visual structure –as found in the first and second order gradient and curvature features. A schema of these features is depicted in figure 2.9 Then, for a given color image I the covariance descriptor associated to the whole picture can be obtained as the regular covariance matrix introduced in previous sections:

$$CovMed\left(\Phi_{2D}(I)\right) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\phi\_med_i - \mu\right) \left(\phi\_med_i - \mu\right)^T, \tag{2.23}$$

where μ is the vector mean of the set of vectors $\{\Phi_{2D}\}$ within the image I.

The resulting 11×11 matrix $CovMed$ shares the same structure and properties as previously defined covariance-based descriptors: it is a symmetric matrix where the diagonal entries will represent the variance of each feature channel, and

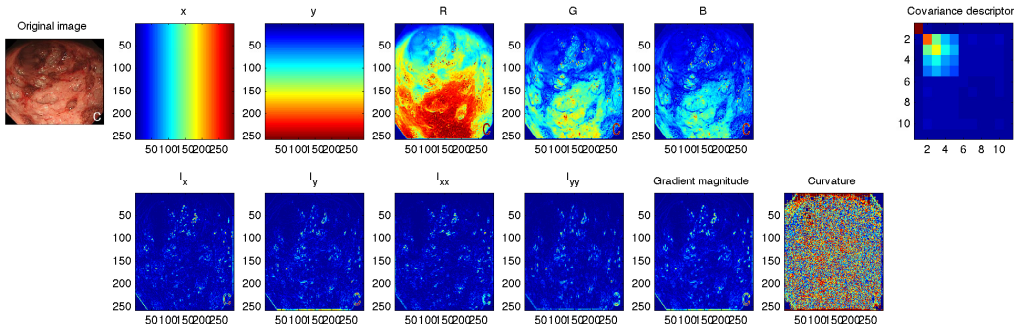


Figure 2.9: Different cues involved in the descriptor building for an image of the endoscopy class (leftmost subimage). The resulting $CovMed$ covariance descriptor is shown in the rightmost sub-figure. Images of the same class share similar covariance descriptor signatures, while images from classes with different color distributions and shape features have differentiated descriptors.

the non-diagonal elements represent their pairwise covariance. This descriptive signature is robust to intraclass spatial transformations, such as rotations and scale transformations: even if pixels are translated or scaled, their unstructured collection for the covariance description will share the second order moment statistics leading to an equivalent matrix. Finally, this image level family of descriptors does not depend on computationally loading intermediate stages, such as keypoint extractions, and provides a compact signature for images of any given size.

2.5.2 Manifold-regularized Sparse Representation for Classification

For the classification of the proposed 2D color feature based $CovMed$ covariance descriptors this section introduces a sparse representation based classification method, taking into account their Riemannian geometry for learning the dictionary of the different class topologies found in the descriptor space. The submission of this method to the ImageCLEF medical classification task has served two purposes: in a first instance it tests the performance of this approach in the heterogeneous class distribution found in the provided medical image dataset. In a second place it compares the performance of the presented approach against other participants which use other textual-based retrieval methods. This can provide an unbiased, quantitative comparison scale for proofing the concept that a purely visual classification method can be on par with more complex text-based retrieval approaches.

The topological layout of the proposed covariance-based descriptor yields to focus on a geometrically sensitive classification method which can exploit the Riemannian spatial distribution of the descriptors. Sparse representation based methods [Wright et al., 2010; Zhang et al., 2011] have shown a recent rise in the

machine learning community in the context of face recognition. In this application, two key concepts are very relevant: sparsity and collaboration. They are related to the complexity of the model learning: not only because a complete set of learning samples is hardly available, but also because an unknown element can share characteristics from different classes. As this also the case in medical image retrieval, where images from a particular class might be scarce and the low-level visual cues provide a complex class definition, we propose a new sparse method formulation adapted to the manifold of 2D color based covariance descriptors.

In its general formulation, sparse representation based classifiers propose to consider a test sample c as a linear combination of elements in a dictionary A of training samples from different classes: $c = A\alpha$, where α is the sparse vector indicating the weight coefficients for each element in A. As the sample c should ideally be represented by using the less number of samples, and as accurate as possible, α is found forcing its sparsity via its L1 norm minimization constrained as follows:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left\{ \|\alpha\|_1 + \|c - A\alpha\|_2^2 \right\} \tag{2.24}$$

Then, given $\hat{\alpha}$, the classification label for c is determined by the subset of training samples of a given class i which provides the minimum representation error:

$$class(c) = \underset{i}{\operatorname{argmin}} \left\{ error_i \text{ s.t. } error_i = \|c - A_i\hat{\alpha}_i\|_2 \right\} \tag{2.25}$$

This initial approach shares similar fundamentals than the classical nearest neighbour or nearest subspace classifiers. Equation 2.24 intuitively represents an unknown sample as a possible combination of all elements in A, but this "collaboration"' is discarded afterwards as the minimization of the residuals in equation 2.25 determines the closest distance to only a single class with the minimal representation error, for a unique class decision.

This suggests a main concern: if some subsets of different classes i and j in the training set, A_i and A_j, are correlated due to similarities in the elements of each class then the distance between two class reconstructions $\|error_i\|_2$ and $\|error_j\|_2$ could be very small leading to possible misclassifications. A solution would be to avoid the L1 norm sparsity minimization constraint in equation 2.24, and express the test sample c collaboratively on all the dictionary samples $X = [A_1, A_2, ..., A_n]$ without forcing any class sparsity prior: then the linear representation solution could be treated as a classical least squares minimization problem:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left\{ \|c - A\alpha\|_2^2 \right\} \tag{2.26}$$

The main problem is that the solution to this minimization may become unstable and computationally expensive if the number of classes is too big (more details can be found in [Zhang et al., 2011]).

Considering all these details, this section is providing a sparse representation minimization expression taking into account the prior knowledge found in a learning dictionary of samples in the manifold, by means of their tangent space projection atlas and a regularization term taking into account the distances on the manifold. Figure 2.10 depicts a schema of this classification paradigm.

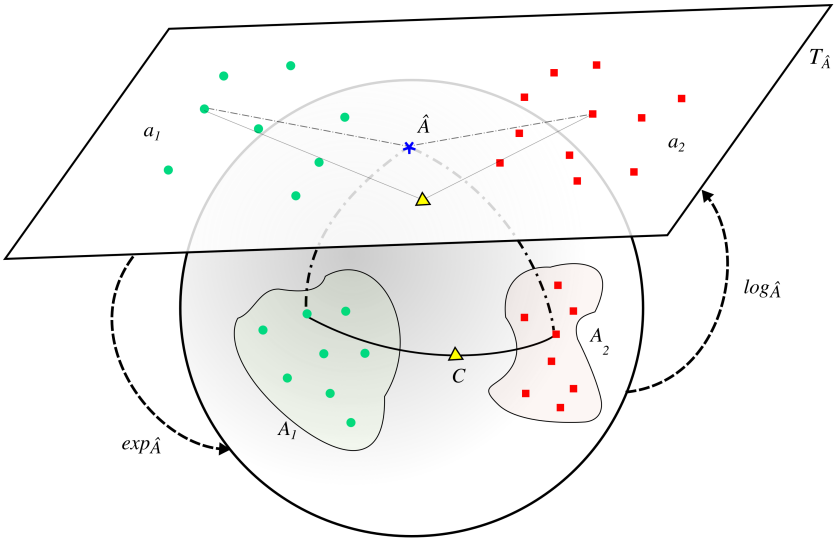


Figure 2.10: Schema of the sparse classification method on top of the covariance descriptor manifold.

Let A be the whole set of n training samples, in its vectorized form according to equation 2.11, from K different classes: $A = [A_1, A_2, ..., A_K] \in \mathbb{R}^{66 \times n}$, where each $A_i = \left\{\text{vect}(\log_{\hat{A}_i}(CovMed_i))\right\}$ is the set of vectorized covariance descriptors which form the subset of training samples for the class i. And let $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K]$ be a vector of weights corresponding to each one of the training samples in A. Then, the sparsity restriction on α can be achieved via its L2 norm minimization, proposing a manifold-aware minimization constraint which relaxes the computational expense of the method and adds numerical stability:

$$\hat{\alpha} = \underset{\alpha}{\text{argmin}} \left\{ \|C - A\alpha\|_2^2 + \|D\alpha\|_2^2 \right\} \tag{2.27}$$

where D is a diagonal matrix of size $n \times n$ which allows the imposition of prior knowledge on the solution with respect to the training set, using the Riemannian metric defined in equation 2.8. This term contributes also on making the least

squares solution stable, and on introducing forehand sparsity conditions to the vector $\hat{\alpha}$ as well. D is defined as:

$$D = \begin{pmatrix} \delta(A'_1, C') & & 0 \\ & \ddots & \\ 0 & & \delta(A'_n, C') \end{pmatrix} \tag{2.28}$$

where A'_i and C' are the unvectorized covariance descriptors for training and test samples respectively. The solution to the sparse collaborative representation, $\hat{\alpha}$, can be calculated by the following derived expression according to [Zhang et al., 2011]:

$$\hat{\alpha} = \left(A^T A + D^T D\right)^{-1} A^T C \tag{2.29}$$

Finally, the classification label of the test sample C can be obtained by observing the regularized reconstruction residuals from the resulting sparse vector $\hat{\alpha}$:

$$class(C) = \underset{i}{\operatorname{argmin}} \left\{ \frac{\|C - A_i \hat{\alpha}_i\|_2}{\|\hat{\alpha}_i\|_2} \right\} \tag{2.30}$$

2.5.3 Results

In order to evaluate the method, the medical classification task in the ImageCLEF challenge provided a set of more than 4500 images of the different classes for learning, which were used for modelling the dictionary of covariance descriptors for each class and the candidate test labels after the closed form of equation 2.29. Participants were allowed to submit the classification decision labels on a test set of 2244 images, which were estimated according to the provided formulation. A cloud-based service was provided for collecting the participation and evaluating each participant submission performance. The evaluation score used on the task performance assessment is the classification accuracy ratio for all the classes, computed as the ratio of true positives and negatives over the total number of samples. The top results are collected in table 2.1, which are also publicly available on the challenge website [1].

Before the submission of the approach to the ImageCLEF image retrieval challenge, the presented method was tested on the provided training data set, using a 10-fold cross-validation. Each fold was adapted so at least 20% of samples of each class were kept in each subset. In classes with a very low number of samples which would cause to have some folds without class representation, some samples where duplicated. Therefore, classes with very few samples where guaranteed to be balanced and represented on the training set of the classification method. After

[1] http://www.imageclef.org/2015/medical

Method	Features	True positive ratio
Participants 1	Visual + text	67.60
Participants 1	Only visual	60.91
Our method	Only visual	52.98
Participants 3	Only visual	45.63

Table 2.1: Top accuracy performances after submission evaluation of the ImageCLEF medical classification task. The presented method accuracy is placed after the most accurate method. Using only visual features it is close to the best method, which also exploits textual information associated to the training samples.

iterating the cross-validation runs, an average accuracy of 73.24 % was obtained. As it has been commented in the methodology description, the presented classifier arises as a method for expressing unknown samples as the best sparse representation regarding to a learning set. Therefore, this increase on the accuracy in this preliminary test evaluation is explained as a direct effect of the balancing preprocessing of those classes with very few elements.

Table 2.2 presents the different precision and recall values for each class once the groundtruth annotations of the testing set were made publicly available, and observes if there is a particular correlation between these values and the different sample size of each class. As it can be observed, mainly due to the nature of the data or the difficulty on acquiring and labelling images from certain classes, the set of images is clearly unbalanced, which affects the method performance. Besides the accuracy evaluation of the presented approach, this provides a valuable focus on a previous balancing stage which we will integrate into the presented method in future applications, as described in chapter 6.

2.6 Conclusions

This chapter has provided an introduction to covariance-based descriptors for 2D images. This descriptor translates the domain of d-dimensional feature observations within regions of interest to the space of feature covariances, capturing the joint distribution of those feature variabilities. Properties of this family of descriptors include intra-class variability tolerance due to its construction, loss of structural information which leads to invariance to scale and spatial transformations, and a compact notation. The particular structure of covariance-based descriptors is also meaningful, as they lie on the Riemannian manifold of symmetric positive definite matrices, Sym_d^+.

Class	D3DR	DMEL	DMFL	DMLI	DMTR	DRAN	DRCO	DRCT	DRMR	DRPE
Class #	112	60	312	266	77	7	27	6	43	4
Precision	0.5300	0.1584	0.6629	0.6810	0.3875	0	0	0	0.1579	0
Recall	0.4732	0.2667	0.7436	0.5376	0.4026	0	0	0	0.1395	0

Class	DRUS	DRXR	DSEC	DSEE	DSEM	DVDM	DVEN	DVOR	GCHE	GFIG
Class #	0	20	0	4	1	12	4	17	8	764
Precision	0	0.0526	0	0	0	0.3333	0.1250	0.0217	0.1667	0.6600
Recall	0	0.0500	0	0	0	0.1667	0.2500	0.0588	0.5000	0.8154

Class	GFLO	GGEL	GGEN	GHDR	GMAT	GNCP	GPLI	GSCR	GSYS	GTAB
Class #	6	116	173	52	8	34	0	13	66	32
Precision	0	0.4806	0	0.0857	0	0.2143	0	0.0833	0	0.1707
Recall	0	0.5345	0	0.0577	0	0.0882	0	0.0769	0	0.2188

Table 2.2: Analysis of the cardinality of different classes in the testing set and their associated precision and recall values. These are clearly affected by the unbalanced class sets, which has a direct impact on the presented method due to its underlying formulation.

Two methodologies have been tested in order to proof the conception of this descriptor in two applications: first, in human body image detection, the part-based nature of this classification task has been the perfect excuse for exploring the implementation of an existing atlas-based boosting algorithm, where several areas of the manifold have been considered both for part-oriented classifiers and for error minimization of each weak classifier. The approach of [Tuzel et al., 2008b]has been selected in order to develop our own expertise so further methodologies could be formulated in our future research. In the second part of this chapter, covariance-based descriptors including color and other low-level visual cues have been tested in medical case image-based retrieval. In this case, whole images are represented by single descriptors with the goal of capturing their joint distribution of color and structure information, as the discriminative signature for identifying their class. Considering full-image descriptors provides the basis for modelling a dictionary of all the possible available classes, and a sparse representation based classifier has been developed. This second part of the chapter has been experimentally evaluated by its submission to the ImageCLEF medical classification challenge, which provides a testing benchmark and comparison framework with respect to classification methods from other participants.

Several conclusions can be derived: in terms of overall accuracy on both methods, the covariance of feature distributions has demonstrated to be a characteristic signature of image regions, allowing to use any particular feature extraction function under a same descriptor conception. These results settle the basis of the framework described in this dissertation and encourage the continuation of this research line in extended application directions, as 3D region descriptors with

a special focus in feature fusion for combining shape and texture information. Furthermore, covariance-based descriptors provide a meaningful, compact space which has straightforward geometric relation on class feature modelling. They use only low-level visual features and require very low computational cost for its construction. The conducted research has dealt with atlas-based and sparse dictionary approaches which are believed to suit other applications, such as modelling the complete space of 3D shaped textures or dense tissue in medical images. The participation on the ImageCLEF medical classification challenge has provided practical outcomes as well, such as opening the possibility of fusing textual and visual information on case retrieval, and identifying the need of balancing classes in particularly challenging datasets as patient and medical images.

Chapter 3

3D Scene Understanding

> "I ran into Isosceles. He had a great idea for a new triangle!"
>
> — WOODY ALLEN

DESCRIPTION, DETECTION AND MATCHING of points from different complex scenes is a challenging task for many computer vision applications on 3D point clouds such as object modelling, recognition or scene reconstruction. Existing approaches make use of all the available cues in the usual two channels of information: visual photometry such as color or textures, and shape and depth information from 3D sensors. While state-of-the-art methods have given successful outcomes in both areas, as further reviewed in this chapter, it is still encouraging to find a global method which can fuse information from both two worlds, and provide a descriptive unit which is able to encode surface definition and its correlated texture or pattern information together by adapting the covariance-based descriptor framework features. This will be supported with a global matching procedure specially aimed to complex scene understanding, so it can be observed from an overall perspective, providing scene-aware geometric constraints. This is crucial in order to cope with challenging conditions such as locally repetitive patterns or symmetries. The aim is to avoid ambiguities which can be reduced at all levels: both locally if a shape is defined in conjunction with its associated visual cues, and globally with a holistic match refinement procedure.

3.1 Introduction

This chapter presents a threefold contribution: first, the formulation of covariance-based descriptors is tailored in order to be able to gather shape and visual information together within a radial 3D area (see figure 3.1 for a concept example).

Figure 3.1: Example of a coherent visual and shape aware descriptor for matching in a 3D scene. While the visual appearance of the ball is similar on its real appearance and the paper printed representation, the matches should be correctly considered only on the true 3D points, since shape information should also encoded on the used descriptor for matching.

Thanks to its fundamentals, the descriptor is robust to noise changes, rigid spatial transformations or even resolution variations; and because of its low computational cost it can be extended to a multi-scale context for better discrimination performance. In a second place, intrinsic properties of the descriptor are reviewed: thanks to them it offers a procedure for keypoint extraction. Therefore, salient points in the scene (in terms of major color and shape variation areas) can be detected at the same stage where descriptors are being obtained. And finally, a game theory based solution method which integrates local descriptor similarity with global 3D geometric consistency for a possible registration of several scene views is provided. This method efficiently looks for the global minimal reconstruction error, taking into account all the available descriptor matches and avoiding local minima which other methods could reach due to symmetries or local repetitions. Indeed, the main key point of this chapter is showing how the descriptor framework can be extended with additional layers of geometric constraints in order to solve problems more related to the nature of described data, which are beyond the own descriptor formulation.

This chapter is organized as follows: section 3.2 reviews the state-of-the-art approaches and related work for the particular application of 3D scene understanding and registration. Sections 3.3 and 3.4 introduce the presented contribution in two separate sections: the covariance-based descriptor formulation for 3D shape

and texture fusion and scene analysis framework itself, and the game theory based solution approach for scene reconstruction. Section 3.5 presents and discusses the results, before concluding in section 3.6.

3.2 Related Work

3D scene registration is currently an active topic in the computer vision literature, recently compelled by advances in the sensors technology which have provided some affordable devices and acquisition techniques. This has eased the capture of 3D information to the mass public and also produced an increase in the processing proposals for this kind of images during the last years.

This topic has been however studied since some time ago from several perspectives. One of the first proposals, which is still currently considered as one of the main methods in the 3D registration area, is the Iterative Closest Point (ICP) [Besl and McKay, 1992]. Their method estimates the registration between two 3D point clouds, performing an iterative process in order to minimize the mean square distance between two sets. The main problem of this algorithm is the need of a good initialization if we desire that the iterative process converges to a global minimum and not to a local minimum. In order to achieve this initial approximation, the typical procedure consists in the establishment of some correspondences between specific points of the two 3D points clouds. Once these correspondences have been established, both subsets of points can be registered by solving the classical problem of absolute orientation [Horn, 1987].

Other state-of-the-art approaches for point set registration commonly use iterative algorithms such as RANSAC [Fischler and Bolles, 1981] or its variants [Chum et al., 2004; Chum and Matas, 2005, 2008] which allow the integration of geometric consistency as a measure for minimizing the correspondence error. This is basically an heuristic for comparing how a set of points fits within some geometric constraints: projections to a coordinate system, error measurements regarding a rigid transformation, or relative distances amongst connected point sets. Despite the existence of other possibilities, RANSAC is undoubtedly the predominant algorithm for the geometrically consistence situation in 3D registration, thanks to its good results and its standardized implementations. Authors like Johnson, in his Spin Images approach [Johnson, 1997], also propose his own geometric consistency algorithm based in the same conceptualization of the Spin Image, but he also finally refers to RANSAC as the appropriate technique for more problematic cases. In fact, RANSAC has also been used as the basis for well-known methods of 3D scene registration which do not even use the correspondences information and only rely on the iteratively search of the RANSAC algorithm, as shown in [Chen et al., 1999].

In any case, it is obvious that any registration process must rely on a previous search for correspondent points, which must take into account the similarity amongst these candidate matches. During last years, this selection of candidate correspondences has been achieved by using descriptors which encode exclusively the 3D information from the scene points and can provide similitude measures amongst points. Inside this category, Spin Images [Johnson and Hebert, 1999] is probably the most known method, representing the neighbourhood of each 3D point into a 2D image and later comparing it against other Spin Images by a simple correlation factor. Other popular 3D descriptors are the point signatures [Chua and Jarvis, 1997], 3D shape contexts [Frome et al., 2004], THRIFT [Flint et al., 2007] or, more recently, the Fast Point Feature Histograms (FPFH) [Rusu et al., 2009].

However, thanks to the availability of 3D scanners which can also capture texture information, some descriptors which encode simultaneously information from the 3D shape and the color have been recently published in the literature. Textured Spin Images [Brusco et al., 2005] are a good example of this trend. Novel approaches include the MeshHOG descriptor [Zaharescu et al., 2009], which performs a histogram of gradient of a neighborhood of a 3D point by using separately the texture information and the 3D curvature. In order to include both cues in the final descriptor, both representations can be directly concatenated. This same methodology is also used from the authors of the CSHOT descriptor [Tombari et al., 2011], which concatenates their SHOT descriptor [Tombari et al., 2010] and the color information. Other contributions as [Kovnatsky et al., 2012] follow a more geometric perspective, where manifold embedding procedures are used and photometric information is implicitly encoded as part of the coordinate projection parameters.

Once the different correspondences have been established by comparing the descriptors, this first set of matches can be filtered by the aforementioned iterative registration methods, in order to discard the correspondences which can be incorrect or not accurate enough. A current challenge at this stage is posed by effects like symmetries or repetitive patterns in the scene: a correspondence between two 3D points could seem correct if we look individually, but incorrect in a more global context. A global algorithm taking into account all the previously found correspondences must be defined, allowing to obtain at the end of the process a subset of the initial group of correspondences which are geometrically consistent between them keeping local similitude as well.

3.3 Covariance Framework for Scene Analysis

As previously mentioned in chapter 2, covariance matrices as descriptors have been applied to several domains in the computer vision context. Regarding 3D surface description [Fehr et al., 2012] was the only approach, up to the beginning of the research conducted in this thesis, which explored different combinations of features obtained from range images related under a covariance analysis framework. Taking this as preliminary work, the present dissertation has extended it in order to deal with 3D point cloud scenes, making use also of correlated color information and with a formulation which is proven to be invariant to rotations, viewpoint, noise and density variations.

3.3.1 Fusion of Shape and Visual Features

The choice of features for unstructured 3D point clouds area description has to be carefully designed in order to cope with the inherent viewpoint dependency in 3D scenes, in addition to the own fusion with associated texture information. To achieve this goal, a feature selection function $\Phi(p, r)$ for a given 3D point p and its neighbourhood within radius r in the scene is defined as:

$$\Phi(p, r) = \{\phi_{p_i}, \forall p_i \text{ s.t. } |p - p_i| \leq r\} \tag{3.1}$$

where ϕ_{p_i} is the vector of random variables obtained at each one of the points p_i within the radial neighbourhood, and is defined as:

$$\phi_{p_i} = [R_{p_i}, G_{p_i}, B_{p_i}, \alpha_{p_i}, \beta_{p_i}, \gamma_{p_i},] \tag{3.2}$$

This feature selection function includes the following observations that are robust to spatial transformations, as they are computed relatively to the point for which the descriptor is being obtained: first of all, the visual information is taken into account in terms of R, G and B color space values. α, β and γ values are angular measures which encode the surface information of the points within the descriptor center neighbourhood in the following way:

- α is the angle between the normal vector in p and the segment from p to p_i, and encodes the global concavity of the surface regarding the center of the descriptor.
- β is the angle between the same segment and the normal vector in p_i, and measures the local curvature at this point in the neighbourhood relative to the center p.
- γ is the angle between both normal vectors in p and p_i. Being a 3D angle, it helps encoding the local surface curvature in a non-ambiguous way.

Figure 3.2 shows an example of how these measures are obtained. As these selected features are relative measures in terms of shape description, their usage in the covariance descriptor formulation guarantees a rotation and view invariance, which is a desired behaviour in descriptor performance. RGB space color values also lose structural information and become observations of a sampling distribution within the covariance descriptor formulation, therefore they will become invariant to rigid transformations in the scene. Even if in a more formal sense an intermediate colour invariant projection must be performed for minimizing the impact of illumination variations and offering a true robustness to view changes, this is considered beyond the scope of the presented approach and could be considered as a future extension -thanks to the ease of the presented descriptor for including new features. In any case, for small descriptor localities, RGB color space values have demonstrated to be significant enough. Finally, variables are normalized in order to have an equivalent range both for angular and color measure.

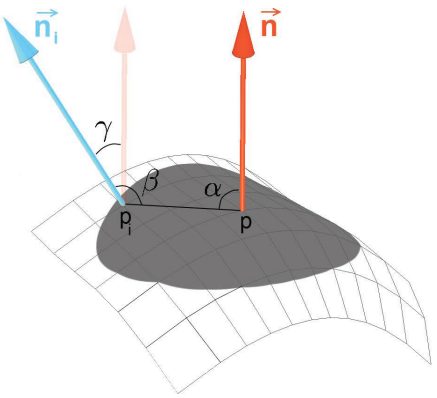


Figure 3.2: Scheme of the used features for shape information encoding. For each p_i in the neighbourhood of p, α, β and γ are the rotational invariant angular measures.

Then, for a given point p of the scene the covariance descriptor can be obtained by the usual covariance formulation as:

$$C_r\left(\Phi(p,r)\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_{p_i}-\mu\right)\left(\phi_{p_i}-\mu\right)^T \tag{3.3}$$

where μ is the vector mean of the set of vectors $\{\phi_{p_i}\}$ within the radial neighbourhood of N samples.

The resulting 6×6 matrix C_r will be a symmetric matrix where the diagonal entries will represent the variance of each one of the feature distributions, and the non-diagonal entries will represent their pairwise correlations.

A covariance descriptor can be seen as a high level and abstract representation which treats the observed features as samples of joint distributions, and loses all the spatial notion (information about the number of points and their ordering) within the region. This compactness provides a combination of flexibility -feature distributions will contribute to the descriptor still preserving their inner characteristics even under changes of scale and rotation in data- and robustness -according to central limit theorem, as long as a significant enough number of samples is used the data within a certain range within the features distribution will be correctly represented. In addition, these two facts yield a valuable performance boost in comparison to other descriptors based on more rigid representations such as histograms. Figure 3.3 shows an example of a covariance descriptor building from the proposed shape and texture features.

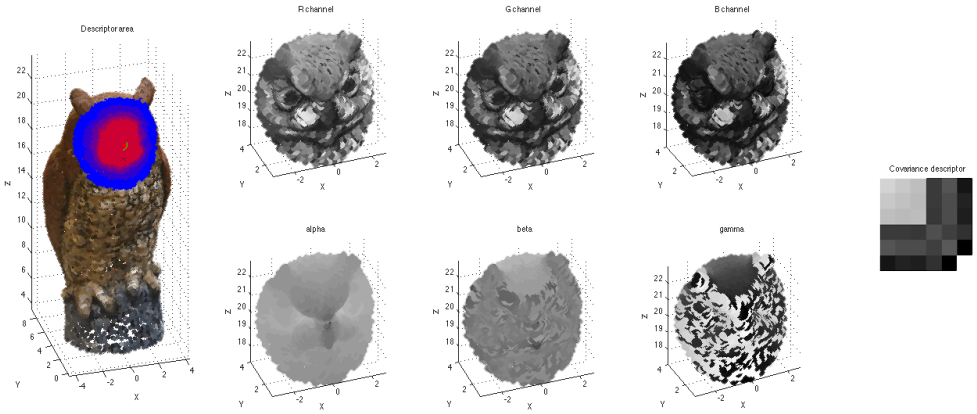


Figure 3.3: Example of a scene view where a multi-scale covariance descriptor is extracted on the face of an owl model. The left image shows the original 3D scene where the overlap gradient of colors from red to blue depicts 5 different scales used for obtaining a multi-scale descriptor. The 6 central subfigures show the different used features, in terms of color (upper row) and shape description (bottom row). Finally, on the right, a single scale 6×6 covariance descriptor is graphically represented.

3.3.2 Manifold Topology of the Descriptor

A remarkable consideration about the proposed descriptor is its geometrical topology. Covariance matrices, being symmetric positive definite matrices, do not lay on a Euclidean space, but on a Riemannian manifold. Indeed, covariance descriptors form the $d \times d$ dimensional space of symmetric positive definite matrices, where d is the number of used features ($d = 6$ in our descriptor approach), and

the main concern is that this descriptor space is meaningful for scene definition purposes as it abstractly represents a geometrical location of shape and texture distributions within a scene point area. This assertion can be visualized by a proof of concept as shown in figure 3.4. In an instance of a scene, descriptors have been extracted at different areas from different nature in shape and colour. Once the manifold distances amongst the set of descriptors have been computed, Multidimensional Scaling embedding onto a 2D coordinate space has been applied in order to graphically represent the consistency of the descriptor space. The plot demonstrates how different points coming from different areas in the scene are located in the descriptor space.

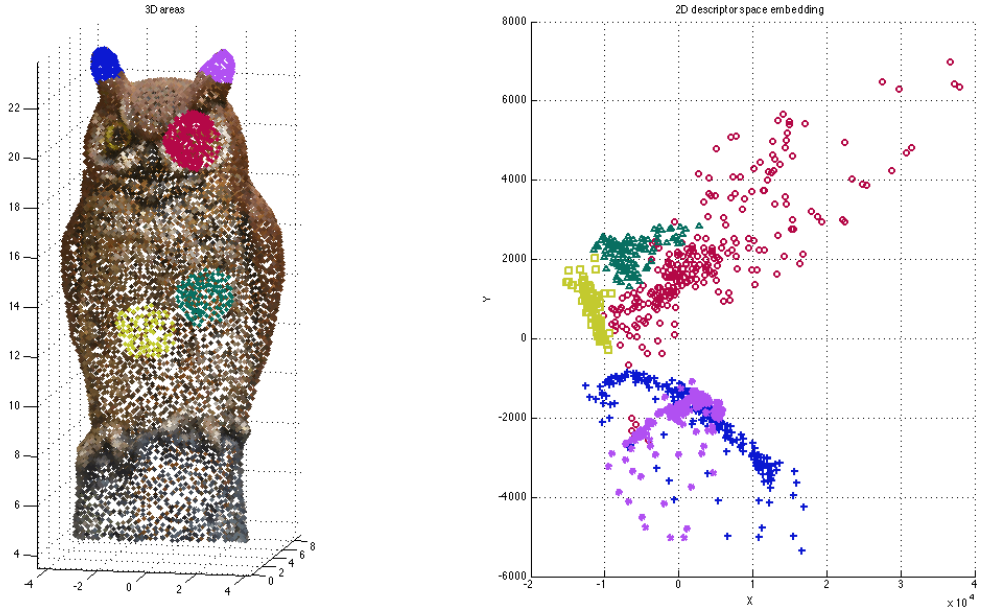


Figure 3.4: Set of scene areas where descriptors are obtained and their embedding to a two-dimensional space. Scene areas include two ears (marked in blue and purple) which are similar and therefore overlapped on the descriptor space plot. As these areas suffer changes in shape, its clusters are visually disperse. The red marked points, belonging to an eye area with changes in both colour and shape, appear separated from other clusters and also disperse due to this intra-area variations. Finally, yellow and green points belong to different homogeneous body areas, therefore they appear close in the 2D descriptor space (with a slight location variation due to slight differences in texture tone), and with a certain cluster compactness (due to their similar shape).

The most important implication of this manifold descriptor space is that it provides a formal way of comparing descriptors, while other approaches are forced to use distance approximations as histogram differences or correlations. While there

exist different approaches in the literature based in local Euclidean approximations ([Cherian et al., 2011; Arsigny et al., 2006]) with an efficiency compromise in mind, the presented method opts for the use of the geodesic distance proposed by Förstner in [Förstner and Moonen, 1999], already commented in chapter 2. This is adequate to this application as no prior knowledge might be available between two arbitrary descriptor points in a context of scene description and matching, therefore a local Euclidean approximation might not be accurate in most of the cases. Furthermore, with such a low dimensionality in the descriptor definition, the computational expense regarding the accuracy gain of a manifold aware distance is permissible.

Therefore, in order to measure the similarity of two arbitrary descriptors, the metric for computing distances between two covariance matrices C_r^1 and C_r^2, is defined as follows:

$$\delta(C_r^1, C_r^2) = \sqrt{\sum_{i=1}^{6} \ln^2 \lambda_i(C_r^1, C_r^2)} \tag{3.4}$$

where $\lambda_i\left(C_r^1, C_r^2\right)$ is the set of generalized eigenvalues of C_r^1 and C_r^2, whose magnitude express the geodesic distance between the compared points, preserving its curvature along the manifold.

3.3.3 Multi-scale Covariance Descriptor, MCOV

As computing covariance descriptors does not involve any major operation, it is easy to extend them to a multi-scale framework by just adding several radius magnitudes for the neighbourhoods around the descriptor center point. Therefore, each point in the scene will receive not one, but a set of descriptors:

$$C_M(p) = \left\{C_r\left(\Phi(p, r)\right), \forall r \in \{r_1..r_s\}\right\} \tag{3.5}$$

The idea behind using several neighbourhood radii is that discrimination performance can be improved if a point is supported by more than one descriptor, regarding a narrow to coarse set of surrounding areas, as depicted in the most left sub-image in figure 3.3. Then, we are intentionally seeking matches of points which are locally similar, but also related in a more global area. This can help to avoid repeatability problems and improve detection of points in edges or borders of scene objects. The radius estimation procedure, which is the only parameter the proposed method needs, is self-contained in this approach and commented below in section 3.3.4.

Finally, in the multi-scale descriptor framework, it is easy to extend the metric defined in equation 3.4 in the following way:

$$\delta_M(C_M^1, C_M^2) = \sum_{i=r_1..r_s} \delta(C_i^1, C_i^2) - \max_{j=r_1..r_s} \left[\delta(C_j^1, C_j^2)\right] \tag{3.6}$$

where C_i^1 and C_i^2 are the covariance descriptors belonging to each one of the $i = r_1..r_s$ radius scales, at each one of both scenes respectively. The formulation behind equation 3.6 takes into account the similarities of all scales except the one providing a lower similarity j, which is ignored because it might contain a major dissimilarity at a given scale -due to a possible dissimilarity on a border, an occlusion, or other artifacts.

3.3.4 Covariance Descriptor Properties for Scene Pre-analysis

Covariance matrices as descriptors have still other desirable outcomes thanks to their mathematical underlying fundamentals which allow several scene analysis stages. One of them is that they can be also used as keypoint detectors in a direct way. As defined after equation 3.3, a covariance matrix C_r contains the variance of the observed features on its diagonal, and the covariance on the other entries. Computing the determinant of a covariance matrix is equivalent to obtaining the so-called "generalized variance", which can be interpreted as a measure of the degree of homogeneity of each point in the scene [Wilks, 1932]. As the used features have been previously normalized, there is no range variation which could interfere on this analysis. Starting with an arbitrarily big radius parameter at a single scale (empirically we determined this as the magnitude corresponding to the 5% of the scene coordinates volume range) the covariance descriptor matrices for all the points in the scene can be computed, and all their determinants can be observed. Then, the ones with higher values can be interpreted as the points which belong to real interest areas, with inner significant variation in visual texture and 3D shape changes. It is worth to notice that these interest points are selected implicitly from a global point of view, combining both visual and shape saliency. Therefore, even in the case of an homogeneously coloured object like the one in figure 3.5, keypoints are still obtained on significant parts such as eye holes or borders.

Due to the nature of the proposed descriptor radial neighbourhood, relevant points might tend to form small clusters as samples could be shared for closer points, therefore producing similar descriptors. This can be reduced with relevance sampling procedures like the one proposed in [Torsello et al., 2011]. In this approach, this is naturally related to the aforementioned concept of generalized variance and also exploited as the associated relevance of a point in the scene. We want to explore all possible saliency clusters and isolate those points with major

relevance in a similar formulation as the one introduced in [Torsello et al., 2011]: for each one of the previously obtained keypoints kp at each scene S, we will compute its relevance region R_{kp} as:

$$R_{kp} = \{q \in S \,|\, \hat{\sigma}_{kp} - \hat{\sigma}_q > T, \forall q \text{ s.t. } |kp - q| \leq r\} \tag{3.7}$$

where $\hat{\sigma}$ is the generalized variance of each point, r is a radius parameter and T is a threshold parameter which we empirically set to 0.7 times the maximum generalized variance found in the points of the scene. Finally, a measure of distinctiveness can be assigned to each one of the relevance regions R_{kp}:

$$f(p) = \|R_{kp}\|^{-k} \tag{3.8}$$

where $\|R_{kp}\|$ is the 2-norm of the points in R_{kp} and k is an equalization parameter in order to change the relative weight of really distinctive points (the larger its value, the more distinctiveness of points in a small patch is emphasized). It has been empirically set k to 1. We can finally keep the points with maximal values according to the distinctiveness features and observe how these belong to local isolated points within the original saliency clusters as depicted in the right image in figure 3.5.

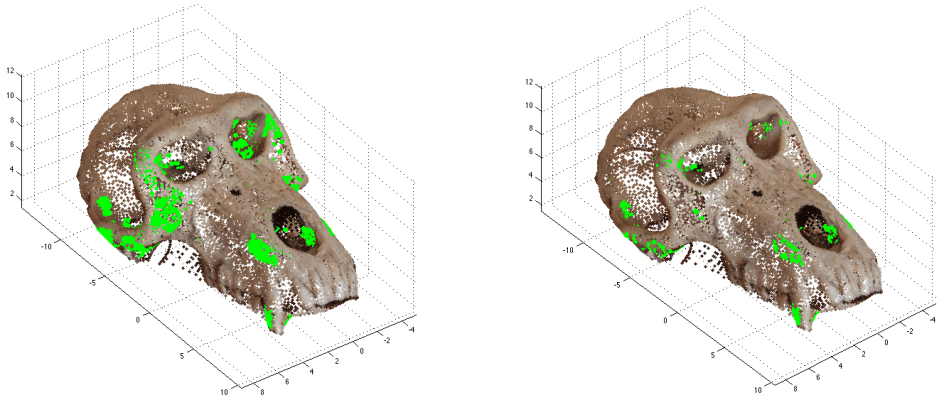


Figure 3.5: Visual example of keypoint analysis by generalized variance. The left sub-figure shows the 1500 most significant points of the scene, marked by sorting their covariance descriptor determinants (generalized variances) in descendant order. Even if the color information of the object is homogeneous, interest points have been detected on salient areas of the scene. The computational cost of such task is minimal. The right image shows the set of points after the relevance sampling procedure, which in this case isolates 488 salient points with a major degree of sparsity regarding the previous saliency clusters. This can help reducing further registration errors.

This saliency analysis can also be preceded by a point suppression stage thanks to the analysis of the rank of the covariance matrix descriptors. If different feature observations within the neighbourhood of a given point are correlated, which is not desirable, the rank of the descriptor matrices will be lower than the number of used feature dimensions. This straight criterion allows discarding uninteresting points where the covariance descriptor does not capture any significant differentiation between features.

Finally, an estimation procedure of a more narrow radius value can be integrated, taking into account the nature of the scene in order to fit its probabilistic definition of points with a more accurate area sensitivity. From statistics theory it is known that the sample mean is a good estimator of the population of a random variable distribution, and its sampling size parameter in order to lay within a confidence interval is modelled by Chebyshev's inequality with the following expression:

$$P\left(\mid \bar{X} - \mu \mid \geq \epsilon\right) \leq {}^{\sigma^2}/_{\epsilon^2 n} \tag{3.9}$$

where μ and σ^2 are the mean and variance of the distribution that is being considered; $\bar{X}$ is the sample mean according to the number of samples n that are being observed; and ϵ is the threshold on data representation.

As an example, if we want to infer the number of samples such that data will lay within 0.1 units the original distribution, with a confidence of the 95%, this can be expressed as $P\left(\mid \bar{X} - \mu \mid < 0.1\right) \geq 0.95$.

This is equivalent to $P\left(\mid \bar{X} - \mu \mid \geq 0.1\right) \leq 1 - 0.95$, therefore it can be related to Chebyshev's inequality and generalize the following expression for an arbitrary feature distribution:

$$n \geq \frac{\sigma^2}{\epsilon^2 \left(1 - p\right)} \tag{3.10}$$

where p is the desired confidence value. Usually a threshold value $\epsilon = 0.1$ and a confidence interval of $p = 0.95$ will be used. This will provide an upper bound on the minimum needed number of samples n necessary to limit the sampling mean confidence error to a given value. Relating this to the presented framework, while calculating the covariance descriptors for the proposed scene pre-analysis we are already observing each one of the six used feature distributions at each scene point neighbourhood, and this is being kept encoded at the diagonal values of the set of descriptors. Therefore, the boundary equation defined in 3.10 can be applied to each one of the feature variances, defining a set of 6 candidate sampling sizes. As this provides several lower boundaries, the maximum value of the candidate sizes will be kept. While this is a scene-dependant, quite flexible methodology, it allows for an adaptive method in the case there are areas with specific high variation. Its

formulation is coherent along the points of the whole scene and provides specific descriptor constructions, rather than using a static radius parameter for the whole scene. Usual analyses for scenes with average homogeneity of shape and color (as most of the ones depicted in figure 3.8, whose point clouds have densities ranging from 20000 to 30000 points) reflect the need of taking around 400-500 samples within the radial neighbourhood. This sampling size can be translated to a radius magnitude according to the density of the scene point cloud. For the multiscale approach, the usage of 5 different scales is proposed, with radius scales $s = \{1, 1.1, 1.3, 1.6, 2\}$ times the single-scale descriptor radius. This scaling distribution focuses the attention on narrow neighbourhoods, while coarse areas are still present for disambiguation.

All these inherent benefits reinforce the idea that the proposed descriptor methodology is not only suitable for the core task of 3D scene point definition, but also integrates a set of possibilities on the statistical analysis of data, as gathered up on this section, which provide an added value to the framework.

3.4 Globally Aware Scene Registration by an Evolutionary Game Theory Approach

The descriptor introduced so far has proven to be discriminative enough for a local recognition of a point in different views of a scene. The associated salient point detector, which pre-selects a set of relevant points according to what has been explained on previous section, is also helpful on this high descriptiveness level. Nevertheless, if the scene contains unavoidably similar points due to facts as repetitive patterns of an object or symmetries, this could pose a bigger challenge for the descriptor which could only be addressed with the help of scene-wise knowledge taking into account these particularities. This focuses the interest on the proposal of a descriptor matching methodology which does not only takes into account relationships between local point similarities, but also encodes a set of global restrictions in order to avoid possible ambiguities in the whole scene. The proposed framework will perform a rejection of all the points which do not fit into this set of scene-wise constraints, leaving only a selection of correctly considered point matches. A global heuristic for match evaluation based in the so called geometric consistency will be proposed.

As previously stated in section 3.2, state-of-the-art approaches in scene registration are commonly based on top of iterative algorithms such as RANSAC or its later variants [Fischler and Bolles, 1981; Chum et al., 2004; Chum and Matas, 2005, 2008]. However, there are several aspects in these algorithms which do not suit our purposes. The most important one is the presence of a possibly high

number of outlier matches which could have an impact in the performance of the method due to the number of needed iterations, or even worst, the achievement of a convergent solution which is not a correct registration. In the context of registering complex scenes with a huge number of descriptor matches to be discarded, the impact of high amounts of outlier candidates might become computationally intractable as well. Other drawback considerations include the need of input parameters which must be tuned in order to obtain a valid solution, and the need of a high number of random evaluations of possible combinations producing, in the specific case of registration, an unoptimized performance of the method and probably different solutions for two different executions limited in time.

This section introduces a registration method for the context of matching 3D scene views which entails a significant conceptual innovation regarding the aforementioned methods. In a first place, the change of paradigm implies not to compute implicitly the spatial transformation between scenes at each iteration (and to temporarily evaluate it), but to perform a rejection of all those matches which do not satisfy a set of constraints. The spatial transformation will be computed in a final instance, as the result of a limited set of leftover candidate points. In a second place, the presented method will allow to enclose a formulation expressing the adequacy of each match within the final solution: it is proposed to combine the geometric consistency together with the point descriptor likelihood (which, as a reminder, is well defined by the metric in equation 3.6). Therefore the set of constraints will be joining both local and global information. And, last but not least, the proposed method will be independent of the presence of outlier candidate matches, and theoretically guaranteed to converge asymptotically towards the solution: at each iteration, it will discard one candidate match. Therefore, the maximum number of trials will depend on the number of match samples, which is a clear advantage in a high presence of noise or outliers regarding an approach like RANSAC (this will be commented hereafter in an experimental set-up in section 3.5).

This method proposes to translate the global scene matching problem to the game theory field using the so called "evolutionary stable strategy" (ESS) solver introduced in [Albarelli et al., 2009]. This approach presents a framework where a set of abstract candidates of a system are successively discarded in order to obtain the best remaining combination of them according to a defined heuristic function. While the ESS method defined in [Albarelli et al., 2009] is a standard methodology for game-theoretic problem solving, Albarelli *et al.* have used this approach in a more scene registration focused application context in [Albarelli et al., 2010] and [Rodolà et al., 2013]. In these cases, they use the game theoretic solver for refining matches that have already been filtered by a standard descriptor pairing algorithm –MeshHOG descriptor and associated MeshDOG keypoint locator [Zaharescu et al., 2009]. Therefore, their heuristic functions for game definition are

limited to simple spatial constraints for a final discard of those matches. The underlying idea is to consider all the pairwise matches of two compared scenes, and calculate a penalty value associated to the cost of hypothetically choosing each one of these correspondences as part of the final registration solution. In the game theoretic framework these values are named payoffs, and can be computed at once for a set of correspondences resulting from a descriptor matching stage. In an analogy to a game, these payoffs will be the set of "rules", each scene view will be a player and each match candidate a game turn. Therefore, the best play of the game (the best registration between scene views) will take place by the best set of turns for both players -that is, the best set of matches at each scene incurring on the best global cost. The aforementioned set of payoffs is codified in a matrix notation where all the possible pair choices are being taken into account.

The following subsection presents the definition of a payoff term which is able to integrate both global scene geometric structure constraints and descriptor similarities. A game theoretic based solver is powerful enough for taking into consideration similarity constraints of point descriptors, and scene-wise geometrical structure restrictions for relevant challenges as symmetries or repeated areas, in a single game definition. This is different to other current approaches which depend on a previous descriptor extraction and correspondence stage and use the game theory framework as a gathering of rules in order to reject match candidates which do not fit some local surface characteristics.

3.4.1 Modelling the Game

The main complexity of the game theoretic matching framework lies on the payoff matrix building step, which must take into account all the possible pairwise affections between candidate matches of both scenes. It is emphasized that each match payoff must encode all the information that must be assigned to a pair of points, both in a positive or a negative way: in this sense, the costs related to local likelihood, geometric consistence, and relative distance of points will be considered together. The latter term helps in a better discard of undesired matches: the ideal keeping of point candidates is that set which is sparse enough with local similarities of point descriptors. This can be seen as finding the most spaced-out sub-graph common to both scenes graphs of coordinates, where the vertices are similar enough (see figure 3.6).

A game theory based solution is proposed as a holistic way of grouping all the information available both in terms of descriptor similarities and scene-whise geometric prior knowledge. With these conditions, the building of the payoff matrix is defined as follows. Let A and B be the scenes that have to be registered. Let $\{a_m\}$ be the set of points in A and $\{b_n\}$ the set of points in B. Then there exists a set of k candidate pairs $\{(a_i, b_j)\}$ which have been preselected according to the

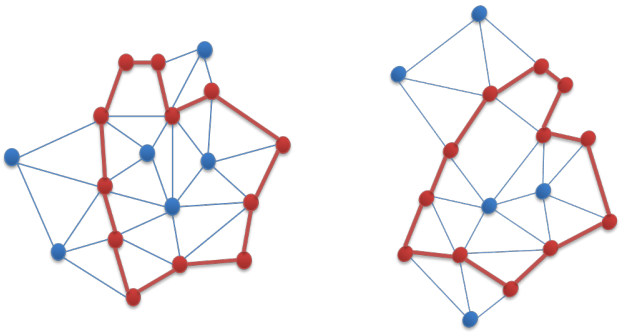


Figure 3.6: Schema of how a common sub-graph must be selected by the game theory solution. On the left: the graph obtained by the cloud points from the first scene; on the right; the graph obtained by the cloud points of the second scene. Marked in red there is the most suitable common sub-graph found within both graphs.

best covariance descriptor likelihoods between scenes. For each pair, its game payoff can be evaluated regarding any other pair of matches, exhaustively. Therefore, a matrix C of game payoffs, of size $k \times k$, is defined for all combinations of pairs $\{(a_i, b_j), (a_k, b_l)\}$, and it will take into account all the relationships in the scenes with the corresponding incidence over the global registration error:

$$c_{(a_i,b_j)(a_k,b_l)} = P_{\text{desc}} \cdot P_{\text{geom}} \tag{3.11}$$

where P_{desc} is the payoff related to the covariance descriptor similarity and P_{geom} is the payoff related to the geometric consistency. In more detail, both values are defined as follows:

$$P_{\text{desc}} = f(a_i, b_j) \cdot f(a_k, b_l) \tag{3.12}$$

$$f(a, b) = e^{-\delta_{\text{M}}(C_M(a), C_M(b))} \tag{3.13}$$

where δ_{M} is the multi-scale Förstner distance between the covariance descriptors of a and b. This term is taking into consideration a normalized payoff value associated to the local similarity of points.

The geometric consistency constraint is defined as:

$$P_{\text{geom}} = \frac{\min(d(a_i, a_k), d(b_j, b_l))}{\max(d(a_i, a_k), d(b_j, b_l))} \cdot g(a_i, a_k, b_j, b_l) \tag{3.14}$$

$$g(a_1, a_2, b_1, b_2) = e^{-|\text{d}(\text{a}_1,\text{a}_2) - \text{d}(\text{b}_1,\text{b}_2)|} \tag{3.15}$$

where $d(x, y)$ is the Euclidean distance between 3D points x and y. As we are working with 3D information, we are able to ensure that the Euclidean distance between points of the object is the same from every point of view.

The first *min/max* term in this geometric payoff was originally used in [Albarelli et al., 2010] and [Rodolà et al., 2013], and penalizes elements which are closer in the scene as they would incur in more error if they where selected as a wrong part of the registration solution. Note that these approaches are based on a previous keypoint detection and matching stage, therefore a single spatial constraint such as this one is enough for the rejection of erroneous pair candidates, provided the keypoints are correctly matched on controlled scenes. As this approach is focused onto the registration of complex, textured scenes with still many point candidates at this point, a second term in this payoff value is added, as defined in equation 3.15. This adds a normalized coefficient which indicates the structure similarity between points in both scenes. All these constraints, together with the aforementioned P_{desc} term in equation 3.13, define a single game which is capable of selecting registered point pairs taking into account both texture and shape information. See figure 3.7 for a clarification of the elements involved in such calculations.

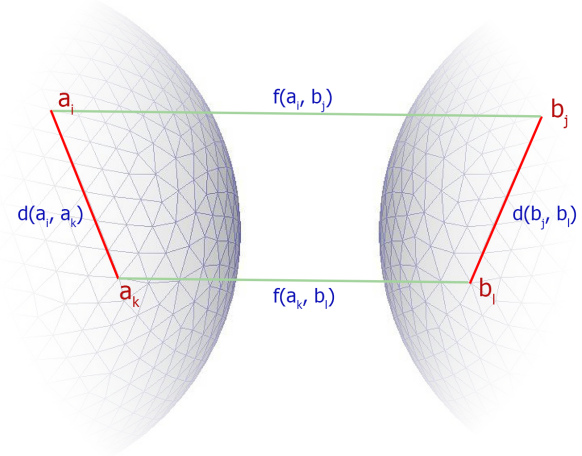


Figure 3.7: Scheme of the elements involved in payoff calculations. $f(a, b)$ expresses the descriptor likelihood between a pair of matches. $d(a_1, a_2)$ evaluates the geometric consistency on the match candidates within the pair of matches which is being evaluated.

A visual representation of a general payoff matrix remains as follows:

$$C = \begin{array}{c} \begin{array}{ccccc} \dots & a_i & \dots & a_k & \dots \\ \dots & b_j & \dots & b_l & \dots \end{array} \\ \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots \\ \dots & 0 & \cdots & c_{(a_k b_l)(a_i b_j)} & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \dots & c_{(a_i b_j)(a_k b_l)} & \cdots & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array} \begin{array}{cc} \vdots & \vdots \\ a_i & b_j \\ \vdots & \vdots \\ a_k & b_l \\ \vdots & \vdots \end{array} \tag{3.16}$$

3.4.2 Playing the Game

Finally, it is necessary to find a stable solution to the game represented by the payoffs modelled in C, which are in fact the implicit restrictions of the candidate matches of the scene registration. According to [Albarelli et al., 2009], the evolutionary stable solution of the game is the so-called support vector x whose response to the game is maxima: $x^T C x \geq y^T C y \ \ \forall y \in \Delta$, where $\Delta = \left\{ x \in \mathbb{R}^k : \sum_{i=1}^{k} x_i = 1 \text{ and } x_i \geq 0 \right\}$, the space of all vectors which are solutions to the game.

The support vector x can be found via the Evolutionary Stable Strategy solver algorithm proposed in [Albarelli et al., 2009]. If $x_i > 0$, then the match belonging to column or row i in C is marked as a positive correspondence. The own values x_i can be considered as normalized weights expressing the confidence associated to each correspondence. In [Albarelli et al., 2009] some other valuable details are examined: in a first instance, it is shown that if a mixed solution is wanted (that is, more than one element in x satisfies $x_i > 0$), then it is necessary that $c_{ii} = 0$ and $c_{ij} \geq 0 \ \ \forall i \neq j$. In a second place, the algorithm is proven to converge to a unique and global solution which takes into account all the payoff values associated to all the possible matches between scenes. And finally, this convergence is guaranteed in an asymptotic way and with a linear time complexity per iteration.

Therefore the solution found in the support vector indicates the indices of the subsets of correspondences in both scenes which can be used to find the spatial transformation needed in order to change the coordinates of one scene into the other, by solving the problem of absolute orientation [Horn, 1987] for registration. All the involved steps of this approach are summarised in algorithm 3.1.

Algorithm 3.1: Overview of the proposed registration procedure

input : Two scene views in the form of 3xN point clouds
output: Rigid transformation (R, T) between scene views

Stage 1: *scene pre-analyisis*
- Compute pre-descriptors at view 1 and view 2.
- Perform generalized variance analysis.
- Prune salient areas by relevance sampling.
- Estimate *covRadius* by confidence intervals.

Out: *keypoints kpSc1, kpSc2, covRadius*

Stage 2: *descriptors obtention*
for $kpSc1 \leftarrow 1$ **to** $\#kpSc1$ **do**
    - compute each MCOV(kpSc1)
end
for $kpSc2 \leftarrow 1$ **to** $\#kpSc2$ **do**
    - compute each MCOV(kpSc2)
end
- compute distance matrix between descriptors

Out: *distMatrix size* $\#kpSc1 \times \#kpSc2$

Stage 3: *get candidate correspondences*
for $r \leftarrow 1$ **to** *#rows distMatrix* **do**
    - get best correspondence according to MCOV distances (*inclusive/exclusive ratio criterion*)
end

Out: *set of* n *candidate pair matches*

Stage 4: *registration via Evolutionary Game Theory*
- build payoff matrix C
- apply Evolutionary Stable Strategy Solver [Albarelli et al., 2009]

Out: *set of* m *final matches*

- Compute $[R, T]$, rigid transformation from set of matches (absolute orientation)

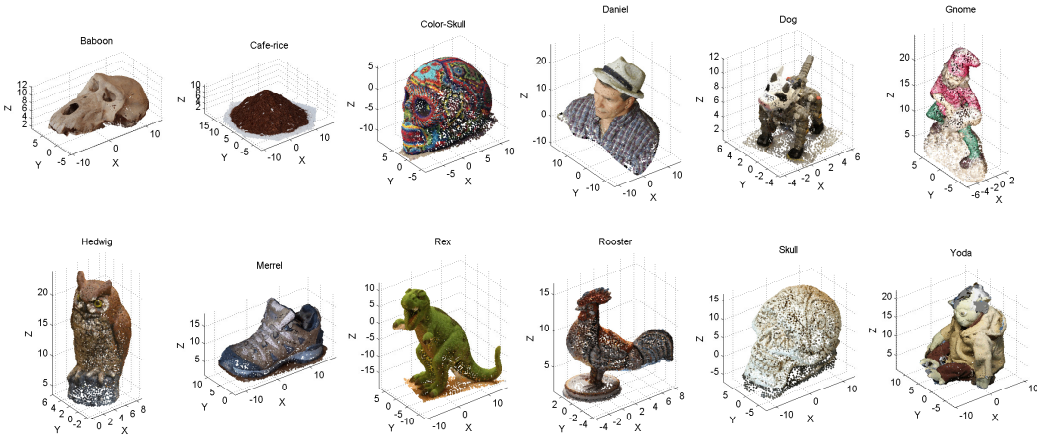


Figure 3.8: 3D plot of the 12 models included on our database. Full scenes are shown without added noise.

3.5 Experimental Results

The proposed descriptor approach is being validated on top of a dataset combining 3D shape with visual information. This dataset contains 12 scenes which have been obtained using Autodesk 123D Catch [1] 3D modelling software. The dataset combines scenes of originally acquired objects and others available on the 123D Catch website under a Creative Commons license. These models are stored as 3D meshes with photometric texture, where each vertex has a unique identifier for experimental ground-truth purposes. See figure 3.8 for a visual representation of the 12 base models used. This dataset has been made publicly available at *http://cmtech.upf.edu/3DVisDatabase*. The contained objects have been particularly selected in order to include challenging handicaps for testing the performance of the presented method: repeated areas, homogeneous surfaces and textures, and symmetries.

3.5.1 Descriptor Comparison

In order to test the descriptor performance, the MCOV covariance descriptor approach will be compared against the state-of-the-art methods MeshHOG [Zaharescu et al., 2012] and CSHOT [Tombari et al., 2011]. In addition, the performance of the Textured Spin Images approach [Brusco et al., 2005] will also be evaluated. Even if this method was presented a decade ago, it is a variation of the original Spin Images approach [Johnson and Hebert, 1999] which is still considered one of the classical 3D descriptors in the literature for successful matching

[1] http://www.123dapp.com/catch

of dense scenes. We want to include its results in our comparison as a base line of a method which set up a standard in 3D scene matching. The compared descriptor approaches are used following the original implementation by their authors, and any needed parameter (radius, bin size) is set according to the recommendations of their original proposals -or to equivalent values regarding our approach in order to provide the most fair comparison as possible.

3.5.2 Performance Over Noise Variations

This experimental evaluation is testing the descriptor tolerance to noise variations. Each model in the database is affected by a variation including *i)* an arbitrary rotation, *ii)* an arbitrary translation, and *iii)* an addition of noise to color and surface coordinates. Noise levels will follow different Gaussian distributions with standard deviations according to 2, 4, 6, 8 or 10% of each one of the data channels. Therefore, for each model the following cross validation procedure including 10 folds of 100 randomly selected points along the surface of the scene has been performed. For each one of the evaluated points, its descriptor likelihood against the same set of points on the variation of the model has been computed. The evaluation method consists on observing the amount of false and true positives, and false and true negatives averaged along the cross validation test, in terms of matching scene points by their according descriptor likelihood measures. For the presented MCOV descriptor, we will use the metric defined in equation 3.6. According to a *ratio* parameter, two criteria for evaluation are presented:

• The so-called *exclusive ratio*, considers a match as a true positive if and only if the descriptor likelihood between the match points is *ratio* times better than the second best match candidate likelihood. This criterion variant is inspired in current approaches as SIFT [Lowe, 2004] and has the particularity of being more restrictive on finding true positive matches, reducing also the apparition of false positives. Due to its behaviour, this selection is suitable for the evaluation of the descriptor performance itself.

• The so-called *inclusive ratio*, considers as true positives all those matches which are within the boundaries of *ratio* times the best likelihood of this set of candidates. In this case the rate of true positive candidates is increased, but this has the expense of increasing the risk of appearance of false positives. This criterion is suitable for a whole registration procedure as a point is associated to many matches as long as they are similar within a range of likelihood measurements, at the expense of requiring a rejection method afterwards in order to deal with elements external to the descriptor itself, as pattern repetitions in the scene or symmetries.

Both matching criteria are presented and evaluated in the experiments as they might be of different adequacy regarding the application context of the descriptor: as stated before, the main difference between both methods is the amount of tolerated false positives they allow. Assuming a descriptor is reliable at representing a given area, the presence of false positives is not a drawback by itself; it is just a side effect due to the possibility of repetitions of visual patterns or surfaces in the scene. Therefore, the *inclusive ratio* criterion is more flexible and is allowing this fact to happen. In some applications, such as object detection or scene registration, this can be a desired feature, but it puts into consideration the needing of some sort of global mechanism which must be capable of finding repetitions, symmetries, etc. and filter out the non-positive matches according to global error minimizing constraints such as geometric consistence, which is why the MCOV descriptor proposal is paired with a scene-wise game theoretic solver definition. Nevertheless, both criteria are complementary, with a common point when both exclusive or inclusive ratio parameter is set to 1. In this case, both criteria are conceptually the same one.

The results of the experiment are presented as follows: for each level of noise the *ratio* coefficient is moved within a range of 1 to 5 and a set of ROC curves is obtained as exemplified in figures 3.9 and 3.10 for *exclusive* and *inclusive* ratio methods respectively. This is useful for comparing the behaviour of the different tested descriptors under all noise variations, for each one of the twelve available models. As it can be observed in the separate figures, due to aforementioned complementarity the ROC curve plots belonging to inclusive criterion are the continuation of the exclusive criterion ones (please note the later ones are zoomed in in order to offer a better visualization). It is agreed that in a more formal sense, the plots should be presented continuously, and with a more extense *ratio* parameter space exploration in order to offer normalized coordinate axis between 0 and 1 values. However, the current figures intend to offer more detail in order to interpret the results: using the current disposition, the figures clearly display that when the *ratio* parameter is set to a defined higher value of 5, some of the descriptors have a false positive rate of 1, while other ones still maintain this value in a lower level.

For a numerical comparison between these curves, their *Area Under the Curve (AUC)* measure can be obtained. This allows to numerically summarize the average performance of the four tested descriptors over all the models in the database, as seen in tables 3.1 and 3.2 for exclusive and inclusive ratio criteria, respectively.

We can see how the proposed MCOV descriptor is more stable regarding the increases on the noise levels. Since other methods are working with local surface neighbourhoods and 3D coordinate histogram representations, they will quickly suffer this distortion on data, i.e: at bin discretisation. On the contrary, the MCOV

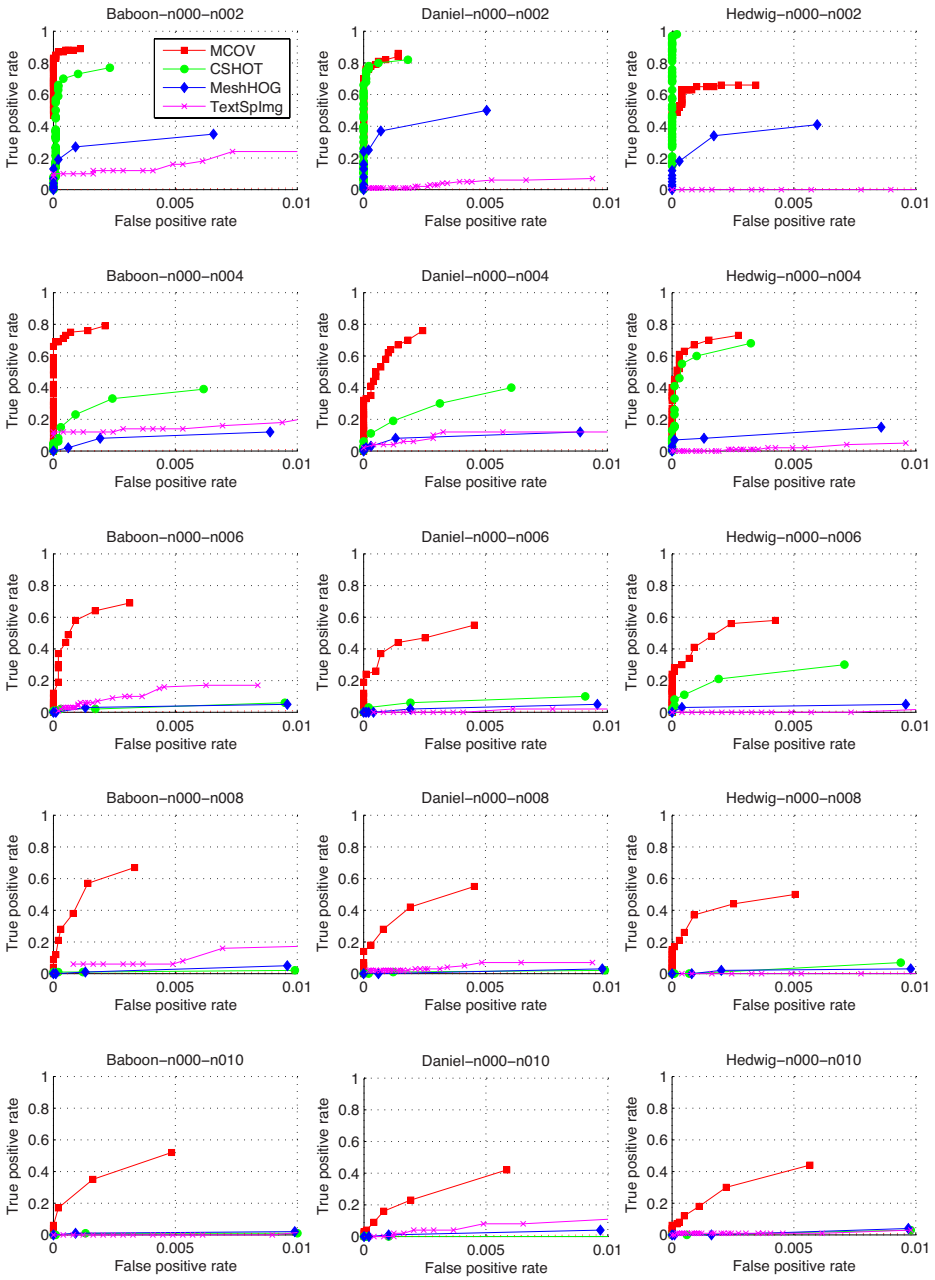


Figure 3.9: ROC curves for comparison of several 3D descriptors, using the *exclusive ratio* criterion. Each column depicts a test on a different model of the database. Each row shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10% of the standard deviation of color and surface coordinates).

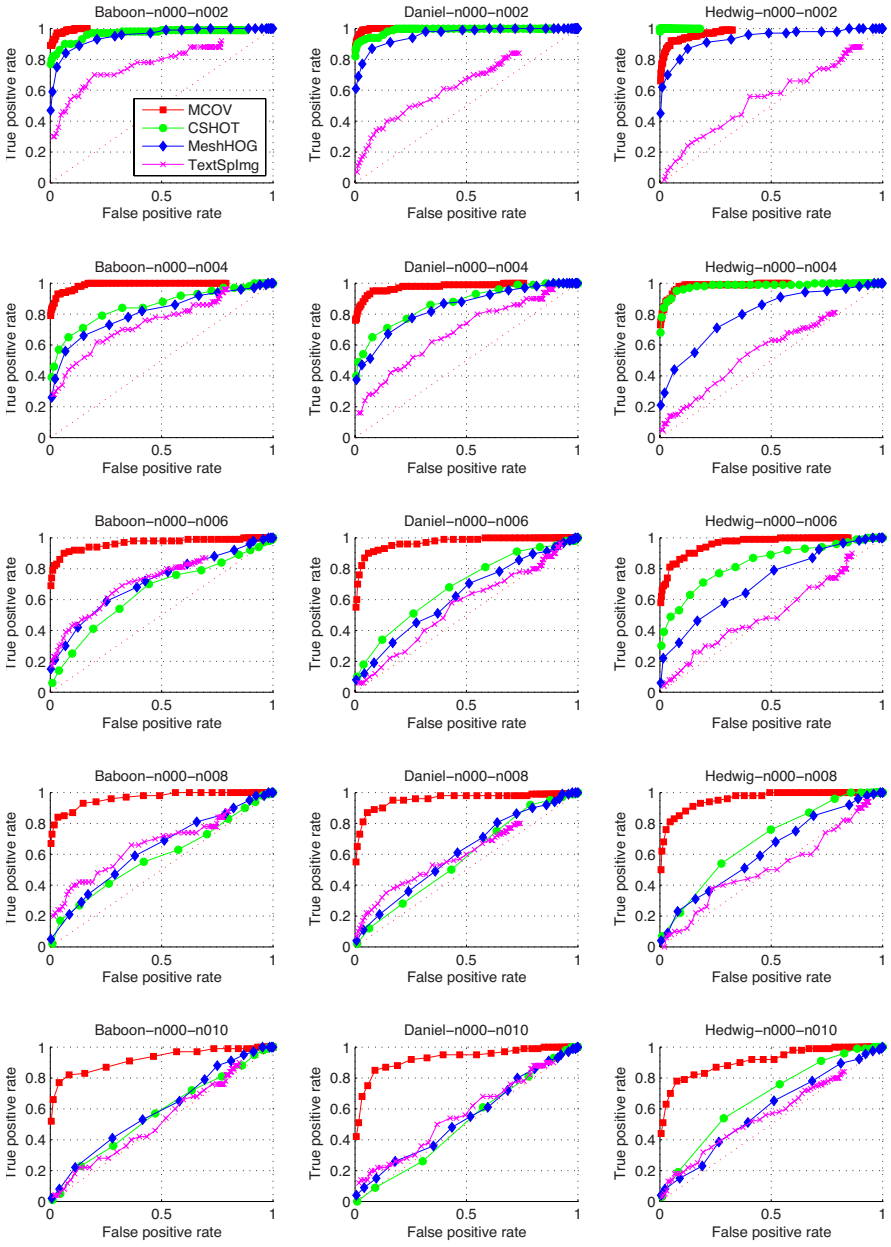


Figure 3.10: ROC curves for comparison of several 3D descriptors, using the *inclusive ratio* criterion. Each column depicts a test on a different model of the database. Each row shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10% of the standard deviation of color and surface coordinates).

	n002	n004	n006	n008	n010
MCOV	0.896	**0.868**	**0.781**	**0.758**	**0.710**
CSHOT	**0.911**	0.799	0.602	0.534	0.511
MeshHOG	0.745	0.703	0.613	0.528	0.506
Text. SpinImages	0.619	0.544	0.540	0.523	0.503

Table 3.1: Average AUC measures for 12 models, *exclusive ratio* evaluation, 100% vs 100% resolution, for 5 levels of noise. Bold values indicate the best performance in each case.

	n002	n004	n006	n008	n010
MCOV	0.991	**0.976**	**0.961**	**0.953**	**0.917**
CSHOT	**0.992**	0.913	0.758	0.616	0.562
MeshHOG	0.963	0.819	0.704	0.607	0.577
Text. SpinImages	0.750	0.614	0.615	0.564	0.533

Table 3.2: Average AUC measures for 12 models, *inclusive ratio* evaluation, 100% vs 100% resolution, for 5 levels of noise. Bold values indicate the best performance in each case.

descriptor offers a more flexible representation since it considers 3D points as samples of a distribution and, by construction, subtracts the mean of this samples distribution: therefore, in case of noise, it will be naturally attenuated.

Thanks to this experimental set-up, a conclusion about fusion of color together with shape information can also be extracted, as the three database models selected to be represented column-wise in figures 3.9 and 3.10 provide three different challenging scenarios in terms of color homogeneity, repetitive patterns or great color variation, respectively. In this sense, one can see how classical histogram representations, as in the basis for MeshHOG or Textured Spin Images, are clearly affected by color variance. The usage of the *Hedwig* model is a clear challenge for the Textured Spin Images approach as the color sparsity is saturating the illuminant binning component of that descriptor. This was in fact identified as a possible drawback by their own authors in [Brusco et al., 2005]. Other more flexible approaches as CSHOT or the presented MCOV descriptor offer more robustness in its representation until bigger amounts of applied noise.

3.5.3 Performance Against Resolution Changes

A very similar experiment to the one presented on the previous section is also conducted by applying a high resolution variation over the models. The aim is to test

the performance of descriptors when matching original models against a down-sampled variation to a 50% of their point cloud density. This down-sampling procedure is applied by randomly suppressing samples over the point clouds.

Again, by moving the *ratio* coefficient within a range of 1 to 5, a set of ROC curves for the 12 tested models can be obtained under the same 5 noise variations as in the previous experiment. Tables 3.3 and 3.4 reflect the associated average AUC measures for exclusive and inclusive ratio criteria, respectively. As in the previous experiment, the ROC curves corresponding to the *Baboon*, *Daniel* and *Hedwig* models are presented for an easier visualization of descriptor performances. These are plotted in figures 3.11 and 3.12, again taking into account the proposed *exclusive* and *inclusive* ratio matching criteria.

	n002	n004	n006	n008	n010
MCOV	**0.874**	**0.813**	**0.732**	**0.657**	**0.599**
CSHOT	0.772	0.651	0.572	0.515	0.510
MeshHOG	0.561	0.547	0.523	0.521	0.511
Text. SpinImages	0.572	0.522	0.527	0.498	0.498

Table 3.3: Average AUC measures for 12 models, *exclusive ratio* evaluation, 50% vs 100% resolution, for 5 levels of noise. Bold values indicate the best performance in each case.

	n002	n004	n006	n008	n010
MCOV	**0.984**	**0.967**	**0.924**	**0.871**	**0.812**
CSHOT	0.906	0.823	0.668	0.614	0.597
MeshHOG	0.616	0.597	0.522	0.517	0.521
Text. SpinImages	0.662	0.613	0.563	0.534	0.520

Table 3.4: Average AUC measures for 12 models, *inclusive ratio* evaluation, 50% vs 100% resolution, for 5 levels of noise. Bold values indicate the best performance in each case.

As it can be seen, both numerical and ROC curve results suggest this is a more challenging experiment, as data is highly altered. Nevertheless, the statistical basis of the presented descriptor is valuable again in terms of resolution robustness: as long as a large enough number of samples is preserved, fact which we are assuring, covariance will still encode the underlying characteristics of feature distributions.

In the other evaluated descriptors the changes on data resolution will incur on a bigger descent of their performance. A special consideration must be taken into account in the MeshHOG method, which requires faces information in order to compute its descriptor. The applied resolution down-sampling implies the

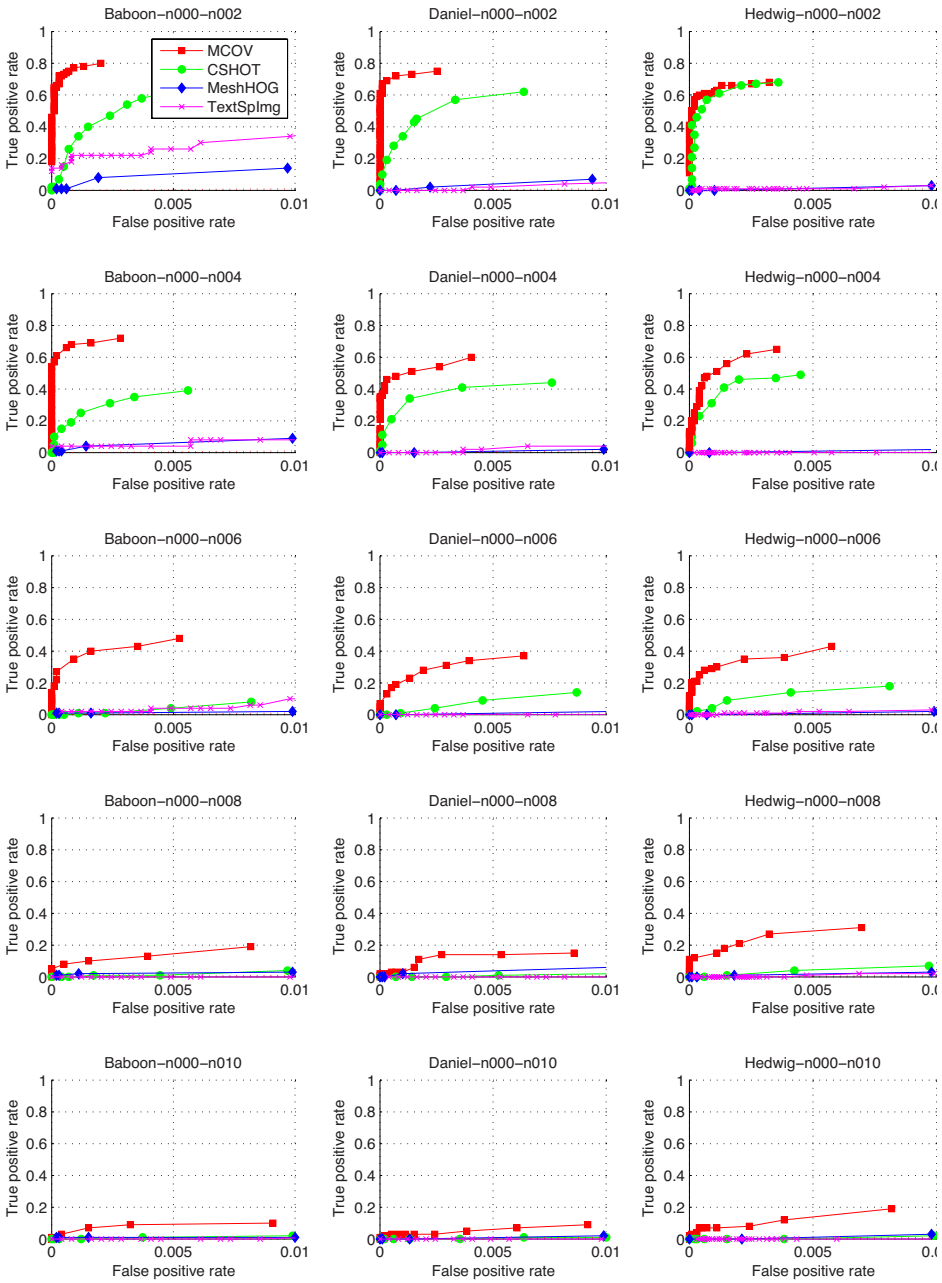


Figure 3.11: ROC curves for comparison of several 3D descriptors, using the *exclusive ratio* criterion and reducing the resolution of the second scene to the 50%. Each column depicts a test on a different model of the database. Each row shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10% of the standard deviation of color and surface coordinates).

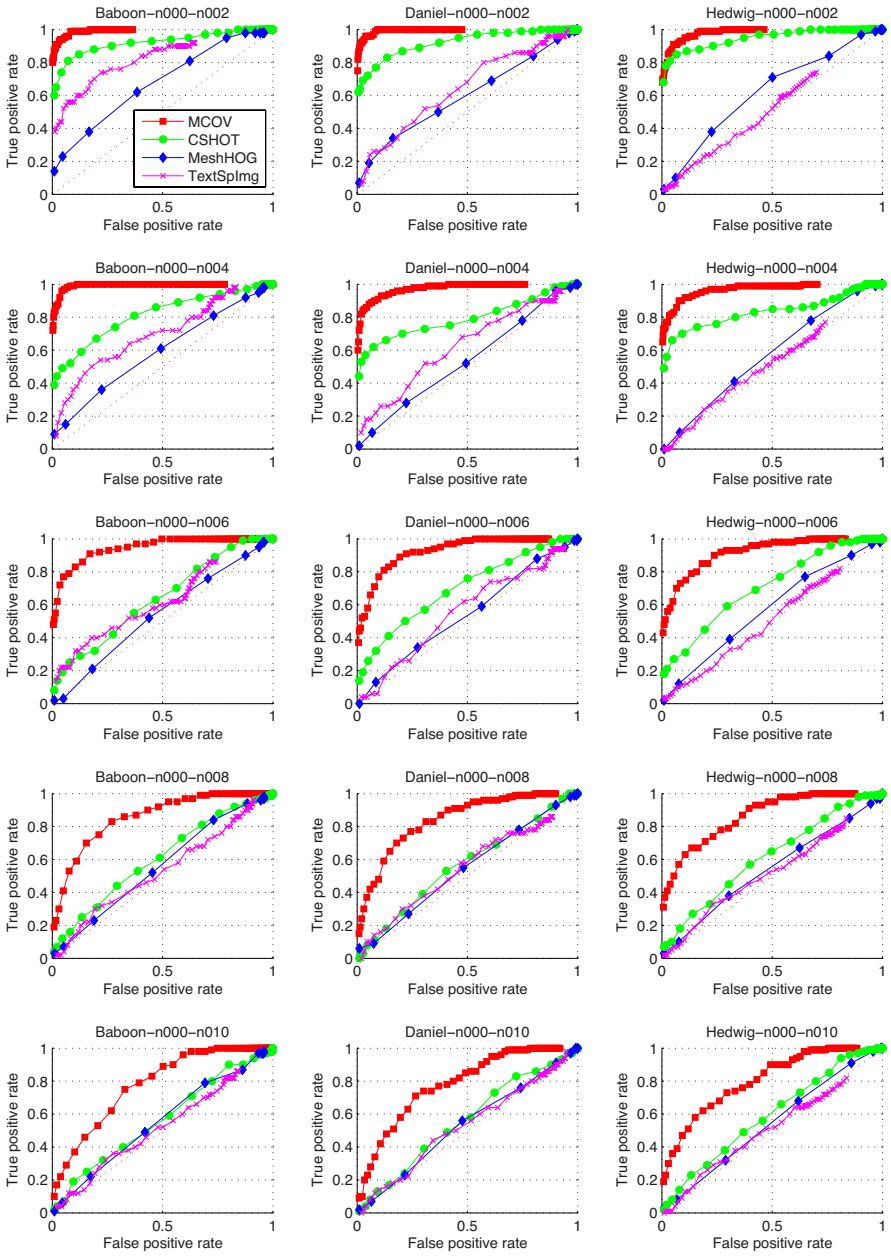


Figure 3.12: ROC curves for comparison of several 3D descriptors, using the *inclusive ratio* criterion and reducing the resolution of the second scene to the 50%. Each column depicts a test on a different model of the database. Each row shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10% of the standard deviation of color and surface coordinates).

computation of an equivalent triangulation by using the edge collapse procedure [Luebke, 2001]. This has a drastic impact on its performance as ROC curves and AUC values suggest.

3.5.4 Exclusive/Inclusive Ratio Matching Evaluation

Figure 3.13 presents a complementary qualitative result for visually observing the different impact of the aforementioned matching criteria on the descriptive performance of this approach. The assignation of different *ratio* values, as well as the performed criterion, affects on the number of established matches. The equivalent case between two methods takes place when *ratio* parameter is set to 1.

While the *exclusive ratio* criterion is a usual procedure found in other approaches, its application is of limited feasibility in the context of registration of arbitrarily repetitive scenes. As several challenging conditions must occur, it is better to intentionally allow a certain flexibility on point matches, in order to keep all the locally similar areas of the scene. Later on, the parts with local similarities will be filtered by the game theory geometric consistence methodology. The conclusion that can be extracted from this experimental set-up is that the most feasible criterion in order to perform this match candidate selection is the *inclusive ratio* criterion.

3.5.5 Game Theory Evolutionary Stable Strategy Solver Validation

The procedure of the game theoretic approach consists in the successive removal of error inducing correspondences, from the set of initial descriptor matches, until the algorithm converges to a limited set of point correspondences. These must be consistent in terms of local descriptor likelihood as well as scene-wise geometric consistency. Since the main reason for adopting this methodology in order to discard the incorrect scene correspondences is its faster convergence to a global solution, and its computational feasibility over scenes with huge amounts of outlier correspondences, it is valuable to provide an experimental set-up which validates the performance of the proposed methodology under different deliberate amounts of outlier descriptor correspondences. For this reason the following procedure is carried out for different models and levels of noise:

- A fixed number of 500 correspondences is established between two instances of a scene. This amount of candidate matches will be intentionally corrupted with increasing levels of outlier correspondences, from 5% to 95% regarding the total amount of candidate pairs. A corruption of a candidate match is considered as an alteration of the coordinates from the two scene views which should have been matched according to a local similarity

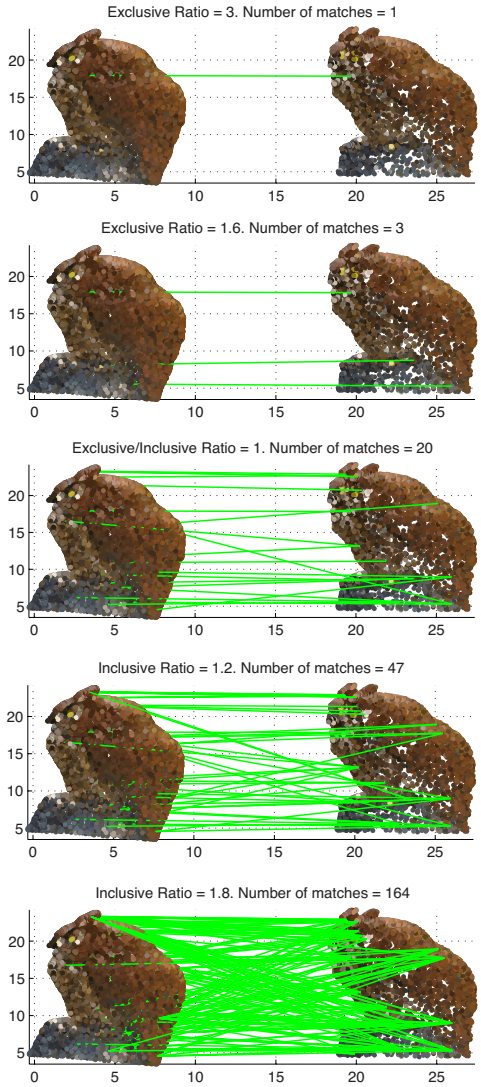


Figure 3.13: Effects of the different matching criteria over the matches of *Hedwig* model, which is considered specially challenging due to homogeneous pattern areas. The test is performed using a variation of the second scene under 50% resolution and 2% noise. The number of keypoints has been limited to 20, as can be seen in the simulation conducted when *ratio* parameter is set to 1.

of the descriptor likelihoods. Thus, an outlier correspondence would pose a challenge to the registration approach in the sense that it will not fulfil the geometric consistency constraints.

- Along the different iterations of the evolutionary stable strategy solver algorithm [Albarelli et al., 2009], successive incoherent correspondences will be removed according to the consecutive game payoff evaluations. As the manually altered correspondences are known, it is possible to evaluate the evolution of the ratio of inlier candidates regarding the remaining set. Ideally, the tendency to keep the correct candidate pairs in a monotonically increasing way must be validated.

- Starting from the initial set of 500 correspondences until a minimum of 3 finally selected matches (which is the minimum amount of correspondences needed in order to estimate a rigid spatial transformation for scene registration), the averaged results of this experimental set-up in the 12 models of the database can be displayed as depicted in figure 3.14. Different simulations are shown, starting with different outlier percentages within the 500 correspondences. The intention is to visualize the evolution of the inlier ratio, which should converge to a remaining set of correspondences where this ratio is 100%. This convergence must be reached in a monotonic increasing way as defined in [Albarelli et al., 2009]. Although there exist unusual cases where the inlier ratio evolution temporarily decreases (indicating the *sacrifice* of a correct correspondence at a given iteration), the overall crescent evolution shown in the figures certifies the correct performance of the approach even in cases where a very low presence of initial correct matches is set up.

An ideal comparative experiment must contrast the performance of the evolutionary game theory based approach against any commonly used iterative RANSAC-based approach. But the difference on paradigms complicates the establishment of any comparative criterion: while the approach presented in this chapter is based in a constraint-based rejection methodology, and we can evaluate the inlier ratio evolution of the remaining set of match candidates at each iteration; any RANSAC-based approach is usually based in an iterative hypothesis evaluation until a minimal error is found. At each iteration, a sub-sampling of point candidates takes place: this means that the ratio of inlier candidates will evolve in an erratic way during iterations. Furthermore, the number of iterations will vary for different executions of the method. And as each iteration hypothesis is a spatial transformation itself, its evaluation will also depend on an error threshold parameter which will directly affect the inlier discrimination. Finally, for this same reason, noise on data will also affect the system. This will not happen on a game theory approach as this method will implicitly select the elements in the payoff matrix with less noise, and the solution will be consistent no matter how many different executions are done.

For all these reasons, we finally compare both approaches via the number of iterations needed, which can be a good justification baseline. The proposed evolutionary game theory method, being a rejection based approach, will not surpass the number of initial candidates -500 in the current set-up- and this upper bound will be constant regardless the initial inlier ratio. On the other side, in an iterative approach the number of needed iterations can not be exactly established a priori: in this sense, [Hartley and Zisserman, 2003] provides a theoretic approximation for an estimation of RANSAC number of needed iterations N in order to solve a registration with a given amount of outlier elements:

$$N = \frac{\log(1-p)}{\log\left(1-(1-\epsilon)^s\right)} \tag{3.17}$$

where s is the sampling size which is taken at each iteration, ϵ is the probability that any match pair is an outlier, and p is the probability that the sampling set is free from outliers (usually set to 0.99). In a registration context the subsampling size for estimating a rigid transformation is set as $s = 4$. This is taken as the baseline of a common and standardized outlier removal procedure and can provide estimations like the ones presented in table 3.5

inlier %	75%	50%	25%	10%	5%	1%
N	8	34	292	**4603**	**36839**	**4605168**

Table 3.5: Estimation of RANSAC needed iterations N according to hypothetical percentages of initial inlier elements.

For the experimental set-up proposed here, a 21% of inliers would suppose the threshold value for which it is advantageous to choose the evolutionary game theory method. As a conclusion we can see the validity of the proposed method in the context of real scenes registration, where a limited set of initial candidate matches can be provided by a descriptor matching, but the nature of the scene itself can still provide a presence of repeated areas or symmetries, resulting in outlier match candidates. The presented approach can provide an efficient solution in a robust way, regardless of challenging conditions such as a high presence of outliers or noise on data.

3.5.6 Global Matching Evaluation

For testing the overall performance of the descriptor in conjunction with the correspondence selection stage, an exhaustive scene registration test has been designed so each one of the twelve models has been split in halves of different common overlap (from 10% to 70% of the surface in common). A random spatial transformation (arbitrary rotation and translation) is introduced to one of the halves. In

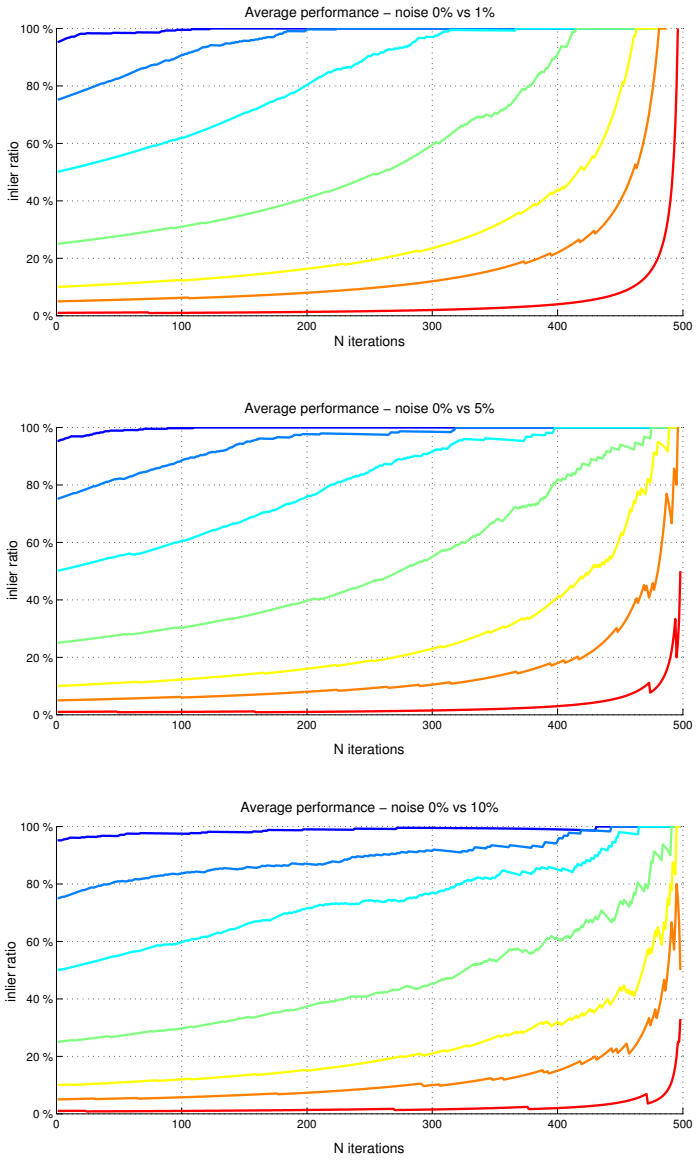


Figure 3.14: Evolutionary game theory approach performance on the removal of correspondences outliers. Each row depicts the average performance on the proposed set-up for all 12 models in the database, for different levels of noise: 1%, 5% and 10% the standard deviation of each feature. Each plot displays the evolution for 7 different initial situations: 95%, 75%, 50%, 25%, 10%, 5% and 1% of inliers: along the different iterations of the evolutionary game theory approach the percentage of correct correspondences is evaluated.

addition, each model is tested under different levels of noise, from 0 to 10% the standard deviation of color and surface coordinate values. For each scene, the experiment is conducted 5 different times so different halves and noise applications are considered. This leaves a total of 4620 registration executions, which have been evaluated as follows.

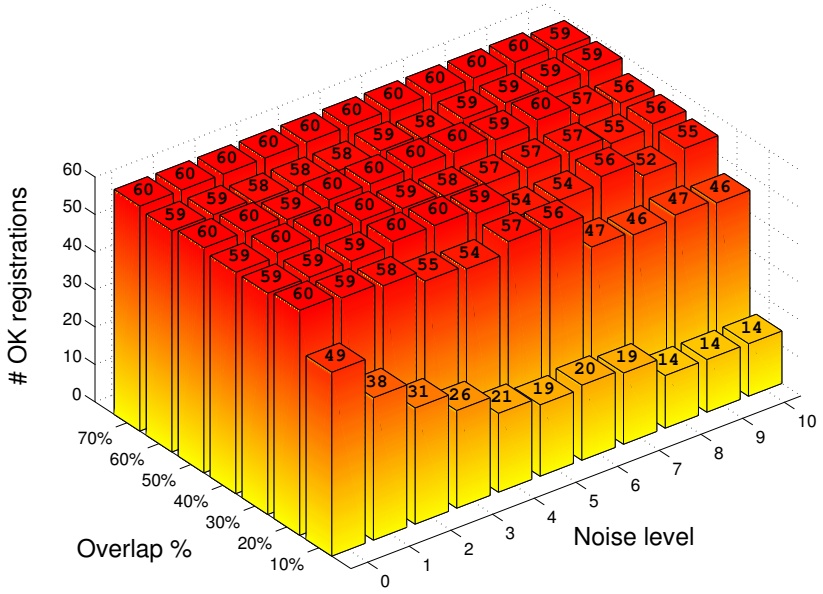


Figure 3.15: Histogram of correct registrations (for an error threshold of 0.02). As it can be see, the performance of the presented approach is rather homogeneous on most of the experimental conditions, even with low overlap between scenes and high levels of noise applied to data.

In order to consider a registration as correct, its registration error measure is evaluated by looking at the average Euclidean distance of ground-truth points in the common overlap surface. This is done after applying the found rotation and translation which undoes the arbitrarily applied rigid transformation. In the case of executions with applied noise, the system is solved using the modified data but the performance is evaluated on the equivalent un-noised scenes in order to be coherent on performance comparison. Object spatial coordinates are normalized so they fit within the boundaries of a prism of unitary volume, therefore the error measure is also normalized and is finally expressed as a percent ratio regarding the overall scene range. This way, results between different size scenes can be coherently compared.

An error acceptance threshold of 0.02 has been chosen, which would mean that objects of one cubic meter of volume should have an average error of 2 centimetres. By establishing this threshold the execution of all registrations can be represented by a histogram of how many of them are considered as correct, for each experiment conditions of noise and overlap. See figure 3.15 for such results representation. By watching this histogram, one can conclude which is the minimum overlap between scenes for which our approach is valid -at a 20% of common surface our method is able to perform most of the scene registrations in a correct way.

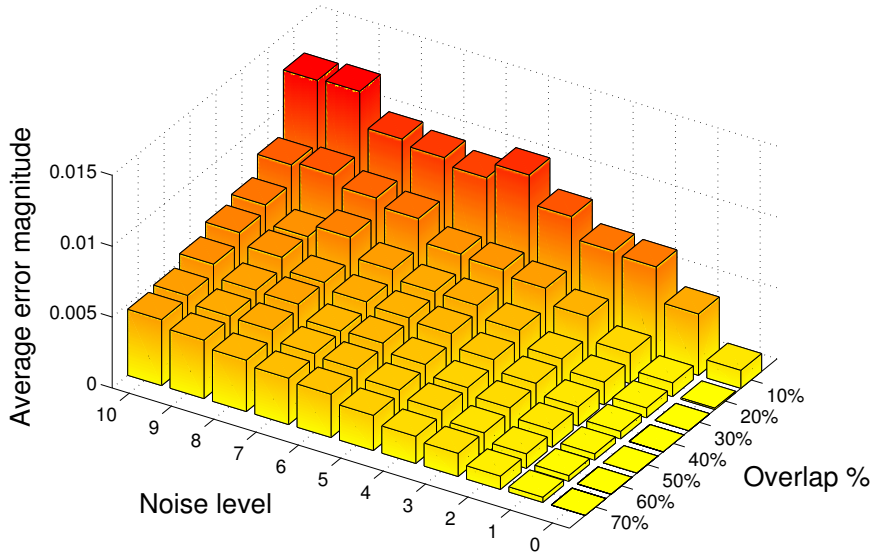


Figure 3.16: Average error distribution of those registrations considered as correct. As it could be expected, major errors occur on the cases of higher noise levels and less overlap.

Figure 3.16 shows the distribution of error magnitudes for the aforementioned correct scene registrations. As expected, the most challenging conditions are those where the system is tested with a smaller overlap and a higher noise. Nevertheless, by watching the value distributions on these figures, it can be conclude that the presented approach is more sensitive to the minimum overlap need rather than to the noise tolerance, which is coped by the descriptor performance -as tested in experiment 4.1. This is easily arguable, as the provided system needs a minimum of heterogeneous observable areas in order to find symmetric or repeated areas.

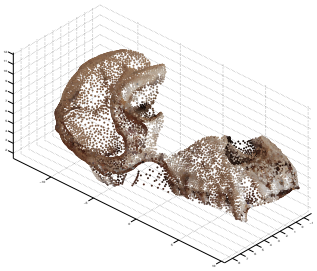
This experiment has been conducted in an Intel Core i5 computer with 4Gb of RAM. As stated before, the implementation of the proposed approach does

not pose major computational demands, and for the models in the database which have a density ranging from 20.000 to 30.000 points the whole registration execution time takes around 140 seconds in a prototype, non-optimized implementation. From this time, the scene analysis stage takes an average time of 17 seconds; the descriptor candidate matching and payoff matrix construction an average time of 90 seconds; and the evolutionary game theory solver algorithm an average time of 30 seconds. Figures 3.17 and 3.18 show some of the steps involved in the registration procedure, as well as a qualitative result on one concrete model of the database. The second view shown in figure 3.17b has been altered with an additive noise representing the 3% of the standard deviation on each of its coordinate and color components; and suffered an arbitrary spatial rigid transformation. The overlapping area between both views represents a 20% of the scene surface.

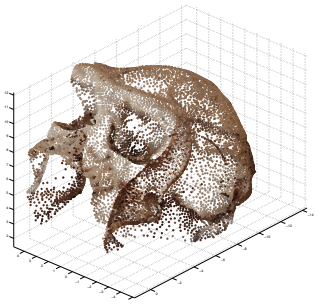
3.5.7 Real-data Matching Qualitative Evaluation

This last experiment proposes to test the complete approach in the context of scenes acquired with a Microsoft Kinect device, which suffer from sensor noise and artefacts along the capture of the different views. Therefore, the registration of such scenes has the drawback of not allowing a direct quantitative evaluation, as there does not exist any direct ground-truth information of correspondences between different views, and this converts this experimental set-up in a mere qualitative evaluation. Nevertheless, there are still several benefits in this framework which can help extracting conclusions about the provided methodology: in a first place, the usage of real data can validate the statement about the performance of the covariance descriptor against noise and resolution changes (in this case, caused stochastically by the acquisition sensor). In a second place, it will validate the application of the method under practical conditions like computational feasibility, or description of differently shaped objects -from planar to round. And finally, it will provide an example of broadening the scope of our approach to other areas such as scene understanding or object indexing under challenging conditions: the query objects belong to different views and can suffer small shape variations or lack of detail in certain parts. The main power of this game theoretic approach is that it will act as a best subset selector of points present both in the object query and the scene, enabling a robust searching procedure.

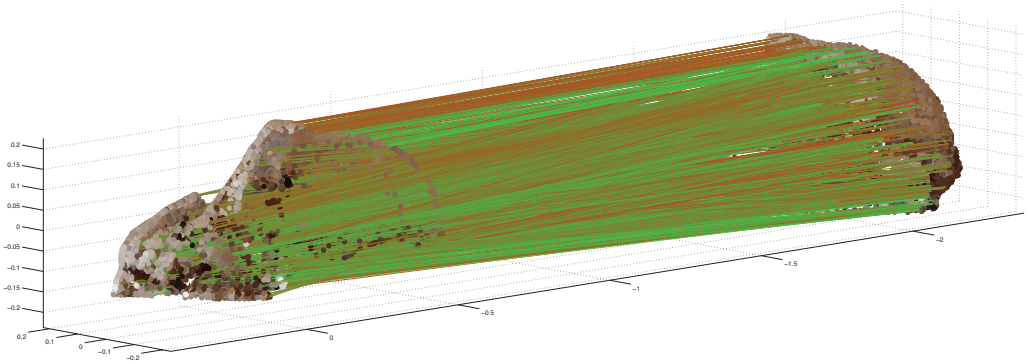
This experiment is performed on top of the publicly available RGB-D dataset presented in [Lai et al., 2011], which contains 300 objects organized into 51 categories as well as 22 different complete scenes. The main interest of using this database is that included objects and scenes suffer unstructured noise due to the own acquisition sensor, and segmented objects may have been acquired at different resolutions and static conditions with respect to the scenes. The goal is to perform a 3D object searching task: segmented objects will be used as query in-



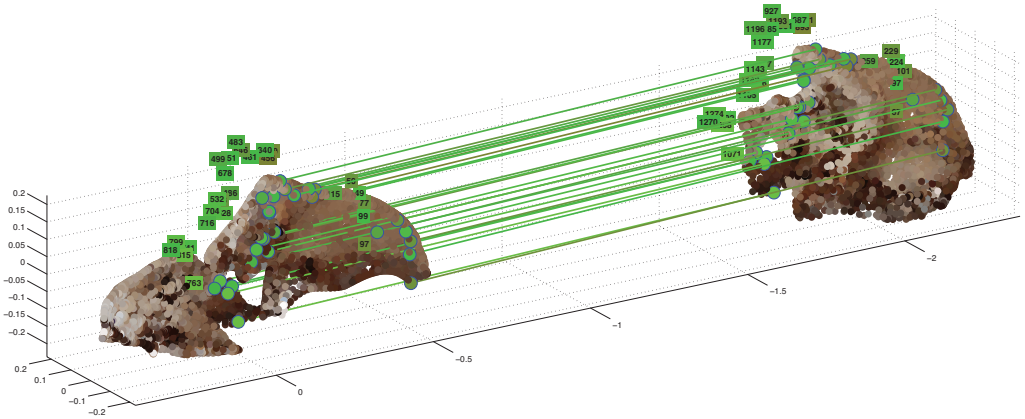
(a) First synthetic view.



(b) Second synthetic view.

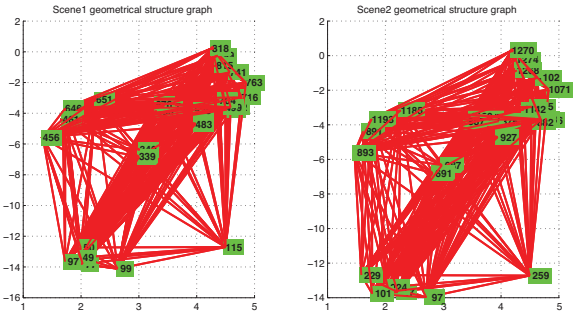


(c) Initial descriptor match candidates via inclusive ratio. The different levels of green on the correspondence lines indicate the likelihood magnitude.

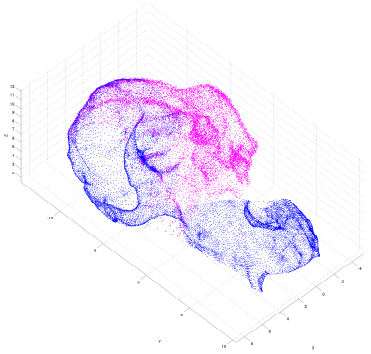


(d) Descriptor match candidates found as the result of the Evolutionary Game Theory solver. The parallellism level on the candidate lines indicates a good performance on registered point sets.

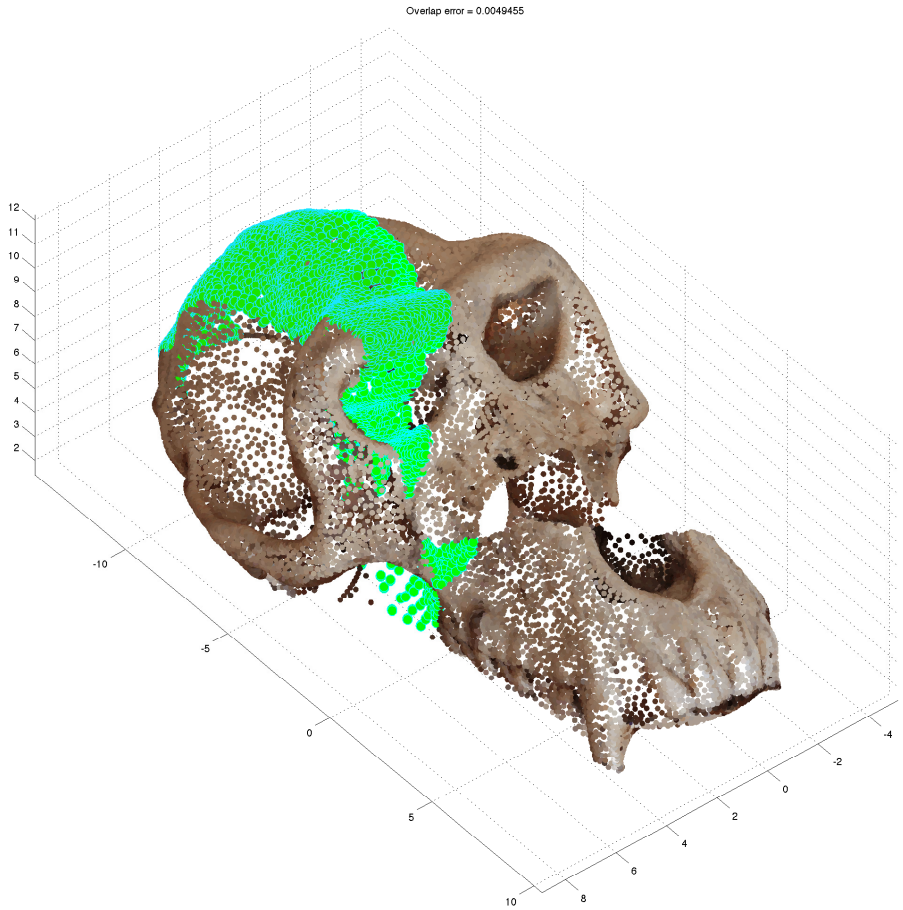
Figure 3.17: Example of some steps of the whole presented registration approach on an instance of the *Baboon* model, which poses a challenge due to its homogeneity in color and shape on some areas. Sub-figures (a) and (b) show the synthetic views to be registered. Sub-figures (c) and (d) represent the amount of match candidates found by the descriptor likelihood and after the evolutionary game theory solution, respectively.



(a) Geometrically consistent structure.



(b) Registered point sets. Each view is plotted on a different color.



(c) Registered point sets with marked groundtruth points and initial overlap area.

Figure 3.18: Sub-figure (a) shows a graph representation by projecting and connecting the final remaining points at each scene view in order to validate the geometric consistency constraints implicitly encoded in the game theory Solution. Sub-figures (b) and (c) show the final registration between scene views.

stances to be found within the whole scenes, where these instances will be mixed with clutter elements and altered by changes on resolution, spatial transformations or incomplete views. In this context, as we know that there will be at most one instance of each object in the scene, it is justified to use the *exclusive ratio* criterion for initial match candidates. The rest of the methodology presented before remains unaltered, but instead of putting two scene views of a similar size into correspondence, we seek a spatial registration for matching a smaller object inside the scene. The spatially translated points from the query model regarding the whole scene will be considered as object identifying points, therefore inferring the presence of the element in the scene.

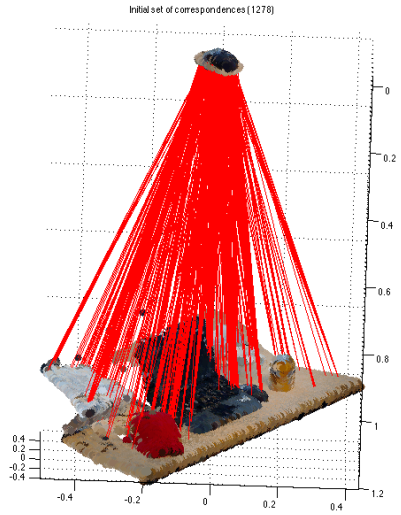
Different cluttered scenes and different query objects from the aforementioned dataset have been used, with different shape and texture distributions. Qualitative results are shown in figures 3.19 and 3.20.

3.6 Conclusions

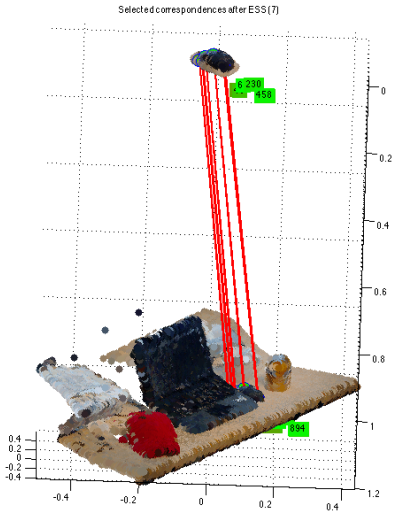
This chapter has introduced a novel descriptor for fusion of 3D shape and visual information which is defined to work under spatial rigid transformations and changes in noise and scene resolution. The rather simple formulation of this descriptor has several benefits: it can be extended with additional features in the future besides texture and surface information; it can be used as a salient point selector thanks to its underlying statistical notions; and the computational cost is low as the descriptor calculation does not involve any major operation than vector products and subtractions. Its flexible, compact and statistical-based conception has been analysed as the main reason of its high representation capabilities.

There are also practical advantages in the presented approach. On one hand, MCOV only requires a parameter for radial neighbourhood, which can be set according to each scene nature in a self-contained manner. Other methods will require fine-tuning of parameters for histogram bins, connection neighbourhood, etc. affecting directly to their performance. On the other hand, the Förstner distance defined in equation 3.6 is a geometrically sensitive metric for the inner topology of MCOV, which is coherent with its theoretic geometrical topology. Other methods are based on correlations or histogram distance definitions, which might add some error drift on the likelihood computation.

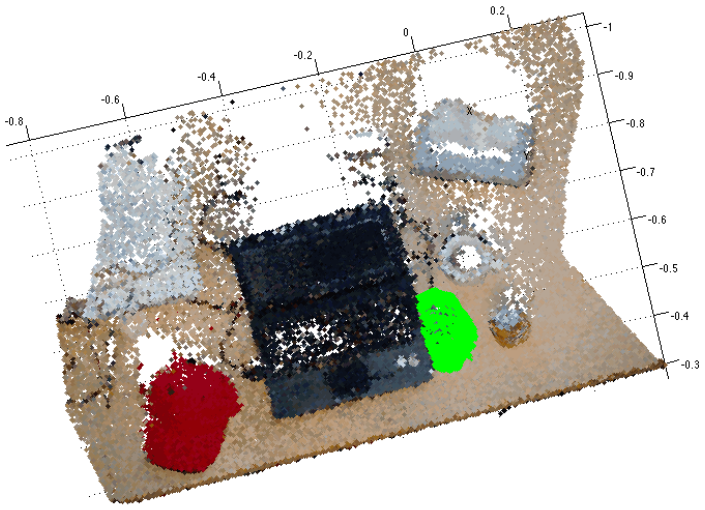
The results have been presented in conjunction with a tailored database of twelve scenes which include variant objects in order to represent challenging handicaps of repeated textures, homogeneous regions and symmetric areas. The proposed descriptor has been demonstrated to have a representative and discriminative capability which outperforms other state-of-the-art methods, specially in the case of noise over data, or density variations. The computational and perfor-



(a) Initial query matches for an instance of “mouse” object in “desk table” scene.

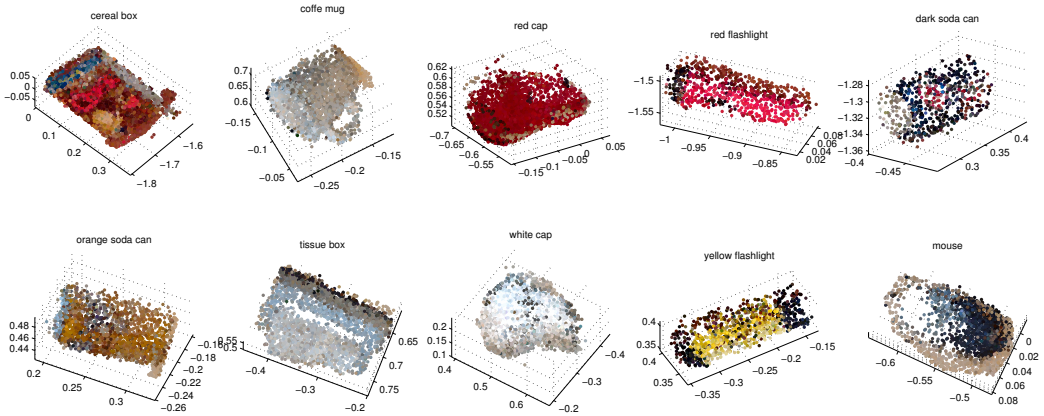


(b) ESS acts as the best subset selector of query matches inside the whole scene.

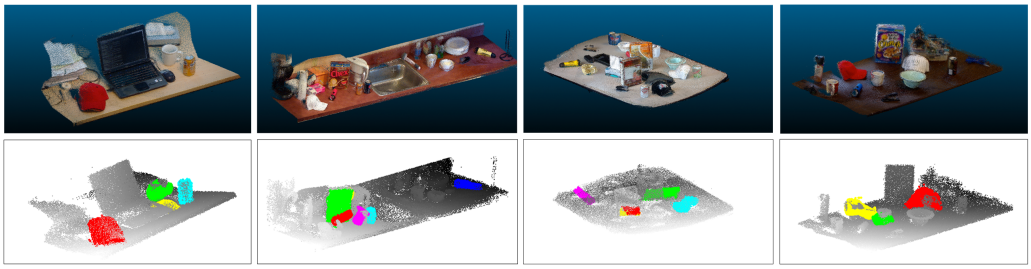


(c) Query instance is transformed according to the rigid transformation found after ESS solution. Inferred query points in the scene are marked in green.

Figure 3.19: Results from the experimental set-up performed on the RGB-D dataset. Sub-figures (a) and (b) show the different stages of our approach, where the set of descriptor candidates (many, due to low resolution of the query instance) are selected by the game theoretic approach. Even if the task is challenging due to changes on the quality and different views of the query object, our defined game achieves the goal of selecting that subset of points in the cluttered scene which is considered to belong to the query instance. In (c), the query instance is projected into the cluttered scene for an inference of its location.



(a) Different object queries in RGB-D dataset. Note the differences on views, resolution and nature of the objects.



(b) Examples of query search results in four cluttered scenes in RGB-D dataset.

Figure 3.20: Sub-figures (a) and (b) show the set of different query instances available in RGB-D dataset, and some examples of the query procedure depicted in figure 3.19 applied to different scenes (query inferences plotted in solid colours on top of grayscale scene point clouds).

mance benefits of the proposed approach suggest it is a flexible and easy descriptor with many practical applications for representation of scenes with current 3D and color sensors. Its associated keypoint detector feature can also be used on problems which require particular computational efficiency or point reliability.

We want to reflect that the different performance of the descriptor under different matching criteria is not a drawback, but the expected behaviour in this context. The choice between *exclusive* and *inclusive* criteria is a matter of knowing the task where the descriptor will be applied. For object recognition task matches, *exclusive ratio* will be suitable as it reduces the number of possible false positives and is more restrictive. For scene reconstruction problems, for instance, *inclusive ratio* will be more appropriate: in this kind of applications, area repeatabilities are a possible known handicap due to the nature of scenes (as they can contain homogeneous patterns). If a descriptor encodes the nature of an area, and this appears on some places along the scene, the repeatedly found points will be unavoidable (indeed, this asserts that the descriptor is doing what it is supposed to do). So, this reflects the needing of a posterior method which will filter false positive matches regarding a more global set of constraints, i.e. geometric consistencies. This has been hereby one of the motivations for the evolutionary game theory approach introduced in this chapter, which demonstrates the easy integration of the proposed covariance-based descriptor with meta-algorithms for taking advantage of its manifold properties.

Chapter 4

4D: Gesture Recognition in Depth Map Sequences

"Never mistake motion for action."

— ERNEST HEMINGWAY

EXTENDING A DESCRIPTOR FRAMEWORK one dimensionality step is not always direct to visualize. The gap from 2D images to 3D point clouds supposed a straightforward transition, selecting pixels within a radial neighbourhood of coordinates instead of pixel regions of interest in image planes. This chapter extends the covariance-based framework to a fourth dimension constituted by time, observing pixels not only on a single image plane but also in a "ray" along its temporal evolution in sequences of frames. Temporal variability is hard to encode so the framework extension needs some design strategies, presenting a nested "covariance-of-covariances" descriptor along the three plane directionalities of a sequence of images. As we will be dealing with the classification of recorded sequences embodying particular gestures with a start and ending rest position (hand sign language patterns, full body interaction and hand motions), the characterization of the motion patterns by the descriptor will be done by flattening the temporal information, fusing spatial features with their temporal variability.

4.1 Introduction

Automatic human gesture recognition is one other of the many challenges in computer vision research. While the traditional data feed has been based on monocular image sequences or not-so-accessible motion capture installations, recently appeared devices as Microsoft Kinect have eased the access to a valuable source of information as 3D depth. This could help not only on the interactive entertainment industry, but also broaden the scope of application to other areas with great

research impact as elder *exergaming* [Gerling et al., 2012], *eHealth* [García et al., 2012] or remote rehabilitation treatment [Lange et al., 2012].

It is encouraging to find a robust approach which is able to avoid usual existing problems: inter-subject and intra-class variations, repetitions in periodic motions, different speeds between different executions of the same gesture and temporal segmentation of a motion sequence (see figure 4.1). The compact, yet descriptive capabilities of covariance matrices of feature variations, rather than encoding the features themselves like other keypoint or histogram-based approaches, can provide a suitable methodology for dealing with these mentioned handicaps. The definition of 3D feature observations together with a covariance-of-covariances notation along the spatio-temporal domain of a gesture scene gives place to the 4DCov descriptor, and leverages the framework with respect to the methodology introduced in previous chapters. Again, this approach benefits from the implications of covariance matrices laying on a specific manifold, as this can provide a natural framework for a classification method adapted to such spatial variety. The spatial distribution of the descriptor has a meaningful nature as similar gestures appear spatially clustered, and samples which share motion patterns may yield to overlaps between class clusters. In this chapter the sparse collaborative classifier formulation introduced in chapter 2 will be used again, exploiting the spatial distribution of a set of descriptors for adding prior knowledge on the discrimination of a new gesture. This assertion is depicted in figure 4.2, where the spatial location of a set of 4DCov covariance descriptors is shown, mapped from their manifold to a Euclidean two-dimensional space via multidimensional scaling. The plotted embedded descriptors represent different sequences from the Gesture3D dataset [Kurakin et al., 2012], belonging to 4 different American Sign Language (ASL) gestures.

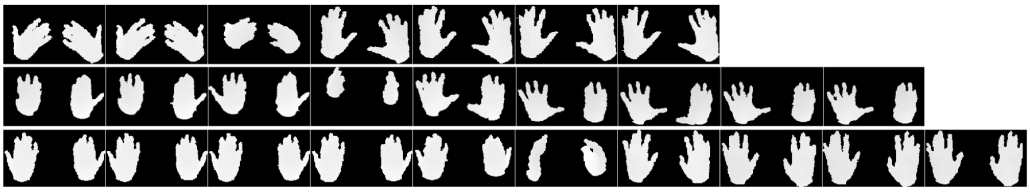


Figure 4.1: Three sequences (one at each row) depicting the same entity in American Sign Language (ASL), "finish". Despite the different number of frames and different cadence on the sign execution, the different sequences share the same hand motion, which is what really characterizes a gesture. This is an example of what should be taken into account by a spatio-temporal descriptor.

This chapter is organized as follows: section 4.2 reviews the related work available on the field of human gesture recognition from depth sequences. Section

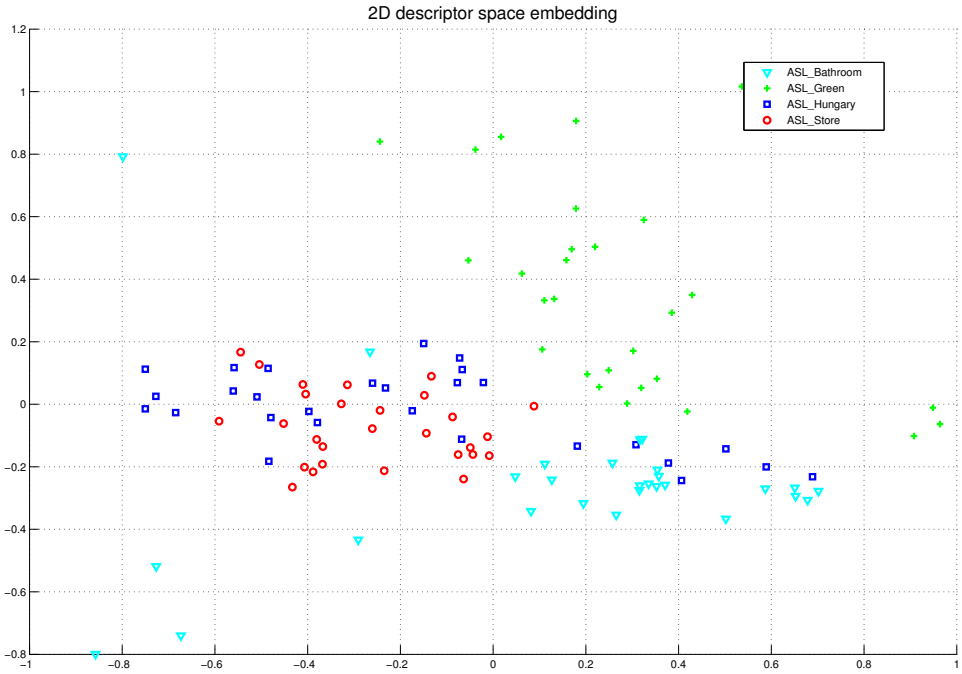


Figure 4.2: 2D space embedding of a set of 4DCov descriptors for 4 different hand sign language gestures. Classes may partially overlap due to similar motions on the represented gestures, but their localization is noticeably clustered which validates our idea of the proposed descriptor forming a natural spatial topology.

4.3 reformulates the covariance-based descriptor framework, which in this application includes the definition of the so-called 4DCov descriptor. In this case, the action sequence of depth images is considered as a volume which can be observed from different planes: not only in the natural frame-wise plane, but also as slices along the temporal axis. Therefore, for a given sequence this idea allows to obtain three sets of descriptors encoding not only the spatial variance along depth values of a frame, but also its variability along frames. This volume of descriptors is considered as a volume of features and nested into an other level of covariance-based descriptors, encoding an action sequence by a triplet of covariance matrices. This nested descriptor space is paired to the sparse representation classifier already introduced in chapter 2, which is recalled again in section 4.3.2. Finally, sections 4.4 and 4.5 evaluate the classification accuracy of the proposed method with respect to state-of-the-art approaches on top of four public human gesture datasets acquired with 3D depth sensor devices, including complex gestures from different natures, and provide conclusions and discussion on the methodology.

4.2 Related Work

Classic approaches on human gesture recognition in the computer vision area usually relied on part-based models for limb detection and tracking in order to infer a higher layer of abstraction where concrete motions could be identified. Survey works as [Turaga et al., 2008; Poppe, 2010] point this fact but also remark a new tendency where methods are starting to avoid human body models and focusing more on feature extraction from the available sequences. This results in approaches which are less dependent on separated segmentation and tracking algorithms.

Current approaches can be grouped into two common categories, whether if they aggregate information of the entire action sequence using different features; or use information from individual frames which are subsequently joined under some temporal structure modelling. Examples from the first group include approaches like [Bobick and Davis, 2001; Davis, 2001; Bradski and Davis, 2002] (using several dense or histogram based motion templates), [Ali and Shah, 2010; Guo et al., 2010] (via optical flow based kinematic features), [Kellokumpu et al., 2008] (LBP-like features) or [Thurau and Hlavác, 2008] (HOG based representations). Examples of the latter works can be found in approaches as [Wang et al., 2012b; Xia et al., 2012] (proposing different descriptor approaches for capturing frame-wise restrictions between 3D joints and a later temporal modelling in the frequential domain or via probabilistic graphical models). Some other methods extract interest keypoints at a frame level, such as STIP (Spatio-Temporal Interest Points) [Laptev, 2005; Bregonzio et al., 2009]. In any case, these later approaches require higher abstraction machine learning techniques for the temporal modelling of gestures, such as neural networks or probabilistic graphical models [Martens and Sutskever, 2011; Xia et al., 2012; Han et al., 2010]. Due to the large amount of parameters which should be estimated, these models require an often infeasible amount of data samples in order to reach acceptable accuracy levels on future estimations.

Novel approaches take advantage of low-cost and portable devices such as Microsoft Kinect for contributing to the concrete area of 3D depth-based gesture modelling. While some methods exploit the ability of extracting human skeleton joints (a feature which is achieved by algorithm packages on top of these devices [Li et al., 2010; Fothergill et al., 2012; Shotton et al., 2013]); others characterize motion patterns from direct depth cues for a later classification [Wang et al., 2012a; Fothergill et al., 2012; Oreifej and Liu, 2013]. Some contributions on this last direction [Liu and Shao, 2013; Negin et al., 2013] also deliver depth gesture datasets which can serve as a benchmark base for novel approaches.

4.3 4DCov Descriptor for Gesture Recognition

Recent approaches have ported the basis of covariance-matrix based descriptors to the temporal domain for characterizing gestures with encouraging results. The authors in [Hussein et al., 2013] propose a natural extension of the work in [Tuzel et al., 2006]: the temporal modelling is done by concatenating descriptors in a hierarchical relationship, and the classification of the resulting elements is performed by means of linear support vector machine classifiers. Even if this provides promising accuracy levels, this approach uses the covariance notion just for its compact representation and does not exploit any of its complementary geometrical benefits. The approach in [Sanin et al., 2013] is based on a similar point of view but the used features are dependent on the human joints estimated by an intermediate algorithmic layer of the Microsoft Kinect device. Therefore, it can not be applied to cases where there is only raw depth information available and joints can not be estimated, such as hand sign language. Because of these reasons, it was considered that there was enough place for proposing novel research and improvements on covariance based descriptors for gesture recognition from 3D depth sequences: in a first place, it is desirable to deal only with raw depth values. In a second place, the geometric properties of the covariance-based descriptor space should be taken into advantage as reviewed in previous chapters.

4.3.1 Spatio-temporal Coding of Features in Nested Covariances

For the task of gesture recognition, it must be identified how features are encoded both at a local frame and at a whole sequence level. Let A be the $W \times H \times N$ depth sequence of N instant frames of $W \times H$ pixels. Then a feature selection function for each point in the sequence $p = x, y, t$, $\Phi_{4D}(p)$ is defined as:

$$\Phi_{4D}(p) = \{\, \phi_p, \;\; \forall p \in W \times H \times N\} \tag{4.1}$$

where ϕ_p is a feature vector of random variables obtained at each one of the points p and is defined as:

$$\begin{aligned} \phi_p = \big[x,\; y,\; A_p,\; |A_p^x|,\; |A_p^y|,\; |A_p^{xx}|,\; |A_p^{yy}|,\; ... \\ ...\sqrt{|A_p^x|^2 + |A_p^y|^2},\; \text{atan}\left(\frac{|A_p^y|}{|A_p^x|}\right)\big] \end{aligned} \tag{4.2}$$

This feature vector includes the information about depth itself combined with other coarse observations such as first and second image derivatives, gradient magnitude and curvature. Features are normalized in order not to be influenced

by the different ranges on the sample space of these variables. The correlation of these cues is encoded inside the covariance matrix formulation as follows:

$$C_{slice}\left(\Phi_{4D}(x,y,z)\right) = \frac{1}{S-1}\sum_{i=1}^{S}\left(\phi_i - \mu\right)\left(\phi_i - \mu\right)^T,\ \forall x,y,z \tag{4.3}$$

where μ is the vector mean of the set of S observed feature vectors $\{\phi\}$.

This formulation provides three sets of covariance descriptors for each one of the "slicing" directions on the action volume. These directions belong to the three planes on the action sequence, depending on the fixed values for x, y and t as sketched in figure 4.3. The notion behind this comes from considering a gesture depth map as a three-dimensional "silhouette", as inspired by existent approaches as [Blank et al., 2005], with associated characteristic variabilities in depth values along three orthogonal planes (horizontal and vertical frames, and the temporal projection of each pixel). Therefore, a given action sequence A will be characterized by three sets of descriptors, $\{C_X\}, \{C_Y\}, \{C_T\}$, encoding the spatial or temporal changes within a row, column, or temporal plane of the sequence, respectively. The goal is to capture not only the spatial changes within a frame, but also the temporal evolution of a given pixel content along its temporal domain.

The three "slice" sets $\{C_{X,Y,Z}\}$ are formed by three sets of W, H and N 9×9 symmetric matrices, according to the sequence width, height and number of frames.. The diagonal entries of these matrices will represent the variance of each one of the feature distributions, and the non-diagonal entries will represent their pairwise correlations, for all planes observed on the sequence along the three directions x, y and t. The loss of structural information along the sequence is positive in order to identify gestures with different cadences due to each subject execution, or with different orderings or repetitions (e.g. waving a hand or indicating to approximate). Also, if there is any individually corrupting sample due to noise or other artefacts, it is naturally filtered out in the computation of the descriptor due to the mean subtraction.

In order to scale up the definition of a descriptor for the whole sequence A, the own covariance magnitudes calculated as the result of these plane-wise descriptors are considered as new features for a set of higher order descriptors, extracting their independent coefficients by the vectorization operation defined in equation 2.11 and repeated here:

$$vect(C) = \left[C_{1,1}\ \sqrt{2}C_{1,2}\ \sqrt{2}C_{1,3}\ ...\ C_{2,2}\ \sqrt{2}C_{2,3}\ ...\ C_{d,d}\right] \tag{4.4}$$

The vectorization of a $d \times d$ covariance matrix is a minimal representation of all its $d\ (d+1)/2$ independent coefficients, which are found in the upper or lower

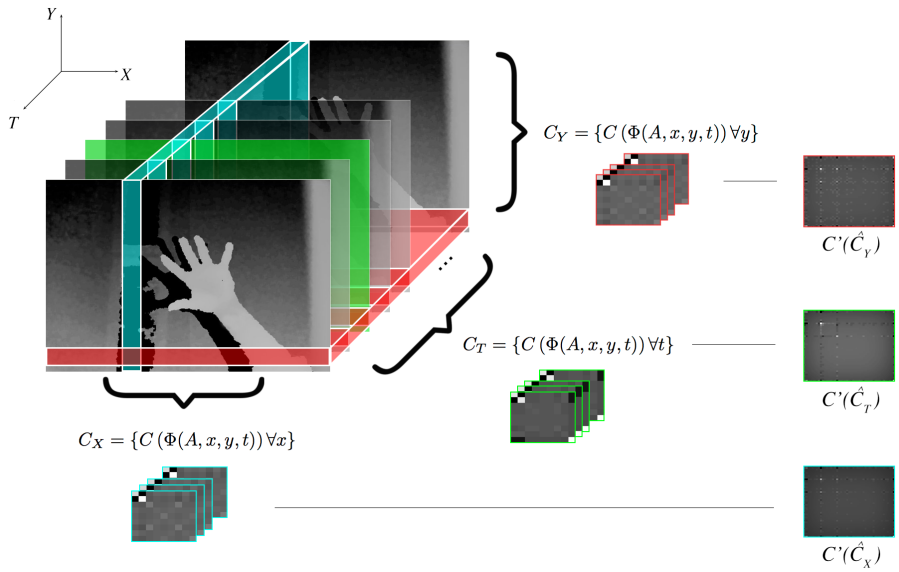


Figure 4.3: Sketch of the computation of covariance descriptors along the three orthogonal planes in x, y and t dimensions. The vectorized sets $\hat{C_X}$, $\hat{C_Y}$, and $\hat{C_T}$ will serve as the observations for three scene-wise covariance descriptors, correlated again in equation 4.5 for a final descriptor.

triangular part of the matrix. As the off-diagonal entries would be counted twice in a norm computation, they are scaled down in this operation by the $\sqrt{2}$ coefficients. Each slice direction yields to n covariance descriptors of size of 9×9 ($\mathbb{R}^{n \times 45}$ in its the vectorized form) where n is the number of descriptors obtained according to the number of columns, rows or frames on the sequence, respectively. The three sets of vectorized covariance descriptors for the three plane directions are the new features of this covariance-of-covariances notion:

$$C'(\hat{C}) = \frac{1}{S-1} \sum_{s=1}^{S} \left(\hat{C}_s - \mu_S\right) \left(\hat{C}_s - \mu_S\right)^T \tag{4.5}$$

being $\hat{C} = \{vect\,(C\,(\Phi_{4D}(x,y,z)))\,\forall x,y,t\}$ the independent coefficients from vectorized covariance descriptors for each one of the planes of the sequence in x, y or t, and μ_S its vector mean. While the "slice"-wise covariance descriptors in equation 4.3 expressed the local variability of the features at each plane, the newly defined global scene covariance matrices gather the whole spatio-temporal evolution of these variabilities and characterize a complete gesture in the depth scene by a triplet of higher order descriptors, $< C'(\hat{C_X}), C'(\hat{C_Y}), C'(\hat{C_Z}) >$ of size $3 \times (45 \times 45)$.

For a compact notation, these three final sequence descriptors can be vectorized ($3 \times \mathbb{R}^{1035}$) and concatenated laying in $\mathbb{R}^{3105}$ regardless of the number and size of frames of the sequence. This notation is reversible and allows to come back to the definition of a sequence by its three higher order descriptors, which will be used later on the classification method (in equation 4.10).

4.3.2 Collaborative Sparse Classification

The representation capabilities of the proposed descriptor have a direct relationship with a topological layout where similar motions stay close in the descriptor space. This yields to consider a geometrically sensitive classification method which can exploit the descriptor manifold distribution to its best extent. Sparse representation based classifiers, as already reviewed in chapter 2, have shown a recent rise in the machine learning community in the context of face recognition [Wright et al., 2009, 2010; Zhang et al., 2011]. The sparsity and collaboration concepts are defended in order to cope with cases were the training set is complex, not only because a low availability of learning samples, but also because an unknown element can share characteristics from different classes. A recap of the methodology already used on image classification is provided and the regularization constraint is adapted to the new compound descriptor nature for gesture sequence classification.

Sparse representation based classifiers propose to consider a test sample y as a linear combination of elements in a dictionary A of training samples from different classes: $y = A\alpha$, where α is the sparse vector indicating the weight coefficients for each element in A. As the sample y should ideally be represented by using the less number of samples, and as accurate as possible, α is found forcing its sparsity via its L1 norm minimization constrained as follows:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left\{ \|\alpha\|_1 + \|y - A\alpha\|_2^2 \right\} \tag{4.6}$$

Then, given $\hat{\alpha}$, the classification label for y is determined by the subset of training samples of a given class i which provides the minimum representation error:

$$class(y) = \underset{i}{\operatorname{argmin}} \left\{ e_i \text{ s.t. } e_i = \|y - A_i\hat{\alpha}_i\|_2 \right\} \tag{4.7}$$

Adding a manifold-aware minimization constraint brings back sparsity conditions and also provides an inclusion of prior knowledge on the descriptors geometric distribution. It also helps on relaxing the computational expense and adds stability to the presented method. There exist several metrics for the symmetric positive definite matrices manifold where 4DCov descriptor lays, which are specifically focused on the retrieval of matrix similarities on close neighbourhoods [Arsigny

et al., 2006; Cherian et al., 2013]. The proposed regularization term will be based on the manifold metric defined by W. Förstner in [Förstner and Moonen, 1999], already used in previous chapters. This distance definition preserves the global geometric relationship of the descriptors as the involved generalized eigenvalues between two covariance matrices express the magnitude of their geodesic distance:

$$\delta(C^1, C^2) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_i(C^1, C^2)} \tag{4.8}$$

where $\lambda_i\,(C^1, C^2)$ is the set of generalized eigenvalues of C^1 and C^2 according to their dimensionality d.

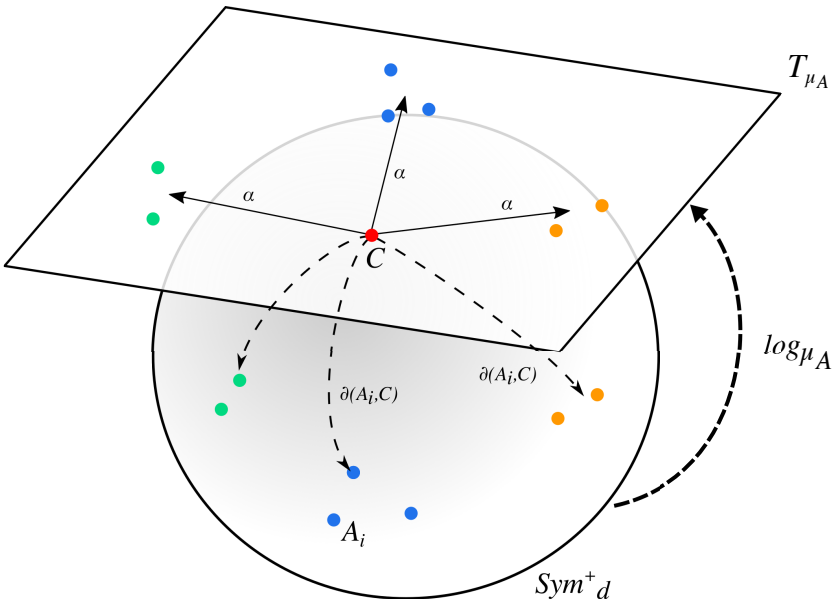


Figure 4.4: Schema of sparse representation based classification method with manifold regularization constraints.

Let A be the whole set of n training samples from K different classes, $A = [A_1, A_2, ..., A_K] \in \mathbb{R}^{d \times n}$, where each $A_i = \left\{\text{vect}(\log_{\mu_A}((C_A))^T\right\}$ is the set of vectorized 4DCov descriptors which form the subset of training samples for the class i and d is the vectorized 4DCov descriptors size (3105). Then, a test sample in the form of a vectorized covariance descriptor $\hat{C} \in \mathbb{R}^{3105}$ can be expressed as a linear combination of the available set of training samples: $\hat{C} = A\alpha$, being $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K]$ a vector of weights corresponding to each one of the training samples in A. See figure 4.4 for a schema of this classification paradigm. Then,

a regularized variation of the minimization expression defined in equation 4.6 is defined as follows:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left\{ \|\hat{C} - A\alpha\|_2^2 + \|D\alpha\|_2^2 \right\} \tag{4.9}$$

where D is a diagonal matrix of size $n \times n$ which allows the imposition of prior knowledge on the solution with respect to the training set, using the covariance matrices metric defined in equation 4.8. This term contributes also on making the least squares solution stable, and on introducing forehand sparsity conditions to the vector $\hat{\alpha}$. D is defined as:

$$D = \begin{pmatrix} \delta(A'_1, C') & & 0 \\ & \ddots & \\ 0 & & \delta(A'_n, C') \end{pmatrix} \tag{4.10}$$

where $\delta(A'_i, C')$ represents the addition of the Förstner distances between each of the three sequence descriptors contained in the vectorized forms A' and C'. The solution to the sparse collaborative representation, $\hat{\alpha}$, can be calculated by the following derived expression according to [Zhang et al., 2011]:

$$\hat{\alpha} = \left(A^T A + D^T D\right)^{-1} A^T \hat{C} \tag{4.11}$$

Finally, the classification label of the test sample $\hat{C}$ can be obtained by observing the regularized reconstruction residuals from the resulting sparse vector $\hat{\alpha}$:

$$class(\hat{C}) = \underset{i}{\operatorname{argmin}} \left\{ \frac{\|\hat{C} - A_i \hat{\alpha}_i\|_2}{\|\hat{\alpha}_i\|_2} \right\} \tag{4.12}$$

4.4 Experimental Results

This section evaluates the presented method on four publicly available datasets which contain depth sequences of different gestures: Microsoft Research *Action3D* [Li et al., 2010] and *Gesture3D* [Kurakin et al., 2012] datasets, respectively, contain gestures involving full body actions and American Sign Language gesticulations. Sheffield Kinect Gesture dataset (SKIG) [Liu and Shao, 2013] gathers a set of different forearm gestures suffering variations on hand pose, background and illumination. Finally, the *WorkoutSU10* dataset [Negin et al., 2013] contains a collection of sequences recorded from different subjects performing several full body exercises for therapeutic purposes. Most of these datasets have established benchmarking conditions for state-of-the-art approaches working on depth based gesture detection. Therefore the presented results will be compared

against state-of-the-art classification methods proposed or already tested in papers related to the different commented datasets.

4.4.1 Tests on Action3D Dataset

The Action3D dataset [Li et al., 2010] is a collection of depth sequences acquired with a Microsoft Kinect device. It contains a total of 20 full-body actions ("high arm wave", "high throw", "draw circle", "hand clap", "tennis serve"...) which are performed 3 times by 10 different subjects. Some examples of depth sequence frames are shown in figure 4.5.

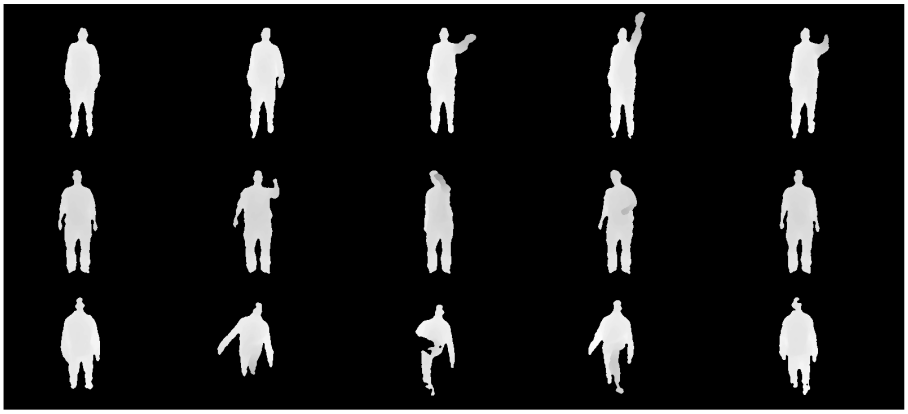


Figure 4.5: Action3D sample depth frames for different gesture classes at each row.

The provided 4DCov method is compared to the results proposed on two recent works [Li et al., 2010; Oreifej and Liu, 2013] which use a graphical model for the temporal modelling of 3D points evolution and a histogram of normals based descriptor respectively. Their respective authors also test other state-of-the-art methods, whose results are collected here as well. The same experimental conditions as the ones applied to the methods tested in these works are used: the classification takes place on the whole set of 20 classes and the training and validation sets comprise folds using half of the samples for each set respectively. The classification accuracy results are represented in table 4.1.

4.4.2 Tests on Gesture3D Dataset

The Gesture3D dataset is defined in [Kurakin et al., 2012] along a graphical model for the real time classification of American Sign Language (ASL) gesticulations. This method shares some similarities with the provided sparse collaborative classifier as states of the presented graphical model are shared between similar ges-

Method	Accuracy %
4DCov + Sparse Collab. Classifier	**93.01** %
HON4D [Oreifej and Liu, 2013]	88.89 %
Wang et al. [Wang et al., 2012b]	88.20 %
HOG3D [Klaser et al., 2008]	81.43 %
Li et al. [Li et al., 2010]	74.07 %

Table 4.1: Comparison results of classification accuracy levels (%) on the complete Action3D dataset

tures, in the same fashion as the presented method allows several samples from different classes to collaborate on the classification decision thanks to the descriptor manifold regularization term. The dataset is a collection of 12 ASL signs performed 3 different times by 10 subjects. Some examples of the provided depth sequences are shown in figure 4.6.

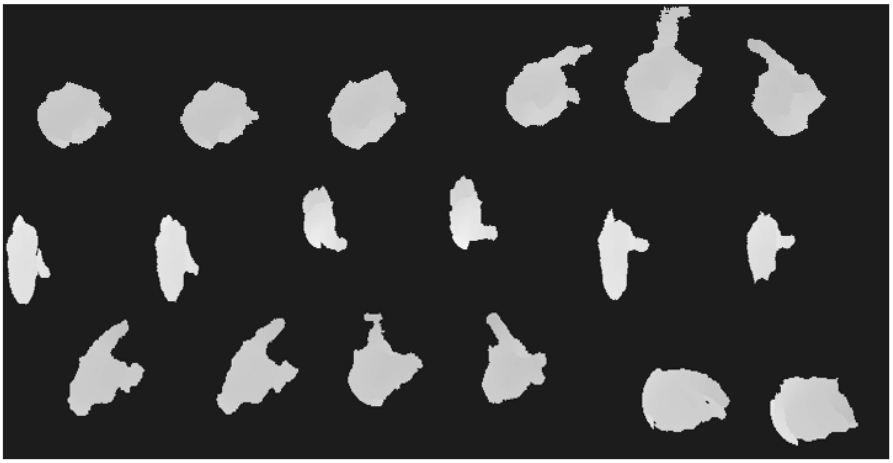


Figure 4.6: Gesture3D sample depth frames for different gesture classes at each row.

The original paper from Kurakin *et al.* proposes several algorithms for the implementation of their graph model-based methodology, reaching a maximum classification accuracy ratio of 87.7 % when using 9 out of the 10 available subjects for training 5 random folds of the dataset. For a stable comparison, results from [Oreifej and Liu, 2013] are gathered, which use the same dataset for their classification method and at the same time contrasts their results with other state-of-the-art gesture classification approaches, using cross-validation folds of half of the subject samples for learning and the other half for validation. Results are

presented on table 4.2.

Method	Accuracy %
4DCov + Sparse Collab. Classifier	**92.89** %
HON4D [Oreifej and Liu, 2013]	92.45 %
Wang et al. [Wang et al., 2012b]	88.50 %
Kurakin et al. [Kurakin et al., 2012]	87.70 %
HOG3D [Klaser et al., 2008]	85.23 %

Table 4.2: Comparison results of classification accuracy levels (%) on the Gesture3D dataset

4.4.3 Tests on Sheffield Kinect Gesture Dataset

The Sheffield Kinect Gesture (SKIG) dataset [Liu and Shao, 2013] gathers a set of 10 different forearm gestures (circular, triangular, up-down, "approach", "turnaround" and other motions) under different hand poses, background and illumination variations from 6 subjects. This dataset poses an interesting testing benchmark as the gestures involve fewer and more rigid subject body parts regarding previously presented databases. In any case, the presence of different actions with different patterns and variations of speed on their executions are of valuable use to proof the discriminative power of the 4DCov descriptor and its paired classification methodology. Figure 4.7 presents some sample frames from these sequences.

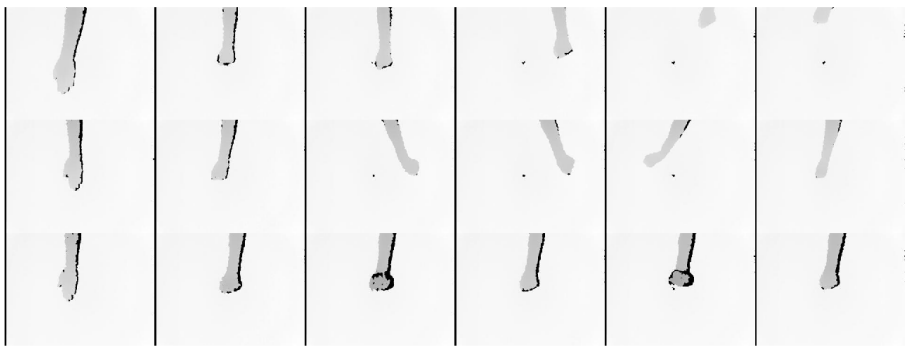


Figure 4.7: Different SKIG dataset sample frames for different gesture classes at each row. Background is noticeably lighter than previous datasets as arm gestures are performed on top of a table close to the subjects.

While the authors in [Liu and Shao, 2013] normalize the sequences to a given size according to their method, this step is omitted as the scale factor both on

frame size and temporal spanning is irrelevant to the 4DCov descriptor. The original paper also compares results from RGB and depth cue recordings used both together and separately, and concludes that depth-only information is way more challenging in discrimination terms. This is also relevant to these experiments as raw depth information is the only cue used in the provided methodology. We can therefore see the performance gain of this method when other state-of-the-art, histogram and keypoint-based approaches, are used on raw depth sequences.

Results on the classification accuracies of the different methods are presented in table 4.3. According to [Liu and Shao, 2013], for a fair comparison, the same experimental conditions have been used: a cross-validation procedure with training sets consisting on 4 subjects and validation sets of the remaining ones. As previously commented, results obtained from depth only information are used in the comparison. As observed, the 4DCov method outperforms the classification accuracies of state-of-the-art approaches presented in [Liu and Shao, 2013], both for the method presented by the authors themselves and for other approaches which are based on linear SVM classifiers on top of state-of-the-art spatio-temporal descriptors.

Method	Accuracy %
4DCOV + Sparse Collab. Classifier	**93.8** %
RGGP [Liu and Shao, 2013]	76.1 %
HOG3D [Klaser et al., 2008]	75.4 %
HOG/HOF [Laptev et al., 2008]	72.1 %
3D-SIFT [Scovanner et al., 2007]	61.3 %
SURF3D [Bay et al., 2008]	55.1 %

Table 4.3: Comparison results of classification accuracy levels (%) on SKIG dataset (depth channel only)

4.4.4 Tests on Workout SU-10 Dataset

WorkoutSU10 [Negin et al., 2013] is a valuable dataset as it provides a real-world environment for testing the introduced methodology: it contains sequences from 15 subjects with different morphologies, which are recorded in a domestic environment with background clutter, performing 10 different fitness exercises under the supervision of professional trainers for therapeutic purposes. Therefore, motion executions suffer variations on time spanning and gesture definition regarding each subject. Some samples are included in figure 4.8.

Together with the dataset, Negin *et al.* provide a methodology for gesture classification from depth sequences via linear SVM classifiers on random decision

Figure 4.8: Sample frames from WorkoutSU10 dataset. Each row represents some frames from different gesture classes.

forests for feature selection. Their method is based on features extracted from the spatial location of human body joints inferred by the Microsoft Kinect SDK. Therefore we can contrast the accuracies of 4DCov depth-based approach against a methodology based on inferred skeletal information, which Negin *et al.* defend as a powerful and richer representation of human body motion.

It is shown how the 4DCov and sparse based representation method performs with similar levels of action recognition accuracy with total independence from the acquisition device software, as this method works with raw depth cues instead of skeletal information. The experimental setup in [Negin et al., 2013] proposes a cross-validation procedure where different folds use the half of the sequences for training and the other half for validation. In that case, and after the normalization of available sequences, Negin *et al.* obtain a $98\% \pm 2.34$ accuracy ratio. Under the same conditions 4DCov obtains a $95.48\% \pm 0.7191$ classification accuracy.

The classification accuracy of the presented method is sensibly lower with respect to the baseline provided by Negin *et al.*, nevertheless our approach deals with raw depth information instead of joint information inferred by the acquisition device software. We consider this as an advantage in terms of method independence, as joint inference is a device-specific algorithmic model layer that can not always be available.

4.5 Conclusions

The strong point of this chapter resides in the demonstrated flexibility of the covariance-based framework for more abstract applications such spatio-temporal modelling of raw depth sequences. The conceptual difference regarding previous chapters is found in the nested conception of the 4DCov descriptor: regular covariances from plane-wise features are obtained in a first instance, and later used as new features for a second layer covariance descriptor. This provides a common-size abstract signature for gesture sequences, regardless of their number of frames. The descriptive capabilities of the presented 4DCov descriptor lay on the characterization of motion patterns from its feature variations along spatial and temporal planes, rather than encoding details of the features themselves along the action sequence volume. A classification method which takes advantage of the geometrical topology of the descriptor in order to reflect prior knowledge from the gesture descriptor space in its formulation is also presented.

Experimental results have demonstrated the performance and power of this enclosed methodology. The different nature of each one of the four tested datasets is valuable in order to extract conclusions about the observed accuracy ratios of the compared methods. Action3D provides depth maps with high degrees of intra-class variation due to full body gestures. The hand sign languages found in Gesture3D represent a more concentrated entity in motion, combining both the overall movement of hands and the inner variability of fingers. SKIG dataset is somewhat similar to Gesture3D but the paper presented by their authors compares accuracy results from 3D keypoint-based descriptors, which is valuable for us in order to contrast the unstructured statistical approach of 4DCov. Finally, WorkoutSU10 allows a comparison of performance on gesture recognition from inferred skeletal joint locations against raw depth data. The common advantage of the presented method is the inherent gesture characterization, "flattening" the gesture time dependence by using the feature variabilities in the nested descriptor formulation. This simplifies the classification approach as the descriptor space contains all the gesture information in a compact notation, and does not require to use temporal modelling approaches. Of course this is valid in bounded contexts such as the ones found in the presented datasets (sign languages, body motion and exercising, hand gestures) as these motions are homogeneous and repetitive, without spurious motion artefacts. While this is feasible in controlled environments, and temporally segmented gestures, it does not suit real-environment situations in which the subjects may interact freely with external inputs. That is the reason why one of the continuity lines of this works is based on returning to a descriptor approach closer to previous chapters, in which descriptors from depth features are obtained at each frame, and the classification takes place by a manifold alignment procedure, simi-

lar to dynamic time warping on the space of symmetric positive definite matrices.

Chapter 5

3D Medical Imaging Analysis

"Once a year, go someplace you've never been before."

—Dalai Lama

OPENING THE DOOR to new applications of a developed framework is always interesting for its validation, and for discovering that a work has practical outcomes besides its original conception. We have had the opportunity to test our covariance-based framework in a complementary field as 3D medical imaging, for the analysis of computerized tomography (CT) imagery. CTs produce dense three-dimensional volume scans of bodies based on their ability to block certain ray beams, according to their density properties. Making use of computer-processed combinations of many scan images taken from different angles, cross-sectional images of specific areas of a scanned object can be produced. This allows to represent internal structures of a body without invasive prospecting.

The parallelism of the nature of this data with the covariance-descriptor framework for 3D scene understanding provided an interesting baseline for developing a dense three-dimensional descriptor based on tissue texture characteristics. With the valuable feedback of expert clinicians, this has been used in several applications as case-based retrieval and tissue classification, segmentation and modelling.

5.1 Introduction

Clinical research has identified morphological tissue properties as indicators of cancer aggressiveness [Yokose et al., 2000] in lung tissue. Texture and size of the solid and ground-glass opacity (GGO) components of a nodule, as observed from CT images, can provide reliable cues in order to assess medical examination criteria [Depeursinge et al., 2015], but region texture delineation and classification is still an open and time-demanding problem. According to the clinical knowledge about the typology of ground-glass opacity and solid tissue, it is established

that the compactness, size, density and homogeneity of a nodule is differentiated from healthy lung regions, despite the large variability of normal lung tissue. The discriminative capabilities of these visual cues can be tackled from a pattern recognition approach, which motivates and settles the basis of the work presented in this chapter.

The remaining part of this chapter introduces an enclosed methodology for tissue texture characterization in lung CT images. The 3D covariance-based descriptor framework is adapted to the usage of 3D Riesz features for tissue morphology characterization: this provides a compact and flexible representation thanks to the use of feature variations rather than dense features themselves and adds robustness to spatial changes. Furthermore, the particular symmetric positive definite manifold of covariance matrices is exploited once again in a classification model following a "bag of covariances" paradigm in order to distinguish three different nodule tissue types in CT: solid (the main part of damaged tissue on a cancerous nodule), ground-glass opacity (the external part of the nodule), and healthy lung. The method is evaluated on top of an acquired dataset of 95 patients with manually delineated ground truth by radiation oncology specialists in 3D, and quantitative sensitivity and specificity values are presented.

5.2 Lung Tissue Classification in CT Images

In computer vision research, several descriptors for 3D object classification have appeared ([Zaharescu et al., 2009; Tombari et al., 2010; Rusu et al., 2009; Flint et al., 2007], amongst others). Nevertheless, these descriptors are usually targeted to 3D surfaces instead of 3D dense volumes as is the case in CT images. In the medical imaging domain, the survey conducted in [Depeursinge et al., 2014] points out relevant techniques in applied 3D solid texture analysis and highlights the importance of multi-scale directional convolutional approaches that are non-separable to characterize subtle and discriminative properties of 3D biomedical textures.

In this area, Riesz-wavelet features have demonstrated great representative capabilities: they characterize the morphology of tissue density thanks to their response to changes in CT intensities. These features are expressed by the response magnitudes to a set of 3D multiscale filters applied to the CT volume. This theoretically solid texture definition is used for proposing its integration into a covariance-based descriptor, with the goal of establishing a paradigm for 3D region definition and classification. The main benefits of covariance descriptors include the robustness to spatial transformations such as rotations, as well as the tolerance to changes in shape, size and resolution in the 3D domain. This is due to the fact that feature variation observations inside a region are used, instead of

absolute feature values, and any structural information about feature location is discarded. Furthermore, as covariance descriptors are embodied by covariance matrices, they lie in a meaningful and geometrically coherent descriptor space: similarly textured regions appear clustered in a low dimensional and analytically operable space. A part-based model is proposed in order to represent the entire possible space of lung tissue types in this particular manifold, and model the underlying three classes of interest (GGO, solid and healthy tissue).

5.2.1 3D Riesz-Covariance Based Descriptors

3D multiscale Riesz filterbanks are used to characterize the texture of the lung parenchyma in 3D at a given CT energy level. The N–th order Riesz transform $\boldsymbol{\mathcal{R}}^{(N)}$ of a three–dimensional signal $f(\boldsymbol{x})$ is defined in the Fourier domain as:

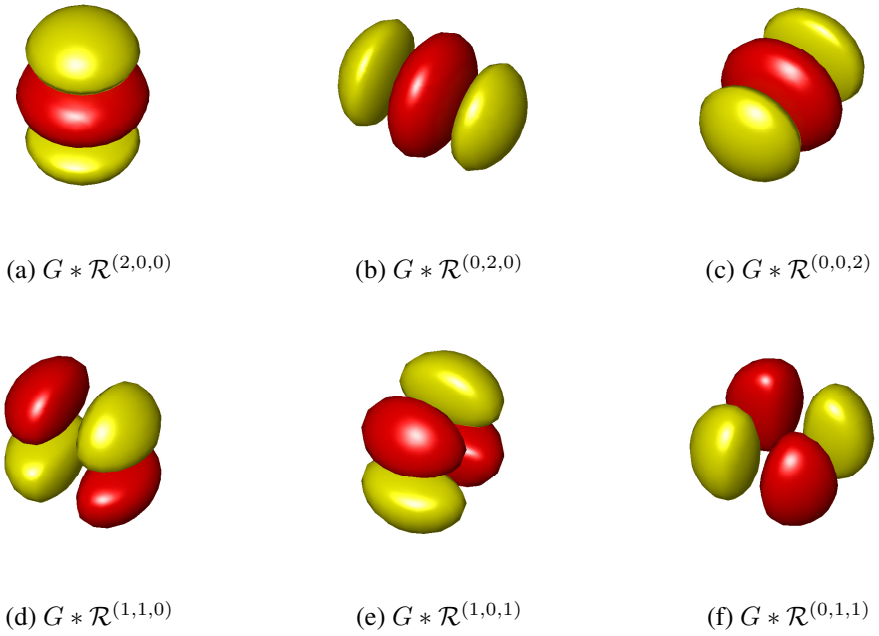
$$\widehat{\mathcal{R}^{(n_1,n_2,n_3)}f(\boldsymbol{\omega})} = \sqrt{\frac{n_1+n_2+n_3}{n_1!n_2!n_3!}}\frac{(-j\omega_1)^{n_1}(-j\omega_2)^{n_2}(-j\omega_3)^{n_3}}{||\boldsymbol{\omega}||^{n_1+n_2+n_3}}\hat{f}(\boldsymbol{\omega}), \quad (5.1)$$

for all combinations of (n_1, n_2, n_3) with $n_1 + n_2 + n_3 = N$ and $n_{1,2,3} \in \mathbb{N}$. Equation 5.1 yields to $\binom{N+2}{2}$ templates $\mathcal{R}^{(n_1,n_2,n_3)}$ and forms multiscale steerable filterbanks when coupled with a multi–resolution framework based on isotropic band–limited wavelets (e.g., Simoncelli) [Unser et al., 2009].

In order to define 3D texture features the second–order Riesz filterbank (depicted in figure 5.1) has been used. Rotation–covariance is obtained by locally aligning the Riesz components $\mathcal{R}^{(n_1,n_2,n_3)}$ of all scales based on the local prevailing orientation. This procedure allows the spatial rotation of each component of the filterbank in order to obtain the maximum response with respect to the principal components of texture intensities, in a consistent analysis for all the samples. The regularized structure tensor for aligning the second order Riesz transform filterbanks $\mathcal{R}_{x^2}$, $\mathcal{R}_{y^2}$ and $\mathcal{R}_{z^2}$ is presented in [Chenouard and Unser, 2011].

2^{nd} order 3D Riesz features yield to a 6-dimensional response to a filterbank according to the texture of the tissue volume as depicted in figure 5.2. For a 3D CT image of size $W \times H \times S$, a new volume with the responses to each one of the second–order Riesz kernels, of size $6 \times W \times H \times S$ can be obtained. Nevertheless, for the task of tissue classification, a more compact and accurate representation is desirable, where feature characteristics can be encoded to a specific common format regardless of the size of any given volumetric region.

Covariance-based descriptors can provide a suitable representation for CT region characterization. Any given voxel of a tissue region is defined by its 6-dimensional Riesz feature filter responses. According to values of density, neighbourhood, orientation and intensity, Riesz transforms provide a characteristic pattern along its correspondent tissue region. Due to their construction, covariance-



(a) $G * \mathcal{R}^{(2,0,0)}$ (b) $G * \mathcal{R}^{(0,2,0)}$ (c) $G * \mathcal{R}^{(0,0,2)}$

(d) $G * \mathcal{R}^{(1,1,0)}$ (e) $G * \mathcal{R}^{(1,0,1)}$ (f) $G * \mathcal{R}^{(0,1,1)}$

Figure 5.1: Second–order Riesz kernels $\mathcal{R}^{(n_1,n_2,n_3)}$ convolved with isotropic Gaussian kernels $G(\boldsymbol{x})$. Responses to a linear combination of the filterbank represented by these kernels are used as discriminative representations of the underlying 3D tissue texture.

based descriptors are robust to noisy inputs and lose structural information about the observed features. Therefore, they are suitable for unstructured, abstract texture characterization inside a region, regardless of spatial rigid transformations such as rotation, scale or translations, and for volumes of any given number of voxels.

In order to formally define the 3D Riesz-covariance descriptors, the usual feature selection function of the framework, $\Phi(ct, v)$, is related to the current domain. Then for a given 3D CT image ct and a selected subvolume region v of arbitrary size and shape inside the boundaries of ct, the previously defined Riesz filter responses can be used as features:

$$\Phi(ct, v) = \{\phi_{x,y,z}, \forall x, y, z \in v\}, \tag{5.2}$$

$$\phi_{x,y,z} = \left(\mathcal{R}^{(n_1,n_2,n_3)}_{x,y,z}, \|\mathcal{R}\|_{x,y,z}, ct_{x,y,z}\right). \tag{5.3}$$

These features include the 6 Riesz features at each one of the coordinates in the set, as well as their norm and the CT intensity values in Hounsfield Units. Ac-

cording to the intuition observed that this feature selection is capable of encoding the texture and the tissue nature.

Then, for a given region v of the CT image, the associated covariance descriptor can be obtained as:

$$RieszCov\left(\Phi(ct,v)\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_{x,y,z}-\mu\right)\left(\phi_{x,y,z}-\mu\right)^{T}, \tag{5.4}$$

where μ is the vector mean of the set of vectors $\{\phi_{x,y,z}\}$ within the volumetric neighbourhood made of N samples.

For a better visualization, figure 5.3 provides a schema of a nodule region slice from a CT image, along with the 6-dimensional Riesz-filter responses for each voxel of the slice. Visually, the difference on the different tissue regions yields to different responses of the 2^{nd} order Riesz transform filterbank. Figure 5.4 complements this visualization with the separate response regions according to 3D manually delineated masks (from clinicians groundtruth annotations) for both solid and GGO regions of a given nodule. Using these masks, only the voxels belonging to each region are used for the computation of the respective Riesz-covariance descriptors.

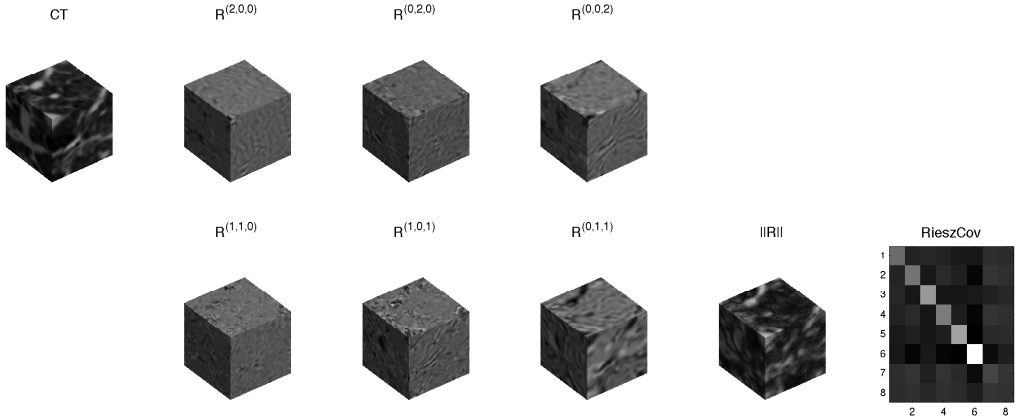


Figure 5.2: Cues involved in the descriptor calculation for a given CT cubic region. The 8 first cubes depict the values within a $40 \times 40 \times 40$ pixel volume, with the CT intensities, 3D-Riesz wavelet responses (for one fixed scale) and Riesz norm features. The 8×8 matrix in the right sub-figure depicts the resulting covariance descriptor, encoding the different correlations between the distributions of the observed cues.

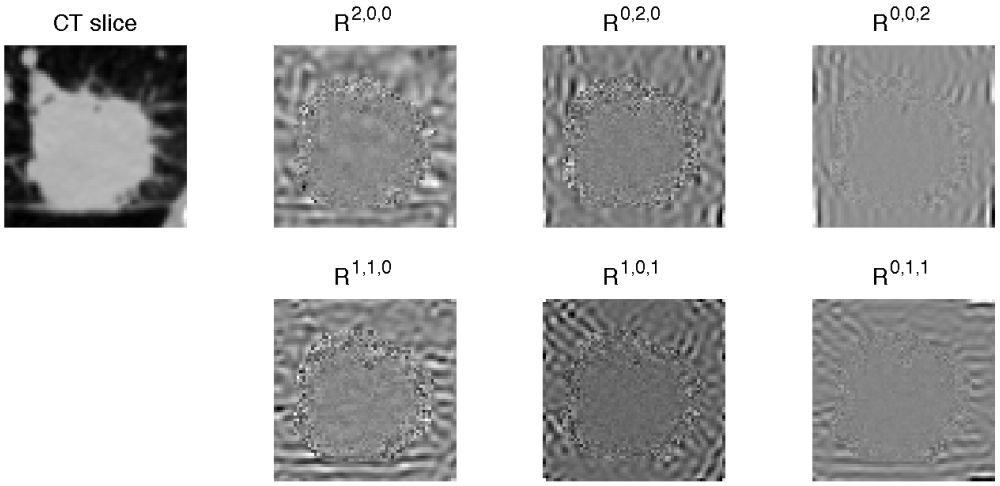
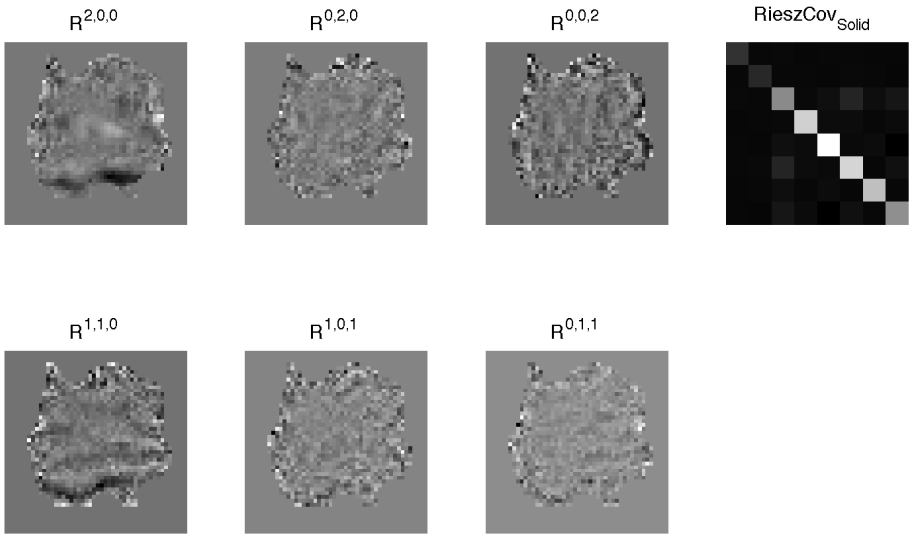


Figure 5.3: Nodule region slice from a CT image, along with the 6-dimensional Riesz-filter responses for each voxel of the slice.

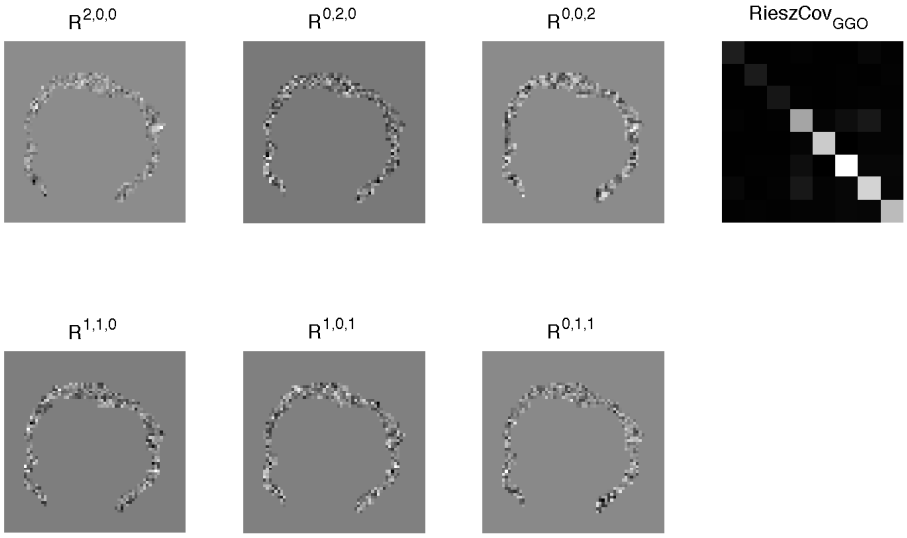
5.2.2 Texture Classification via Bag of Covariances

The "bag of features" paradigm is an established classification technique in the machine learning domain [Lazebnik et al., 2006]. It is conceived as an implicit part-based modeling from a learning set of instances, where collections of different parts of objects can be gathered in order to cover the intra-class variability. Later on, this set of part representations, often referred to as *dictionary*, is used to encode a learning set of instances in terms of frequency histograms of the part repetitions found on these instances. The same representation is done for classification samples and the final decision criteria are made in terms of histogram similarities. This directly suits our classification problem: due to the small number of samples and low resolution of features, we can model a vast dictionary of all the possibilities in tissue types –inner texture and margin of solid and GGO nodule components, vessels, air, blood or even fiducial markers in the healthy lung tissue.

We define the so-called "bag of covariances" in three stages: dictionary learning, modelling of tissue classes by word frequencies, and classification of test regions. In order to build the so-called dictionary, we denote by $P = \{\text{CT}\,{}^{c}_{1:p}\}$ as the set of CT images for all patients p, and their delineated regions for each class c (solid, GGO and healthy lung). From this data, we can obtain the set of vectorized 3D Riesz-covariance descriptors as defined in the previous section: $\hat{x}\,{}^{c}_{v,p} = vect(\log_{I_d}(RieszCov^{C}_{V,P}))$, for a set of learning patients p and classes c. v denotes the set of 3D subvolumes inside the region class for which the descriptors are computed, and it is obtained randomly inside the manually annotated class



(a) Solid region nodule slice.



(b) GGO region nodule slice.

Figure 5.4: Visualization of the response regions according to 3D manually delineated masks for both solid and GGO regions of a given nodule and the corresponding Riesz-covariance descriptors.

regions. See figure 5.5 for a clarification of this learning. All these elements, so-called words, can be stored as a matrix, and data clustering algorithms such as K-means can be applied in order to reduce dimensionality of those over-represented samples.

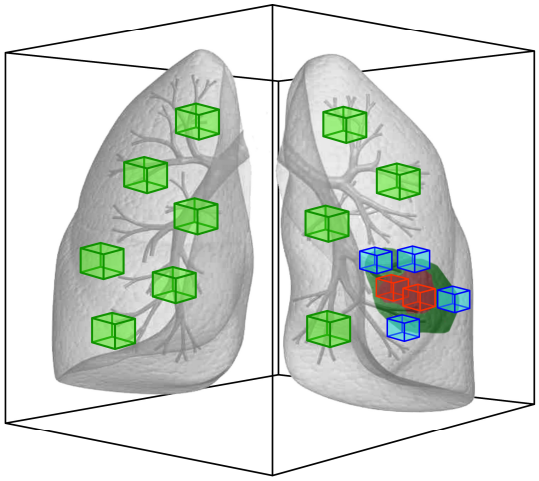


Figure 5.5: Representation of the descriptors at a patient level: cube colors denote the 3D patches used for the construction of the dictionary according to the three tissue classes.

In order to model the different classes, a new set of 3D Riesz-covariance descriptors for different parts of all three classes is again obtained, from a second set of learning patients. The descriptors are mapped to the nearest words in the dictionary via their Euclidean distance, as the parts are projected to the tangent space T_{I_d}. This gives a set of histogram representations in which each one of the tissue instances for all the patients are defined as the frequency of part appearances present in the dictionary.

For the classification of a new sample, a new set of 3D Riesz-covariance descriptors $\hat{y}_{\ v} = vect(\log_{I_d}(RieszCov_{ct,v}))$ is obtained, where v indicates different patches inside the CT image. Again, the descriptors can be quantized in terms of dictionary frequencies, and the final classification criteria are made according to the closest histogram representation in the available model:

$$class(ct) = \underset{i}{\operatorname{argmin}}\, D(h_{ct}, h_i), \tag{5.5}$$

where h_{ct} denotes the histogram representation of the CT test sample, h_i denotes the learned model of dictionary frequency representations, and D is the χ^2 distance used for the comparison of histograms.

5.2.3 Evaluation

100 patients from Stanford Hospital and Clinics with biopsy–proven early stage non-small cell lung carcinoma were used to estimate the performance of our approach. The nodule region present in each patient lungs was delineated in 3D by

the treating radiation oncologist, then the GGO and solid components were contoured separately using lung and mediastinal windows. MATLAB software was used for post-processing of the available CT images and data, including region ground-truth preparation and resampling of volumes in order to have isotropic voxels of $0.8 \times 0.8 \times 0.8$ mm^3 using cubic spline interpolation. 5 patients of the dataset were discarded due to annotation artifacts, therefore the final dataset contains 95 patients.

In order to estimate the classification accuracy achieved by the proposed method, we performed cross-validation over the presented dataset, keeping 35 patients for learning the model and the remaining 60 patients for testing, for 10 iterations. The "Bag of covariances" method was trained by modelling 60 parts for each class, for each patient, therefore creating a dictionary size of 6300 words at each iteration. The classification performance accuracy for the three modelled classes is reported in terms of sensitivity $(TP/TP+FN)$ and specificity $(TN/TN+FP)$, with average values of 82.2% (σ = 2.55%) and 86.2% (σ = 5.85%) respectively, according to the available ground-truth annotations defined by clinicians.

Recent methods as [Song et al., 2013; Depeursinge et al., 2012] reported similar accuracy, which settles our presented approach amongst state of the art performance levels. Even if these methods are focused on interstitial lung diseases rather than nodule separation, their definition of GGO and solid areas is consistent with our approach so comparing against their outcomes is coherent.

5.3 Conclusions

This chapter has intoduced an integrated approach for the characterization and classification of different lung tissue types in 3D, particularly focused to the separation of GGO and solid nodule areas. Despite the high intra-class variability, this method obtained reliable classification results, thanks to a theoretically solid descriptor for encoding feature variations.

The contribution of this work sets the basis for further CT image classification, not only in terms of different tissue classes but also at an intra-class level in order to model temporal stages of a nodule or to learn specific models for characterizing the response to treatment of specific diseases. These techniques may be clinically useful for identifying and characterizing suspicious tissue regions in lesions. One of the handicaps on testing pattern recognition methods with real clinical data is the usual lack of patients, which leads to limited experimental evaluation. Nevertheless, the work presented in this chapter (which derives from a research internship stage at the University of Applied Sciences of Western Switzerland in collaboration with the Department of Radiology and Medicine at Stanford University) has settled the basis for a methodology in ongoing develop-

ment for modelling adenocarcinoma recurrence from texture models. This work uses an extended dataset with complementary patient information provided also by clinicians patient assessment.

On the other hand, clinicians searching through large data sets of multimodal medical information generated in hospitals currently do not fully exploit previous medical cases to retrieve relevant information for a differential diagnose. As identified also after the participation to the ImageCLEF medical image classification challenge, current approaches using textual keyword-based retrieval could be extended by its fusion with texture and visual characterization features. Projects as *Visual Concept Extraction Challenge in Radiology* (VISCERAL [1]) are being developed in order to take advantage of these large data sets and to provide useful information for similar case retrieval and diagnose decisions. VISCERAL in particular provides a cloud–based infrastructure for the evaluation of medical image analysis techniques on large data sets Langs et al. [2013]; Hanbury et al. [2012]. Derived outcomes of the technique presented in this chapter have been submitted to the VISCERAL Multimodal Retrieval in the Medical Domain workshop and obtained results with accuracy levels amongst the top participants. The contribution resides in providing a scoring metric for the comparison of local texture features at particular regions of interest such as specific organs or conflictive areas. An expert clinician has integrated this contribution into a weighted scheme containing anatomical and clinical correlations, as well as RadLex case definition terms (RadLex is a comprehensive lexicon of radiology terms for standardized indexing and retrieval of radiology information resources [Langlotz, 2006]). More details on this contribution can be found in [Jiménez del Toro et al., 2015].

[1] http://www.visceral.eu/

Chapter 6

Conclusions and Future Work

> "Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law."
>
> — DOUGLAS HOFSTADTER, *Gödel, Escher, Bach: An Eternal Golden Braid*

THIS DISSERTATION HAS PRESENTED a common framework for pattern recognition in different feature spaces with a dual contribution. In a first place, one of its values is that the covariance-based formulation translates any desired feature space to a common descriptor manifold. In a second place, this manifold has been explored in different machine learning algorithms for solving particular application problems: part-based modelling, dictionary learning, geometric constraint addition or game theory refinement.

These different methods have been validated by particularly designed experimental set-ups, demonstrating the feasibility of the proposed approaches and providing a comparison against state-of-the-art contributions published in recent years. Despite the enclosed results, having worked with these techniques has opened several continuity lines and improvement possibilities that lay beyond the scope of the current dissertation, but can be commented as future work.

6.1 Summary

The outcomes of this dissertation can be divided in four parts as presented in chapters 2-5. The core of this thesis is twofold, both in the analysis carried on the family of covariance-based descriptor and on its practical side. Besides the applications found on different computer vision tasks, pragmatic details have been discussed: feature extraction and fusion, and machine learning methodology considerations in part-based classification, 3D scene reconstruction via pairwise descriptor matching or sparse dictionary learning. Furthermore, all these tasks have

required computational implementations whose code has been or will be made publicly available.

- In 2D color images, it has been shown how to extract color feature-based descriptors for part-based classes and highly variable images, and how to apply supervised learning techniques as boosting and sparse dictionary learning.

    1. Boosting has been used as an introductory context, implementing one of the pioneer methods of the state of the art on manifold-based learning. The study of the methodology presented in [Tuzel et al., 2008b] was used in order to develop expertise on Riemannian geometry based methods.
    2. Medical image retrieval has been used as a second proof of concept for applying covariance descriptors to complete image regions. The results obtained after the participation to the ImageCLEF image classification challenge placed the presented approach on par with more complex methods.

- The part of 3D area description for scene registration intends to show not only descriptor itself for extended dimensionality applications; it also demonstrates its flexibility for being extended with additional constraints when the descriptor itself is not capable enough of capturing external handicaps, such as scene complexities as repetitive patterns or symmetries.

    1. The so-called 3D MCOV descriptor has been presented and its discriminative performance has been tested under noise and resolution changes. This experimental set-up has been conceived having in mind real scene understanding applications using currently available depth acquisition devices.
    2. Due to real world conditions, such as dense point clouds and repeatability, a game theory based outlier rejection method has been provided for the discard of undesired match candidates. Thanks to its efficiency, this method is much more suitable and computationally efficient than existing iterative methods.
    3. These contributions together have allowed to build an enclosed system for pairwise view registration specially aimed at the reconstruction of 3D scenes with commodity depth acquisition devices. The conducted experiments have shown the benefits of this approach even in highly noisy conditions.

- Gesture recognition for classification of depth image sequences has been tackled in a simplistic formulation via a spatio-temporal covariance description of low-level cues. This enables the classification of motion patterns thanks to the descriptor space, provided that the own descriptors already capture temporal information by a slice-based conception.

    1. The proposed descriptor is suitable for temporally bounded gestures with clear starts and endings, and clear motion execution depicting the needed variability stages along the gestures.
    2. Using balanced datasets of different gestures has provided the opportunity to learn complete dictionaries of gestures.

- In 3D medical imaging, covariance-based descriptors of three-dimensional volumes have provided a natural formulation for encoding tissue characteristics. Low-level tissue features found in 3D Riesz-wavelets were perfectly paired with the unstructured, rotation and translation invariant descriptor formulation in order to provide a part-based bag-of-covariances classification method for the classification of lung nodule tissue in CT images. These descriptors were also used in multimodal case-based retrieval. This work was conducted as part of a research internship at the University of Applied Sciences Western Switzerland (HES-SO Valais) in collaboration with the Department of Radiology and Medicine at Stanford University and is still in an ongoing stage, with the preparation of a publication on adenocarcinoma recurrence modelling from texture information.

6.2 Work Limitations and Future Considerations

The aim of this thesis has been the study of different applications of covariance-based descriptors in order to perform pattern recognition in multiple feature spaces. Along with the research conducted during these years, many research questions have arisen, including seeking for application contexts, design and methodology formulation and experimental set-ups for validation. As presented in the chapters of this dissertation, it has been attempted to provide answers and conclude all the started approaches in order to solve all these questions. However, it is sometimes hard to close a research line when it gives raise to new possibilities and provides new ways to explore. The work introduced in this dissertation is considered to provide a baseline for many extensions: whether it is in new feature selection functions, new machine learning algorithms, or new conceptions of the covariance statistical notion.

In 2D image classification there are many features that could be selected in the future. Colour and gradient visual cues have been used since these are the natural ones in many image processing applications, but novel acquisition devices such as depth map cameras or medical imaging machinery could yield to new feature spaces in the future. The participation to the ImageCLEF medical retrieval challenge also served as a benchmark for considering the use of textual information for its fusion with visual information. Considering text patterns and keyword frequencies associated to given medical image classes could provide a natural feature fusion formulation in the own descriptor level, valuable for case retrieval. This has not been explored to our knowledge, but judging the ImageCLEF challenge results it would be feasible and interesting.

Working with 3D point clouds also left many future work ideas. In order to determine a value for neighbourhood radius, a boundary estimator based on sampling theory has been provided. This is suitable as long as it is consistent in applications as scene reconstruction, as the same criterion is applied in the descriptor computation of any scene view. Nevertheless, for more generalist 3D retrieval applications, this could be improved with geodesic metrics for non-linear neighbourhood selection taking into account the geometry of the point cloud surface. Regarding texture features as color values, RGB colorspace values is used as a simplification. Other color spaces such as CIELab or YUV, or color constancy methods could be explored in order to provide illumination invariance.

In gesture classification, results have been analysed in the controlled environments found in state-of-the-art benchmarking dataset. In these cases, gesture performances are segmented in time and subjects are recorded in ideal conditions regarding camera position, clutter or other external factors. The 4DCov spatio-temporal descriptor is focused on capturing the spatio-temporal variability of depth-map sequence pixels, provided that the used frames belong to bounded gestures –as usual in benchmarking datasets. While this provides homogeneous testing conditions with respect to other existing methodologies, refinement techniques should be used for real applications such as on-line hand gesture translation. In the provided framework, this could be implemented by temporal-modelling methodologies such as manifold alignment. Analogously to a dynamic time warping approach, a gesture could be encoded as a collection of several covariance-based descriptors belonging to the spatial variability in frames. Classifying a gesture could be done by searching for the minimum warping weight of these collection of descriptors with respect to a model of gesture templates. This was in fact a continuation line of research that was left open due to the considerable effort needed on acquiring and annotating a new testing dataset of real, mixed gestures into the wild.

Leveraging the presented framework to medical imaging has been a complementary opportunity to work with different texture features and within a thrilling applied computer vision context: while the work in this area has been clearly focused in a tissue classification application, we have been in direct contact with many other situations that could benefit of this research. Unsupervised tissue segmentation is of real interest in the medical community for computer aided diagnose. It currently requires costly time and human supervision, while a computer based segmentation could provide objective tissue delineation with unbiased texture statistics in order to help clinicians. Immediate continuation lines include an unsupervised 3D segmentation algorithm based on region-growing tissue analysis, and tissue modelling with correlated clinician information, such as patient cancer recurrence and follow-up time, in order to explore the possible correlations between tissue texture observations and clinical outcomes.

In a more conceptual way, covariance matrices can also be explored as something more than descriptors and data signatures. This dissertation has a computer vision research background and therefore has always presented the pattern recognition applied side of the covariance notion for feature definition. Nevertheless, covariance has a broader statistical meaning and could provide a step beyond in causality and prediction models, specially for fMRI or non-stationary signal analysis in medical research as found in recent novel contributions. Furthermore, covariance as a second order moment based descriptor has been explored along this thesis, always with a direct translation to manifold geometry. Along the research with 3D point clouds, a direct translation to higher order moment statistics was also considered, possibly giving birth to Kurtosis or Skewness-based descriptors. At the cost of higher order tensor calculus, intuition leads to believe that these descriptors could encode better shape information in terms of 3D point clouds curvature, constancy and variability. Hopefully, the methods presented along these pages might provide some inspiration for future research.

Bibliography

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Andrea Albarelli, Samuel Rota Bulò, Andrea Torsello, and Marcello Pelillo. Matching as a non-cooperative game. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1319–1326, 2009.

Andrea Albarelli, Emanuele Rodolà, and Andrea Torsello. A game-theoretic approach to fine surface registration without initial motion estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 430–437, 2010.

Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(2):288–303, 2010.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

Alexander C. Berg and Jitendra Malik. Geometric blur for template matching. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–607. IEEE, 2001.

Alexander C. Berg, Tamara L. Berg, and Jitendra Malik. Shape matching and object recognition using low distortion correspondences. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 26–33. IEEE, 2005.

Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 14:239–256, 1992.

Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1395–1402, Oct 2005.

Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(3):257–267, 2001.

Gary R. Bradski and James W. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002.

Matteo Bregonzio, Shaogang Gong, and Tao Xiang. Recognising action as clouds of space-time interest points. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1948–1955, 2009.

Nicola Brusco, Marco Andreetto, Andrea Giorgi, and Guido Maria Cortelazzo. 3D registration by textured spin-images. In *International Conference on 3D Digital Imaging and Modeling, 3DIM*, pages 262–269, 2005.

Pedro Cortez Cargill, Cristobal Undurraga Rius, Domingo Mery Quiroz, and Alvaro Soto. Performance evaluation of the covariance descriptor for target detection. In *Proceedings of the XXVIII International Conference of the Chilean Computer Science Society*, 2009.

Chu-Song Chen, Yi-Ping Hung, and Jen-Bo Cheng. RANSAC-based DARCES: A new approach to fast automatic registration of partially overlapping range images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21:1229–1234, 1999.

Nicolas Chenouard and Michael Unser. 3D steerable wavelets and monogenic analysis for bioimaging. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 2132–2135, April 2011.

Anoop Cherian and Suvrit Sra. Riemannian sparse coding for positive definite matrices. In *European Conference on Computer Vision (ECCV)*, pages 299–314. Springer, 2014.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Efficient similarity search for covariance matrices via the Jensen-Bregman

LogDet divergence. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2399–2406. IEEE, 2011.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Jensen-Bregman LogDet divergence with application to efficient similarity search for covariance matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(9):2161–2174, 2013.

Chin Seng Chua and Ray Jarvis. Point signatures: A new representation for 3D object recognition. *International Journal of Computer Vision (IJCV)*, 25:63–85, 1997.

Ondrej Chum and Jiri Matas. Matching with PROSAC-progressive sample consensus. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 220–226, 2005.

Ondrej Chum and Jiri Matas. Optimal randomized RANSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30:1472–1482, 2008.

Ondrej Chum, Jiri Matas, and Stepan Obdrzalek. Enhancing RANSAC by generalized model optimization. In *Asian Conference on Computer Vision (ACCV)*, volume 2, pages 812–817, 2004.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning (Springer)*, 20:273–297, 1995.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

James W. Davis. Hierarchical motion history images for recognizing human motion. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46, 2001.

Adrien Depeursinge, Dimitri Van De Ville, Alexandra Platon, Antoine Geissbuhler, Pierre-Alexandre Poletti, and Henning Müller. Near–affine–invariant texture learning for lung tissue analysis using isotropic wavelet frames. *IEEE Transactions on Information Technology in BioMedicine*, 16(4):665–675, July 2012.

Adrien Depeursinge, Antonio Foncubierta-Rodríguez, Dimitri Van De Ville, and Henning Müller. Three–dimensional solid texture analysis and retrieval in biomedical imaging: review and opportunities. *Medical Image Analysis*, 18 (1):176–196, 2014.

Adrien Depeursinge, Masahiro Yanagawa, Ann N. Leung, and Daniel L. Rubin. Predicting adenocarcinoma recurrence using computational texture models of nodule components in lung CT. *Medical Physics*, 2015.

Ahmed Elgammal and Chan-Su Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–681. IEEE, 2004.

Duc Fehr, Anoop Cherian, Ravishankar Sivalingam, Sam Nickolay, Vassilios Morellas, and Nikolaos Papanikolopoulos. Compact covariance descriptors in 3D point clouds for object recognition. In *International Conference on Robotics and Automation (ICRA)*, pages 1793–1798, 2012.

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, 2010.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.

Alex Flint, Anthony R. Dick, and Anton Van Den Hengel. Thrift: Local 3D structure recognition. In *Digital Image Computing Techniques and Applications*, volume 7, pages 182–188, 2007.

Wolfgang Förstner and Boudewijn Moonen. A metric for covariance matrices. *Quo vadis Geodesia*, pages 113–128, 1999.

Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *ACM Conference on Human Factors in Computer Systems*, pages 1737–1746, 2012.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing objects in range data using regional point descriptors. In *European Conference on Computer Vision (ECCV)*, volume 3023, pages 224–237, 2004.

Jaime A. García, Karla Felix Navarro, Daniel Schoene, Stuart T. Smith, and Yusuf Pisan. Exergames for the elderly: Towards an embedded kinect-based clinical test of falls risk. *Studies in Health Technology and Informatics*, 178:51–7, 2012.

Alba García Seco de Herrera, Henning Müller, and Stefano Bromuri. Overview of the ImageCLEF 2015 medical classification task. In *Working Notes of CLEF 2015 (Cross Language Evaluation Forum)*, CEUR Workshop Proceedings. CEUR-WS.org, September 2015.

Kathrin Gerling, Ian Livingston, Lennart Nacke, and Regan Mandryk. Full-body motion-based game interaction for older adults. In *ACM Conference on Human Factors in Computer Systems*, pages 1873–1882, 2012.

Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition using sparse representation on covariance manifolds of optical flow. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 188–195, 2010.

Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *International conference on Machine learning*, pages 376–383. ACM, 2008.

Lei Han, Xinxiao Wu, Wei Liang, Guangming Hou, and Yunde Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 28(5):836–849, 2010.

Allan Hanbury, Henning Müller, Georg Langs, Marc André Weber, Bjoern H. Menze, and Tomas Salas Fernandez. Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 24–29. Springer, 2012.

Mehrtash T. Harandi, Conrad Sanderson, Richard Hartley, and Brian C. Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *European Conference on Computer Vision (ECCV)*, pages 216–229. Springer, 2012.

Mehrtash T. Harandi, Mathieu Salzmann, and Richard Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices. In *European Conference on Computer Vision (ECCV)*, pages 17–32. Springer, 2014a.

Mehrtash T. Harandi, Mathieu Salzmann, and Fatih Porikli. Bregman divergences for infinite dimensional covariance matrices. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1010. IEEE, 2014b.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

Simon Haykin. *Neural Networks: A comprehensive foundation*. Prentice Hall, 1998.

Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Optical Society of America*, 4:629–642, 1987.

Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning Euclidean-to-Riemannian metric for point-to-set classification. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1677–1684. IEEE, 2014.

Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *Interational Conference on Artificial Intelligence*, pages 2466–2472, 2013.

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–80. IEEE, 2013.

Oscar Alfonso Jiménez del Toro, Pol Cirujeda, Yashin Dicente Cid, and Henning Müller. Radlex terms and local texture features for multimodal medical case retrieval. In *Multimodal Retrieval in the Medical Domain (MRMD) 2015*, volume 9059 of *Lecture Notes in Computer Science*. Springer, April 2015.

Andrew Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, August 1997.

Andrew Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21:433–449, 1999.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30:509–541, 1977.

Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Human activity recognition using a dynamic texture based method. In *British Machine Vision Conference (BMVC)*, pages 1–10, 2008.

Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference (BMVC)*, 2008.

Stefan Kluckner, Thomas Mauthner, and Horst Bischof. A covariance approximation on Euclidean space for visual tracking. In *Annual Workshop of the Austrian Association for Pattern Recognition*, 2009.

Artiom Kovnatsky, Michael M. Bronstein, Alexander M. Bronstein, and Ron Kimmel. Photometric heat kernel signatures. In *Scale Space and Variational Methods in Computer Vision*, pages 616–627, 2012.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Alexey Kurakin, Zhengyou Zhang, and Zicheng Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *European Signal Processing Conference (EUSIPCO)*, pages 1975–1979, 2012.

Junghyun Kwon and Frank Chongwoo Park. Visual tracking via particle filtering on the affine group. *The International Journal of Robotics Research*, 2009.

Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *International Conference on Robotics and Automation (ICRA)*, pages 1817–1824, 2011.

Belinda Lange, Sebastian Koenig, Eric McConnell, C. Chang, Rick Juang, Evan Suma, Mark Bolas, and Albert Rizzo. Interactive game-based rehabilitation using the Microsoft Kinect. In *Virtual Reality Short Papers and Posters*, pages 171–172, 2012.

Curtis P. Langlotz. RadLex: A new method for indexing online educational materials. *Radiographics*, 26(6):1595–1597, 2006.

Georg Langs, Allan Hanbury, Bjoern Menze, and Henning Müller. Visceral: Towards large data in medical imaging-challenges and directions. In *Medical Content-Based Retrieval for Clinical Decision Support*, pages 92–98. Springer, 2013.

Ivan Laptev. On space-time interest points. *International Journal of Computer Vision (IJCV)*, 64(2-3):107–123, 2005.

Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

John M. Lee. *Smooth manifolds*. Springer, 2003.

Ruonan Li, Pavan Turaga, Anuj Srivastava, and Rama Chellappa. Differential geometric representations and algorithms for some pattern recognition and computer vision problems. *Pattern Recognition Letters*, 43:3–16, 2014.

Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3D points. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9–14, 2010.

Li Liu and Ling Shao. Learning discriminative representations from RGB-D video data. In *Interational Conference on Artificial Intelligence*, pages 1493–1500, 2013.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, 2004.

David P. Luebke. A developer's survey of polygonal simplification algorithms. In *Computer Graphics and Applications*, volume 21, pages 24–35, 2001.

Yui Man Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6):380–388, 2012.

James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. In *International Conference on Machine Learning*, pages 1033–1040, 2011.

Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26:735–747, 2005.

Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

Stefan Munder and Dariu M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(11):1863–1868, 2006.

Farhood Negin, Fırat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In *Image Analysis and Recognition*, pages 648–657. Springer, 2013.

Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24, 2002.

Omar Oreifej and Zicheng Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 716–723, 2013.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision (IJCV)*, 66:41–66, 2006.

Nikolaos Pitelis, Craig Russell, and Lourdes Agapito. Learning a manifold as an atlas. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1642–1649, June 2013.

Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

Fatih Porikli and Tekin Kocak. Robust license plate detection using covariance descriptor in a neural network framework. In *International Conference on Video and Signal Based Surveillance*, pages 107–107, 2006.

Fatih Porikli, Oncel Tuzel, and Peter Meer. Covariance tracking using model update based on Lie algebra. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2006.

Emanuele Rodolà, Andrea Albarelli, Filippo Bergamasco, and Andrea Torsello. A scale independent selection process for 3D object recognition in cluttered scenes. *International Journal of Computer Vision (IJCV)*, 102(1-3):129–145, 2013.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *International Conference on Robotics and Automation (ICRA)*, pages 3212–3217, 2009.

Andres Sanin, Conrad Sanderson, Mehrtrash Harandi, and Brian Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *IEEE Workshop on Applications of Computer Vision*, pages 103–110, 2013.

Robert E. Schapire. The boosting approach to machine learning: An overview. In *Lecture Notes in Statistics*, pages 149–172. Springer (New York), 2003.

Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia Conference*, pages 357–360, 2007.

Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1): 116–124, 2013.

Ravishankar Sivalingam, Vassilios Morellas, Daniel Boley, and Nikolaos Papanikolopoulos. Metric learning for semi-supervised clustering of region covariance descriptors. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–8. IEEE, 2009.

Yang Song, Weidong Cai, Yun Zhou, and David D. Feng. Feature–based image patch approximation for lung tissue classification. *IEEE Transactions on Medical Imaging*, PP(99), 2013.

Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Christian Thurau and Václav Hlavác. Pose primitive based human action recognition in videos or still images. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *European Conference on Computer Vision (ECCV)*, volume 6313, pages 356–369, 2010.

Federico Tombari, Samuele Salti, and Luigi Di Stefano. A combined texture-shape descriptor for enhanced 3D feature matching. In *IEEE International Conference on Image Processing*, pages 809–812, 2011.

Andrea Torsello, Emanuele Rodolà, and Andrea Albarelli. Sampling relevant points for surface registration. In *Conference on 3D Imaging, Modeling, Processing, Visualization and Transmssion (3DIMPVT)*, pages 290–295, 2011.

Pavan Turaga, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technologies*, 18(11):1473–1488, 2008.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. *European Conference on Computer Vision (ECCV)*, pages 589–600, 2006.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Human detection via classification on riemannian manifolds. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Learning on lie groups for invariant detection and tracking. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008a.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(10):1713–1727, 2008b.

Michael Unser, Daniel Sage, and Dimitri Van De Ville. Multiresolution monogenic signal analysis using the riesz–laplace wavelet transform. *Image Processing, IEEE Transactions on*, 18(11):2402–2418, 2009.

Mauricio Villegas, Henning Müller, Andrew Gilbert, Luca Piras, Josiah Wang, Krystian Mikolajczyk, Alba García Seco de Herrera, Stefano Bromuri, M. Ashraful Amin, Mahmood Kazi Mohammed, Burak Acar, Suzan Uskudarli, Neda B. Marvasti, José F. Aldana, and María del Mar Roldán García. General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science. Springer International Publishing, 2015.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3D action recognition with random occupancy patterns. In *European Conference on Computer Vision (ECCV)*, pages 872–885, 2012a.

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, 2012b.

Samuel S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24:471–494, 1932.

John Wright, Allen Y. Yang, Arvind Ganesh, Shankar S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(2):210–227, 2009.

John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S. Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proc. of the IEEE*, 98(6):1031–1044, 2010.

Lu Xia, Chia-Chih Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3D joints. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20–27, 2012.

Liang Xiong, Fei Wang, and Changshui Zhang. Semi-definite manifold alignment. In *European Conference on Machine Learning*, pages 773–781. Springer, 2007.

Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(12):2878–2890, 2013.

Jian Yao and Jean-Marc Odobez. Fast human detection from videos using covariance features. In *European Conference on Computer Vision (workshop on Visual Surveillance, ECCV-VS)*, 2008.

Tomoyuki Yokose, Kenji Suzuki, Kanji Nagai, Yutaka Nishiwaki, Satoshi Sasaki, and Atsushi Ochiai. Favorable and unfavorable morphological prognostic factors in peripheral adenocarcinoma of the lung 3 cm or less in diameter. *Lung Cancer*, 29(3):179–188, 2000.

Andrei Zaharescu, Edmond Boyer, Kiran Varanasi, and Radu Horaud. Surface feature detection and description with applications to mesh matching. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 373–380, 2009.

Andrei Zaharescu, Edmond Boyer, and Radu Horaud. Keypoints and local descriptors of scalar functions on 2D manifolds. In *International Journal of Computer Vision (IJCV)*, volume 100, pages 78–98, 2012.

Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision (ICCV)*, pages 471–478, 2011.

Barcelona, July 2015