pacman

Probabilistic and Compositional Representations for Object Manipulation

FP7-IST-60918

1 March 2013 (36months)

DR1.2:
Learned 3D Visual Vocabularies of Visual Object Shape

Krzysztof Walas, Mete Ozay, Mirela Popa, U. Rusen Aktas, Vladislav Kramarev, Justus Piater, Aleš Leonardis, Jeremy L. Wyatt

School of Computer Science, University of Birmingham, United Kingdom.

⟨k.walas@cs.bham.ac.uk⟩

Due date of deliverable:	2015-02-28
Actual submission date:	2015-02-28
Lead partner:	BHAM
Revision:	final
Dissemination level:	PU

This report describes the algorithms proposed regarding Deliverable D1.2.

Executive Summary

This report presents work carried out in WP1 on Compositional Hierarchies of object categories observed from 2D and 3D images. The work addresses Tasks 1.1, 1.2 and 1.3. We describe an approach to learning and modeling the relationship between 2D shape deformations of compositional hierarchical models and geometric transformations employed on object shapes for shape recognition as defined in Task 1.1. The work led to two publications; *i*) a conference publication submitted to IEEE CVPR 2015 [9] (see Annex 2.1 for a Technical Report version of the conference paper), and *ii*) a journal publication which will be submitted to IEEE TPAMI, 2015 [10] (see Annex 2.2 for a Technical Report version of the journal paper).

Regarding Task 1.2, a hierarchical compositional architecture is described which learns a vocabulary to capture 3D structural elements of the objects, where depth disparities constitute the first layer of the hierarchy. The work has been exploited to perform discrimination and this aspect is reported in a conference publication submitted to SCIA 2015 [4] (see Annex 2.3).

In Task 1.3, we have proposed methods for statistical modeling of the relationship between geometric properties of shape representations obtained from 2D and 3D compositional hierarchies which have been developed in Task 1.1 and 1.2. The proposed probabilistic models and experimental analyses have been provided in a technical report [12] (see Annex 2.6).

The theoretical framework proposed in Task 1.1 has been used for the analysis and modeling of geometric properties of 3D models in Task 1.2. In addition, the approach proposed in Task 1.1 will be used for modeling and incorporation of the geometric relationships into the constrained optimization problems of the statistical learning methods which have been developed for the integration of 2D and 3D models in Task 1.3.

Role of Learned 3D Visual Vocabularies of Visual Object Shape in PaCMan

In WP1, we focus on learning 2D compositional hierarchical models using 2D images of shapes which are deformed under different geometric transformations (Task 1.1), on learning a 3D compositional shape vocabulary (Task 1.2), and on the integration of multi-modal information in Task 1.3 and WP2. Incremental learning methods are considered in the proposed 2D and 3D compositional hierarchical algorithms which will be used to process actively acquired data to support WP3. The proposed algorithms will be utilized in the grasping and dishwasher-scenario tasks which are described in WP4 and WP5.

The proposed algorithms will be used to obtain compositional models of shape representations which will be employed for the development of algorithms for grasping and scene understanding using multi-modal data. First, the proposed methods and algorithms support Task 1.3 for the development of algorithms in order to integrate 2D and 3D representations of shapes of objects. Second, the 3D compositional models are able to be integrated with haptic properties of the models (shape and friction) in Task 2.1. Third, learned 3D visual vocabularies will be utilized for the improvement of grasping capabilities of a robotic system by providing compositional models of shapes in order to model the relationship between parts and compositions of shapes, and contact points (Task 2.2). Finally, 3D compositional models will be used for the development of algorithms for scene analysis and understanding (Task 2.4) and in consequence they may support grasp planning algorithms (Task 4.4).

Contribution to the PaCMan scenario

The algorithms developed for modeling hierarchical compositional representations of 2D and 3D shapes in WP1 will be used for modeling compositional representations of objects and scenes using multi-modal data (WP2) which have been collected using active data acquisition methods (WP3). Then, the hierarchical compositional representations will be utilized for the development of grasping and scene understanding algorithms that will be employed in the dishwasher-scenarios addressed in WP4 and WP5. More precisely, the algorithms developed for inference of missing and occluded parts of 2D and 3D shapes will be used for analysis of cluttered scenes using multi-modal data, and for modeling graspable object parts and grasp selection (WP2, WP3, WP4 and WP5).

1 Tasks, Objectives, Results

1.1 Planned Work

DR 1.2 is supposed to address compositional hierarchies of object categories observed from 2D and 3D images. Planned work mainly concerns Task 1.1 regarding modeling relationships between deformations of shape models and geometric transformations of shapes of objects, Task 1.2 regarding Compositional 3D models of objects, and Task 1.3 regarding integration of 2D and 3D representations.

The objective of Task 1.1 is the analysis and exploration of the relationship between deformations of shape models and geometric transformations employed on the shapes of objects using 2D images observed from different view points. Theoretical results obtained from the analysis were to be used for the development of methods for statistical learning of the mappings between transformations of the shapes and the shape deformations of the models in compositional hierarchies.

In Task 1.2, it was planned to develop a hierarchical compositional representation that captures 3D structural elements of the objects and employs a 3D shape vocabulary. The proposed algorithms should learn subsequent layers of 3D structures by progressing in complexity from absolute depth, to relative depth, planar surfaces of various slants and tilts, and curved surfaces of various degrees of convexity and concavity.

In Task 1.3, it was planned to develop and implement algorithms for modeling the statistical and geometric relationship between 2D and 3D representations obtained from the compositional hierarchies described in Task 1.1 and 1.2.

1.2 Actual Work Performed

In this section, the main achievements related to the topic of this deliverable are briefly described. For detailed descriptions of the work performed the reader is referred to the papers attached in the annex to this deliverable.

1.2.1 Task 1.1 Multi-view Learning of Compositional 2D Models of Object Appearance

We have analyzed the relationship between deformations of hierarchical compositional shape models and geometric transformations employed on shapes using 2D images. In our IEEE CVPR [9] (see Annex 2.1) and IEEE TPAMI [10] (see Annex 2.2) submissions, we addressed a problem of modeling representations of shapes with multiple-disconnected components when i) data and shape spaces are not manifolds and/or ii) the data and shape manifolds contain multiple disconnected components. We have analyzed and designed models that represent shape spaces with four different invariance properties, namely invariance to Euclidean Similarity Transformations (EST), Reflection(R), EST-R and Affine Transformation (AT).

At the first layer $n = 1$ of a compositional hierarchy, a non-manifold data space is obtained using measurements on objects (i.e. 2D images) in scenes where occluded objects, objects with multi-component structures and multiple objects are observed. We model each part as a manifold represented by a Gabor Feature (GF), and a set of centered and normalized GFs represents a preshape space which is not a shape space due degeneracy and singularity occurring in the space. Degenerate points exist if different measurements are used to represent the same shape in the shape space. Singularity is observed if there are undefined points in the shape space, in other words, if a measurement on an object in the data space is not modeled in the shape space. In order to obtain a manifold (i.e., Projective Space) at the next layer, we remove degeneracy by using inhibition, and singularity by using group actions of general linear, symmetry, orthogonal and special orthogonal groups. Therefore, we resolved problem i) by reformulating the relationship between data and shape spaces considering the interaction between Receptive Fields and shape models in a shape vocabulary in our IEEE CVPR paper [9]. More precisely, a Receptive Field (RF) has been defined as codomain of a shape manifold (space) represented in the shape model but not domain as studied by the state-of-the-art methods, such as Density Representation methods [2, 11, 14]. Next, we developed a method to design RFs considering both the topological structure of the models and statistical properties of the data. In our IEEE TPAMI paper [10], we learned the relationship between geometric transformations employed on shapes in data spaces and shape models in shape spaces by statistical learning and inference of shape manifolds defined by RFs.

If singularity and degeneracy occur in a shape space at $n > 1$, then the distance between parts and compositions cannot be computed, and hierarchical models may not be formed appropriately with respect to the invariant properties of the shape space. Manifold structure is required for learning statistical relationships between parts using distance functions for the analysis and formation of different shape spaces satisfying the invariant properties. In other words, singularity and degeneracy problems may propagate through the hierarchy, and may inhibit us from the formation of the manifolds leading to undetected or undefined parts and compositions, e.g. mugs and pans without handles, or with multiple handles, at different layers of hierarchical models. Therefore, in a hierarchical model, we need to assure that a space consisting of a recursive composition of parts constructed at the next layer is a manifold, $\forall n > 1$. For this purpose, problem ii) is considered by designing an indexing mechanism for modeling the formation of shape shapes (manifolds) in the vocabulary. The indexing mechanism is used in order to model compositional relationships between shape spaces and preserve their manifold structure among layers.

In addition, we have been working on the development of a *Descriptive-Discriminative-Generative learning* (D^2G) approach for localization, detection and recognition of occluded objects in a cluttered scene using multi-view 2D images extending our Compositional Hierarchy of Parts (CHOP) algorithm [1]. In the D^2G approach, a hierarchical compositional vocabulary of the CHOP is jointly optimized for recognition and reconstruction of shapes of the objects. In the inference procedure, D^2G first estimates the likelihood for missing parts of the shape of an object using the prior knowledge modeled in the hierarchical compositional vocabulary. Then, pose and category of the object are estimated by updating the hypotheses of the relationship between parts and compositions of the shape which are modeled in shape spaces (i.e., manifolds) of the vocabulary using the likelihood information extracted from the measurements residing in data spaces.

The methods and algorithms proposed in our IEEE CVPR [9] and IEEE TPAMI [10] papers have been implemented by extending our Compositional Hierarchy of Parts (CHOP) algorithm which was introduced in our ECCV paper [1]. In addition, a dataset of multiple view images of 60 different IKEA kitchen objects have been prepared using the robot head and turntable setup designed in the BHAM Intelligent Robotics Lab. The dataset consists of 4200 images captured at 72 different view points, $\theta = 0°, 5°, \ldots, 360°$, which belong to 7 categories, namely Bowl (15 objects), Box (9 objects), Can (4 objects), Mug (15 objects), Pan (5 objects), Plate (4 objects) and Utensil (8 objects). The code of CHOP and the dataset have been published in the PaCMan GitHub repository and in a public repository which are available on "http://rusen.github.io/CHOP/".

1.2.2 Task 1.2 Compositional 3D Models of Objects

In Task 1.2, we have worked on modeling 3D shapes for shape reconstruction, localization and categorization. Following the theoretical principles of hierarchical compositional models of 2D shape representations developed in Task 1.1 [1, 9, 10], we proposed a compositional hierarchy for representation of view-based 3D shape models in our paper submitted to SCIA [4] (see Annex 2.3). At the first layer $n = 1$ of the proposed compositional hierarchy, depth features were extracted from 3D data. Then, we developed part selection and combination methods for construction of 3D compositions at each subsequent layer, $n > 1$, of the hierarchy.

Two part selection methods were proposed in our SCIA [4] submission. First, we employed a Minimum Description Length (MDL) principle in order to construct compositions by maximizing reconstruction performance of the proposed view-based 3D compositional hierarchy. Second, an entropy based part selection method was proposed in order to select parts by maximizing categorization performance of the proposed hierarchy. The code of the proposed methods have been published in the PaCMan GitHub repository.

Subsequently, we studied properties of the geometry of the surfaces. Our investigation is described in the technical report [13] (see Annex 2.5). The concepts described in the report allows us to improve the hierarchy by adding geometry based part selection criteria. Moreover, we are now able to bind the progression in abstraction between hierarchy layers with the changes in the 3D scale space.

In addition, we have organized one of the tracks in SHREC 2015 - 3D Shape Retrieval Contest[1]. The objective of this track is to propose a new shape retrieval scenario in order to examine state-of-the-art algorithms for detection and recognition shapes of occluded objects in cluttered scenes. In the contest, we prepared a dataset consisting of RGB-D images of 30 objects captured at three different view points. A new procedure was proposed for the evaluation of the performances of the algorithms submitted by participants to the contest (see Annex 2.4 for a Technical Report).

1.2.3 Task 1.3 Integration of 2D and 3D Representations of Objects

The research performed in Task 1.3 was focused on the integration of 2D and 3D representations provided by the work performed in Tasks 1.1 and 1.2. Furthermore, we have designed and implemented methods for binding the representations based on computing the conditional probabilities of parts obtained from a 2D compositional hierarchy (Task 1.1) *given* the parts obtained from a 3D compositional hierarchy (Task 1.2), and vice-versa.

The integration algorithm was designed by first defining a spatial neighborhood of parts and compositions obtained from 2D and 3D compositional hierarchies in a Euclidean space defined by image coordinates. Next, we modeled the statistical relationship between the parts and compositions by computing the conditional entropy of distribution of 2D parts which belong to the neighborhood of a given 3D part, and vice-versa. The proposed methods are provided in a technical report [12] (see Annex 2.6).

We examined the proposed methods using our Ikea dataset [2] consisting of both RGB and RGB-D data, which has been also used in WP5. In the analyses, we have observed that the parts and compositions obtained from different hierarchies are not spatially adjacent in a Euclidean space defined by image coordinates (see Annex 2.6). Therefore, geometric properties of 2D and 3D hierarchical compositional representations, such as curvatures of curved 2D shapes and 3D surfaces, should be considered in the development of the probabilistic model.

For this purpose, we propose two approaches as future work. First, we will design different feature spaces, such as the curvature scale space and

[1]http://www.cs.bham.ac.uk/ walask/SHREC2015/

[2]https://github.com/mirelapopa/IkeaObjectDataset

B-spline curves [7, 8], in order to examine the geometric properties of compositional representations following the theoretical principles of hierarchical compositional representations analyzed and introduced in Task 1.1. Second, we will develop quantitative measures to guide the formation of a 3D compositional hierarchy proposed in Task 1.2 by considering the geometric and discriminative properties of representations in order to model the geometric relationship between 2D and 3D representations, and boost the categorization performance of the algorithms using statistical learning methods.

1.3 Relation to the State-of-the-art

In IEEE CVPR [9] and IEEE TPAMI [10] papers (see Annex 2.1 and 2.2), we first address and resolve the challenges of the state-of-the-art manifold based 2D shape representation and classification algorithms, such as Density Representations (DR) [2, 11, 14]. DR model each shape using a distribution as a point on a manifold whose domain is defined by a Receptive Field unlike the proposed definition. Therefore, topological relationships between parts and compositions are not modelled by DR. Since correspondence between shapes may not exist or multiple degenerate correspondences may exist due to the lack of information on the topology, DR may not perform shape matching. Moreover, if a manifold structure of a shape space is assured by employing our proposed indexing mechanism in a compositional hierarchy, the DR can be used in hierarchical models for statistical learning of relationships between parts and compositions using parametric models such as Gaussian Mixture Models [11].

In addition, results on benchmark shape classification outperform state-of-the-art i) shape representation methods, such as Contour Segments & Skeleton Paths, Inner-distance Shape Context, Contour Shapes with Class Segment Sets, ii) shape classification algorithms such as Support Vector Machines (SVM) with Integrated Shape Descriptors, SVM with Procrustes Kernel, SVM with the projection Gaussian kernel, Multiple Kernel Learning with the projection Gaussian kernel, Shape Manifold Classifiers, and iii) hierarchical learning methods such as Denoising, Contractive and marginalized Denoising Auto-enconders, and Deep Belief Networks.

The proposed 3D compositional hierarchy was tested in a classification task using publicly available Washington RGB-D dataset [6] and SHREC'07 dataset [3] (see Annex 2.3). We obtained state-of-the-art categorization performance and a significant improvement over our previous results [5].

References

[1] U. R. Aktas, M. Ozay, A. Leonardis, and J. Wyatt, "A Graph Theoretic Approach for Object Shape Representation in Compositional Hierarchies using a Hybrid Generative-Descriptive Model," in *Proc. of*

European Conf. on Computer Vision (ECCV). Springer International Publishing, 2014, vol. 8693, pp. 566–581.

[2] D. Bryner, E. Klassen, H. Le, and A. Srivastava, "2D Affine and Projective Shape Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 998–1011, May 2014.

[3] D. Giorgi, S. Biasotti, and L. Paraboschi, "SHape REtrieval Contest 2007: Watertight Models Track."

[4] V. Kramarev, K. Walas, and A. Leonardis, "Categorisation of 3D Objects in Range Images Using Compositional Hierarchies of Parts Based on MDL and Entropy Selection Criteria," in *Technical Report, an extended version was submitted to Scandinavian Conference of Image Analysis (SCIA)*, 2015.

[5] V. Kramarev, S. Zurek, J. L. Wyatt, and A. Leonardis, "Object Categorization from Range Images using a Hierarchical Compositional Representation," in *Proc. of the 22nd Int. Conf. on Pattern Recognition (ICPR)*, 2014.

[6] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multiview RGB-D object dataset," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011, pp. 1817–1824.

[7] N. Laiche and S. Larabi, "2D Shape Matching based on B-spline Curves and Dynamic Programming," in *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 1, Lisbon, Portugal, 5-8 January, 2014*, 2014, pp. 484–491.

[8] P. Mongkolnam, C. Nukoolkit, and T. Dechsakulthorn, "Represent Image Contents Using Curves and Chain Code," in *Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA 2007), May 16-18, 2007, Tokyo, Japan*, 2007, pp. 355–358.

[9] M. Ozay, U. R. Aktas, J. Wyatt, and A. Leonardis, "Compositional Formation of 2D Shape Models for Multi-component Shape Representation," in *Technical Report, an extended version was submitted to IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] ——, "Statistical Learning of Geometric Transformations for Modeling Multi-component Shape Representations," *Technical Report, an extended version will be submitted to IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

[11] A. Peter and A. Rangarajan, "Information Geometry for Landmark Shape Analysis: Unifying Shape Representation and Deformation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 337–350, Feb 2009.

[12] M. Popa, M. Ozay, V. Kramarev, R. Aktas, K. Walas, J. Piater, A. Leonardis, and J. Wyatt, "Statistical Learning of Multi-modal Representations of Objects using Multi-view 2D and View-based 3D Images," IIS Group, Department of Computer Science, University of Innsbruck, Tech. Rep., 2015.

[13] K. Walas, M. Ozay, and J. L. Wyatt, "Analysis of 3D Surface Properties for Modeling 3D Compositional Representations," School of Computer Science, University of Birmingham, Tech. Rep., 2015.

[14] Q. Xie, I. Jermyn, S. Kurtek, and A. Srivastava, "Numerical Inversion of SRNFs for Efficient Elastic Shape Analysis of Star-Shaped Objects," in *Proc. of European Conf. on Computer Vision – ECCV 2014*. Springer International Publishing, 2014, vol. 8693, pp. 485–499.

2 Annexes

2.1 Compositional Formation of 2D Shape Models for Multi-component Shape Representation

Bibliography Mete Ozay, U. Rusen Aktas, Jeremy L. Wyatt and Aleš Leonardis Compositional Formation of 2D Shape Models for Multi-component Shape Representation, submitted to IEEE Computer Vision and Pattern Recognition (CVPR), 2015.

Abstract We address the problem of design and construction of shape representation models on manifolds when i) data and shape spaces are not manifolds and/or ii) the data and shape manifolds contain multiple disconnected components. We resolve problem i) by reformulating the relationship between data and shape spaces considering the interaction between Receptive Fields (RFs) and shape models in a shape vocabulary. Then, we suggest a method to design RFs considering both the topological structure of the models and statistical properties of the data. Problem ii) is considered by designing a disjoint union topology using an indexing mechanism for the formation of shape models on shape manifolds in the vocabulary, recursively. To this end we introduce a framework to implement the indexing mechanisms for i) the employment of the vocabulary for structural shape classification, and ii) designing inhibition methods to remove redundant measurements avoiding degenerate models in the vocabulary. The proposed approach is used to construct different shape representations with different invariance properties such as invariance to Euclidean Similarity Transformations (EST) and reflection. Results on benchmark shape classification outperform state-of-the-art shape classification methods.

Relation to WP The paper addresses Task 1.1 by introducing a mathematical framework for representation of occluded and deformed shapes of objects in a compositional hierarchical architecture.

2.2 Statistical Learning of Geometric Transformations for Modeling Multi-component Shape Representations

Bibliography Mete Ozay, U. Rusen Aktas, Jeremy L. Wyatt and Aleš Leonardis, Statistical Learning of Geometric Transformations for Modeling Multi-component Shape Representations, Technical report, in preparation for submission to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

Abstract We introduce a theoretical framework for statistical learning of geometric transformations for modeling multi-component shape representations in compositional hierarchies. For this purpose, we address the problem of analysis of shapes of objects with multiple components *i*) in non-manifold measurement spaces at the first $n = 1$ layer of a hierarchy, and *ii*) in some well-known shape spaces which are not manifolds at some set of points that represent parts and their compositions constructed at the higher layers, $n > 1$, of the hierarchy.

The first problem is considered by reformulating the relationship between data and shape spaces considering the interaction between Receptive Fields (RFs) and shape models in a shape vocabulary. Then, we suggest a method to design RFs considering both the topological structure of the models and statistical properties of the data. Problem *ii*) is resolved by designing a disjoint union topology using an indexing mechanism for the formation of shape models on shape manifolds in the vocabulary, recursively. In addition, we suggest an approach to design inhibition methods in order to remove redundant measurements avoiding degenerate models in the vocabulary. The proposed approach is used to analyze different invariant properties of the shape spaces such as invariance to Euclidean Similarity Transformations (EST), Reflection(R), EST-R and Affine Transformation. In addition, results on benchmark shape classification outperform state-of-the-art shape representation methods, shape classification algorithms and hierarchical learning methods.

Relation to WP The paper addresses a theoretical approach for learning and modeling of relationships between deformations of shape models represented in shape spaces and geometric transformations employed on shapes of objects in measurement (i.e. data) spaces (Task 1.1).

2.3 Categorisation of 3D Objects in Range Images Using Compositional Hierarchies of Parts Based on MDL and Entropy Selection Criteria

Bibliography V. Kramarev, K. Walas and A. Leonardis Categorisation of 3D Objects in Range Images Using Compositional Hierarchies of Parts Based on MDL and Entropy Selection Criteria submitted to SCIA 2015.

Abstract This paper presents a novel approach to object categorisation for range images by using a hierarchical compositional representation. Three dimensional objects are described using surfaces of growing size and complexity. The atomic elements at the bottom layer of the hierarchy are small depth differences which are combined to form the second layer parts. Subsequent layers are formed in the recursive manner, each higher layer is statistically learnt on the layer below via a growing receptive field. Applying compositional hierarchies in this manner allows us to deal with partial occlusions – which are challenging for purely global shape descriptors – and to deal with intra-category variations of shape – which are hard to model using purely local shape descriptors. The compositional hierarchy learns and in a compositional way combines both descriptors in one structure.

Moreover, in this paper we focus on the part selection problem which concerns the choice of the optimisation criterion which provides the information on which parts should be promoted to the higher layer of the hierarchy. Namely two methods based on Minimum Description Length and category based entropy are introduced. The proposed approach was extensively tested on two widely available datasets for object categorisation with the results which are of the same quality as the best results achieved for those datasets.

Relation to WP The paper addresses Task 1.2 and presents an algorithm for learning of subsequent layers of the hierarchical 3D shape vocabulary.

2.4 SHREC2015 - Retrieval of Objects from Highly Cluttered Scene Captured with Depth-Sensing Camera – Evaluation Procedure

Bibliography K. Walas and A. Leonardis, SHREC2015 – Retrieval of Objects from Highly Cluttered Scene Captured with Depth-Sensing Camera – Evaluation Procedure

Abstract Evaluation measures introduced for the SHREC2015 track on Retrieval of Objects from Highly Cluttered Scene Captured with Depth-Sensing Camera.

Relation to WP The dataset and its evaluation procedure addresses Task 1.2 and allows to test our approaches on the challenging dataset and provides the new dataset to the community through the SHREC contest.

2.5 Analysis of 3D Surface Properties for Modeling 3D Compositional Representations

Bibliography K. Walas, M. Ozay, and J. L. Wyatt, Analysis of 3D Surface Properties for Modeling 3D Compositional Representations, School of Computer Science, University of Birmingham, Tech. Rep., 2015.

Abstract The aim of this report is to provide in-depth study of the intrinsic geometry of the surfaces and the properties of the distance measures on a surface together with surface parametrisation techniques. In consequence it will enable building the compositional hierarchy in 3D in the correct and mathematically sound manner. In order to do so, first we analysed the concepts from differential geometry. Next, the problem of measuring distances on a surface is tackled. Finally, the research on geometric scale space is performed. The topics studied in this report enable us to draw some new conclusions on the type of the representation required for learning compositional hierarchy with different levels of abstraction.

Relation to WP The technical report adresses Task 1.2 by providing in-depth study of the geometry of the surfaces, which provides a guidance in further development of the 3D compositional hierarchy.

2.6 "Statistical Learning of Multi-modal Representations of Objects using Multi-view 2D and View-based 3D Models

Bibliography M. Popa, M. Ozay, V. Kramarev, R. Aktas, K. Walas, J. Piater, A. Leonardis and J. Wyatt, Learning a multi-modal object representation, based on 2D and view-based 3D data – Technical Report

Abstract The objective of this technical report is to introduce a multi-modal object representation using multi-view 2D and view-based 3D models obtained from hierarchical compositional representations of 2D and 3D shapes. Furthermore, a method is proposed for the statistical analysis of the relationship between geometric properties of shape representations obtained from the multi-view 2D and view-based 3D compositional hierarchies.

Relation to WP The technical report addresses Task 1.3 by presenting and analyzing the developed algorithms for statistical feature binding in terms of joint probabilities between 2D and 3D parts.

Compositional Formation of 2D Shape Models for Multi-component Shape Representation

Mete Ozay, Umit Rusen Aktas, Jeremy L. Wyatt and Aleš Leonardis
School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, United Kingdom.
{m.ozay, u.aktas, j.l.wyatt, a.Leonardis}@cs.bham.ac.uk

Abstract

We address the problem of design and construction of shape representation models on manifolds when i) data and shape spaces are not manifolds and/or ii) the data and shape manifolds contain multiple disconnected components. We resolve problem i) by reformulating the relationship between data and shape spaces considering the interaction between Receptive Fields (RFs) and shape models in a shape vocabulary. Then, we suggest a method to design RFs considering both the topological structure of the models and statistical properties of the data. Problem ii) is considered by designing a disjoint union topology using an indexing mechanism for the formation of shape models on shape manifolds in the vocabulary, recursively. To this end we introduce a framework to implement the indexing mechanisms for i) the employment of the vocabulary for structural shape classification, and ii) designing inhibition methods to remove redundant measurements avoiding degenerate models in the vocabulary. The proposed approach is used to construct different shape representations with different invariance properties such as invariance to Euclidean Similarity Transformations (EST) and reflection. Results on benchmark shape classification outperform state-of-the-art shape classification methods.

1. Introduction

Statistical Shape Theory (SST) is a mathematical framework for the construction and analysis of *shape representations* of objects by analyzing their statistical properties [13, 22, 24, 38]. In the SST, shapes are represented in shape spaces which are usually considered as manifolds. This assumption fails if the shape spaces are not manifolds because of the *local singularities* that occur in the *shape spaces* [5, 22, 38].

In addition to this observation reported in the literature [5, 22, 38], we address two cases where the manifold assumption of data and shape spaces fails. First, the singularities may be observed in some components of representations because of the representation of landmarks or configurations using *redundant measurements* obtained from the images. Second, shape spaces may consist of multiple disconnected components, therefore the corresponding *shape manifolds* may be *disconnected* and non-smooth [29]. In this case, the methods which model shape deformations in connected shape manifolds may fail [37].

In the multi-component shape representation problem, we address the aforementioned cases in which the manifold assumptions of the state-of-the-art methods fail because of the multiple disconnected component structures of measurements and shape spaces. In other words, we provide a framework where the manifold assumptions can be relaxed using multi-component approximations in order to model shape representations. In this work, we address the problem of design and construction of shape representation models on manifolds when $i)$ data and shape spaces are not manifolds and/or $ii)$ the data and shape manifolds contain multiple disconnected components.

In order to solve the first problem, the relationship between data and shape spaces is reformulated considering the interaction between Receptive Fields (RFs) and shape models in a shape vocabulary. For this purpose, we first define an RF as a co-domain of a shape manifold that is represented by a shape model in a shape vocabulary. Then, we suggest a method to design structures of RFs considering both the statistical properties of the data, geometric structure of the shape manifolds, and topological structure of the models. We theoretically analyze the conditions that should be satisfied by an indexing mechanism to construct components of shape representations with manifold structures in shape vocabularies using the proposed method, even if the measurements are obtained from overlapping RFs.

A disjoint union topology [11] of components and compositions of representations is designed using an indexing mechanism for the recursive formation of shape models on shape manifolds in the vocabulary. For this purpose, we introduce a mathematical framework to implement the indexing mechanisms for $i)$ designing inhibition methods to remove redundant measurements in order to avoid degenerate models in the vocabulary, and ii) the employment of the vocabulary for structural shape classification. The proposed approach enables us to construct different shape representations with various invariant properties using different distance functions for statistical learning of models, and employ topological shape manifolds in the models.

1.1. Related Work and Background

Prior works in SST used landmarks or configurations to represent geometric information of *local* structures of shapes [13, 23, 22, 38]. In this approach, *shape* of an object is defined as all the geometrical information that remains when location, scale and rotational attributes are filtered out from the object [13, 24]. Mathematically speaking, shape of an object is a representation which is invariant under the Euclidean Similarity Transformations (EST) [13].

In order to construct EST-invariant representations, first a set of landmarks or primitive features is extracted from the measurements on objects, e.g. images [23]. Then, a scale and translation invariant representation, called *pre-shape*, is obtained using a set of centered unit-size landmarks. The complete set of all such pre-shapes is called *pre-shape space*, and its members lie on a unit sphere S_D^N [22]. Finally, an equivalence class (orbit) of the landmarks is computed under action of the special orthogonal group $SO(D)$, which is the set of all $D \times D$ rotation matrices Γ with $\Gamma^T\Gamma = \Gamma\Gamma^T = I_D$, where I_D is a $D \times D$ identity matrix [22]. The elements of the orbits reside in a space Σ_D^N called *shape space*. For $D = 2$, Σ_2^N is a compact smooth manifold, or *shape manifold*, and a shape is represented as a point in $\mathbb{C}P^{N-2}$ or a complex line in $\mathbb{C}^{N-1}$. Since the geometry of shape manifolds is non-trivial, development of recognition and statistical inference algorithms on shape manifolds is a challenge [20].

The methods which employ an *explicit* analysis on shape manifolds first embed the manifolds into some Euclidean spaces [5, 22]. Then, statistical learning and inference algorithms are implemented on the embedded submanifolds [5]. Kernel methods were employed in the implicitly mapped and embedded submanifolds [19, 20]. However, computation of the mappings is challenging because of the kernel parameter and type estimation problems. Statistical learning and optimization methods are employed on shape manifolds for *intrinsic* analysis of shape representations [1, 5, 22]. Statistical analysis of manifolds has been performed by (i) first mapping shape structures to tangent spaces, (ii) then computing the statistical properties on tangent spaces, and finally (iii) mapping the computed statistical models into shape manifolds [7, 26, 39]. However, the mappings may not exist, and the assumption of the algorithms to construct the models using the mappings may fail. Moreover, the mappings may not be invertible or unique [46]. The algorithms proposed for computing the mappings have several limitations such that the suggested representations and metrics may not be invariant to translations, and they lack of physical interpretations [26].

Note that, the manifold assumption of the aforementioned algorithms regarding the structure of Σ_D^N may fail in cases where Σ_D^N is not a manifold. For instance, multiple length minimizing geodesics between the *redundant* shape representations are observed on Σ_3^N [5, 22]. In order to solve the singularity problem, Riemannian submersion is employed on the non-singular part of S_D^N [5, 22, 38]. Although removing the *redundant* representations from pre-shape spaces and constructing shape spaces using the non-redundant representations provide a Riemannian manifold, it is not complete. Therefore, intrinsic analysis cannot be employed on the obtained shape manifold, and the methods which use explicit analysis are implemented to provide approximate solutions [5, 19, 20].

Shape representation algorithms which employ statistical analysis of manifolds on data obtained from images assume that the data lie on a manifold [35]. In order to obtain *manifold-like* data spaces, measurements are made on images using data structures that are approximations to local and open subsets such as local patches, sliding windows, or Receptive Fields (RFs) [35]. Therefore, the measurements obtained from RFs are considered as manifolds, and shape representation algorithms are employed on the data based on this assumption. However, the manifold interpretation of RFs may fail in some applications [30]. This problem has been defined as the determination of the structure of RFs and selection of their parameters [10, 18, 21, 27, 34, 41, 47].

In the next section, we first analyze the aforementioned approaches and methods used for Multi-component (MC) shape representation problem in SST. Then, we propose a new approach to resolve the challenges of their assumptions by re-defining the structures used for shape representation and developing mathematical models of inhibition processes, receptive fields, indexing mechanisms and model-based shape representations. In Section 3, we suggest a method to employ a structural classification algorithm using the proposed shape representation approach. In Section 4, the proposed methods are examined on benchmark datasets to perform different tasks in SST such as the construction of shape models equipped with different invariant properties for representation and classification of multi-component shapes observed in images. The results are compared with the state-of-the-art methods. Section 5 concludes the paper.

2. Multi-component Shape Representation

Multi-component (MC) shape representation problem consists of three sub-problems, namely i) measurement problem, ii) statistical learning of shape distributions, and

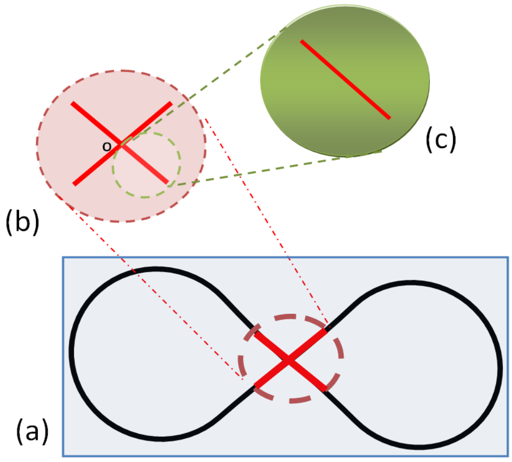


Figure 1: Measurement of shapes in Receptive Fields depicted by blue, red and green regions. The 8 shape given in (a) is not a manifold, and the cross (X) in (b) is not a manifold. However, the line in (c) is a one dimensional manifold.

$iii)$ construction of topological shape manifolds. The methods which solve the MC problem are considered in a compositional shape formation framework.

2.1. Measurement Problem

We will consider the design and utilization of RFs for the formation of shape models considering the statistical properties of the data, and topological properties of the shape manifolds. For instance, we consider a scenario where data measurements are obtained from an RF depicted with blue square region to model a manifold representation of a shape of the number **8** given in Fig. 1.a. However, the data space of the shape is not a manifold. If we make measurements from another RF with different size and structure depicted by the red circular region to obtain the data manifold in Fig. 1.b, we also observe that the red cross × is not a manifold. The reason is: the subspace topology is not locally Euclidean at the intersection point **O**. In other words, if we consider a neighborhood U of **O**, then we observe that $U - \{\mathbf{O}\}$ has four connected components which prevent the existence of a homeomorphism from U to an open ball $\mathcal{B}(\mathbf{x}, r) = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < r\} \subset \mathbb{R}^D$, where $d(\cdot, \cdot)$ is a metric and $r \in \mathbb{R}^D$ [42]. On the other hand, each component of Fig. 1.b, one of which is a line depicted in Fig. 1.c, is a one dimensional manifold.

As we observed in this example, we cannot assure the representation of shapes on shape manifolds by using mappings from data spaces obtained from different RFs with different structures and sizes, since the mappings may not be homeomorphism or their parametrization may not exist. We address this problem by considering a shape space as a manifold, and the measurements obtained from an RF as a co-domain of the manifold.

In our proposed approach, we assume that a set of measurements are obtained from an open ball $\mathcal{B}(\mathbf{x}_c, r) = \{\mathbf{x}' : \|\mathbf{x}_c - \mathbf{x}'\|_2 < r, r \in \mathbb{R}, \mathbf{x}' \in A\}$ which is represented by a Receptive Field centered at $\mathbf{x}_c \in \mathbb{R}^D$, where $A \in \mathbb{R}^D$ is an open subset. In this work, we use two-dimensional images, therefore $D = 2$ will be considered in the rest of the paper, if not stated explicitly. Additionally, we assume that we extract landmarks as Gabor features extracted from images at each image point $\mathbf{x} \in \mathbb{R}^2$ using a two-dimensional Gabor kernel $G(\mathbf{x}, \mathcal{R})$, where $\mathcal{R}$ is a set of parameters [14]. Assuming that the parameters of the Gabor functions are fixed, we propose a method to select the size or structure of the RFs as follows. We define the intensity value of the pixel at $\mathbf{x}$ by $I(\mathbf{x})$. The Gabor filter $G(\mathbf{x})$ and $I(\mathbf{x})$ are convolved using the measurements obtained from an RF centered at $\mathbf{x}_c$ as

$$\mathfrak{C}(\mathbf{x}_c, r) = \mathfrak{F}^{-1}\{\mathfrak{F}(I(\mathbf{x})) \odot \mathfrak{F}(G(\mathbf{x}))\}, \forall \mathbf{x}_c,$$

where $\mathfrak{F}(I(\mathbf{x}))$ is the Fourier transform of $I(\mathbf{x})$ and $\odot$ is the point-wise multiplication. Then, we estimate an optimal radius $\hat{r}$ of the ball $\mathcal{B}(\mathbf{x}_c, r)$ as

$$\hat{r} = \arg\max_{r} \{\frac{1}{2r} \arg\max_{\mathbf{x}_c} \mathfrak{C}(\mathbf{x}_c, r)\}, \forall \mathbf{x}_c,$$

where $\frac{1}{2r}$ is used following the circular structure of the RF. This method was previously suggested for learning receptive fields using three-dimensional tensors in images [47]. In our proposed approach, we employ the method for designing open ball shaped RFs in two-dimensional images. Due to the co-domain interpretation of the proposed RFs, our implementation is faster by order of number of scales since the RFs are used to obtain scale invariant manifold representations which is explained in the next section.

Selection of the RFs using this approach provides us the features with maximum response in a neighborhood or ball $\mathcal{B}(\mathbf{x}_c, r)$, $\forall \mathbf{x}_c$. However, redundant measurements may be obtained from the regions in which the RFs are overlapped. In order to remove the redundancy, pooling [21, 27], weighted selection [34] or clustering [47] methods have been employed on the set of measurements. These methods aim to remove the redundancy according to the statistical properties [27, 10] of the data independent of the requirements of the representation models due to the nonlinearity of the relationship between the models and the measurements.

2.2. Multi-component Shape Representation Model

We consider a Receptive Field (RF) as a co-domain from which the measurements are obtained within open sets using open balls. The co-domain U determined by an RF is defined together with a manifold

$$\mathcal{P} = \{(A_j, \phi_j) | \phi_j : A_j \to U_j \subset \mathbb{R}^D, \forall j \in \mathfrak{I}\},$$

where $\mathfrak{I}$ is a nonempty countable set. We call the manifold $\mathcal{P}$ a *part*, and (A_j, ϕ_j) is a chart of the $\mathcal{P}$, $\forall j \in \mathfrak{I}$. In SST, a part $\mathcal{P}$ can be considered as an element of a shape space, which is also designed to be a manifold. Therefore, we should first show that $\mathcal{P}$ is a manifold which is a non-degenerate representation model. Following the shape space construction approach used in SST and discussed in the previous section, parts will be constructed using equivalence relations between shape representation models through quotient maps of data and shape spaces.

2.2.1 Geometry of Shape Representation Models

In order to employ group operations on the measurements, we denote a matrix representation of the measurements as $X \in \mathbb{R}^{N_j \times D}$, where the i^{th} row of X is $\mathbf{x}_i \in A_j$, and $N_j = |A_j|$, $\forall j \in \mathfrak{I}$. Then, we extend the Euclidean Similarity Transformation (EST) model used in SST [13] to obtain EST and reflection invariant (EST-R) representations. In addition, we suggest an inhibition method to define non-degenerate shape models in different shape spaces.

For this purpose, we first define a scaled and translated (ST) shape as

$$[X]_{st} = \{\beta X + I_D \gamma^T : \beta \in \mathbb{R}, \gamma \in \mathbb{R}^D\},$$

where $\beta \in \mathbb{R}^+$ is a scale variable used for isotropic scaling, and $\gamma \in \mathbb{R}^D$ is a translation vector. Note that the ST space $\mathcal{S}_{st}$, which is the space of all possible pre-shapes, is scale and translation invariant [13, 22].

Inhibition: In order to assure that the ST space $\mathcal{S}_{st}$ is a manifold, we should satisfy the requirement that $rank(X_c) \geq D$, where $X_c \in \mathcal{S}_{st}$. Selection of κ number of independent columns (rows) of a matrix has been studied as a subset selection problem [16, 17, 45]. In a recent work, a sparse encoding algorithm is proposed to provide an approximate solution to the problem [34]. In this work, first the generalized pooling method of [34] is extended with the constraint $D \leq \kappa \leq N$. Then the problem is solved using a feature split distributed optimization algorithm [6] considering the recursive structural relationship between co-domains of parts, parts and their compositions.

In order to obtain EST invariant representations, we employ actions from the rotation group as

$$[X]_{est} = \{X_{st}\Gamma : \Gamma \in SO(D), X_{st} \in \mathcal{S}_{st}\},$$

where $\Gamma \in SO(D)$ is a rotation matrix. The EST shape space consisting of all possible EST Shapes $[X]_{est}$ is denoted by $\mathcal{S}_{est}$.

Note that, if $D > 2$, then the shape space consisting of elements of $\mathcal{S}_{est}$ is degenerate. In order to remove the degeneracy, the parts which have rank-deficient measurements are removed from the space [5]. Although this operation provides a Riemannian manifold, the elements are disconnected and the space is separated into multiple components [5]. Since we consider $D = 2$ in this work, we assure that the space of $\mathcal{S}_{est}$ is non-degenerate. In order to solve this problem, the action of manifold involution is employed on the $\mathcal{S}_{est}$ [5, 12, 22]. Then, we obtain a reflection shape space $\mathcal{S}_{er}$ consisting of all reflection shapes which are defined as

$$[X]_{er} = \{X_{est}R : R \in O(D), X_{est} \in \mathcal{S}_{est}\}, \qquad (1)$$

where $O(D)$ is the group of all $D \times D$ orthogonal matrices. In order to detect and remove the degenerate representations in the $\mathcal{S}_{er}$ space, we should check the rank deficiency property of the measurements $X_{er} \in \mathcal{S}_{er}$. For this purpose we employ the proposed inhibition method.

Once the degenerate representations are removed, we obtain a compact differentiable manifold, and can employ extrinsic and intrinsic analysis on the manifold for statistical learning of shape representations [5]. The distance functions of representations that will be employed in the learning algorithms are designed according to the geometric and topological structure of the spaces. Therefore, we will analyze the geometric and topological properties of shape spaces in the following subsection before introducing the methods proposed for statistical learning of representations.

2.2.2 Topological Properties of Shape Manifolds

We have shown that a part $\mathcal{P}$ is a manifold using quotient maps to construct EST and reflection invariant representations. Now, we will analyze the conditions which enable us to model the manifold structure of a part $\mathcal{P}$ when there is an overlap among the charts of the $\mathcal{P}$, and among the corresponding RFs in the measurement spaces. Given any two charts (A_i, ϕ_i) and (A_j, ϕ_j) on $\mathcal{P}$, where $A_i \cap A_j \neq \varnothing$, the transition maps are defined as [15]

$$\phi_{ji} : \phi_i(A_i \cap A_j) \to \phi_j(A_i \cap A_j),$$

$$\phi_{ij} : \phi_j(A_i \cap A_j) \to \phi_i(A_i \cap A_j),$$

$$\phi_{ji} = \phi_j \circ \phi_i^{-1} \text{ and } \phi_{ij} = \phi_i \circ \phi_j^{-1}, \forall i, j \in \mathfrak{I}.$$

Defining $U_{ij} = \phi_i(A_i \cap A_j)$ and $U_{ji} = \phi_j(A_i \cap A_j)$, we compute a map $\phi_{ji} : U_{ij} \to U_{ji}$ which is called a *gluing map* in [15]. Then, a set of gluing measurements, which is a superset of RFs used on the images, is defined as [15]

$$\mathfrak{RF} = \{\{U_i\}_{i \in \mathfrak{I}}, \{U_{ij}\}_{(i,j) \in \mathcal{L}}, \{\phi_{ji}\}_{(i,j) \in \mathcal{L}}\}, \qquad (2)$$

where $\mathcal{L} = \{(i,j) \in \mathfrak{I} \times \mathfrak{I} : U_{ij} \neq \varnothing\}$ is an *index set*. In order to assure that there is a manifold with co-domain and maps defined by $\mathfrak{RF}$, the maps should satisfy the following conditions [15]:

1. $\phi_{ii} = id_{U_i}$, where id_{U_i} is an identity map of U_i, $\forall i \in \mathfrak{I}$.

2. $\phi_{ij} = \phi_{ji}^{-1}$, $\forall (i,j) \in \mathcal{L}$.

3. If $U_{ji} \cap U_{jk} \neq \varnothing$, then $\phi_{ij}(U_{ji} \cap U_{jk}) = U_{ij} \cap U_{ik}$, $\forall i,j,k \in \mathfrak{I}$, and $\phi_{ki}(\mathbf{x}) = \phi_{kj} \circ \phi_{ji}(\mathbf{x})$, $\forall \mathbf{x} \in (U_{ij} \cap U_{ik})$.

4. There are open balls $V_{\mathbf{x}}$ and $V_{\mathbf{x}'}$ centered at $\mathbf{x}$ and $\mathbf{x}'$ such that no point of $V_{\mathbf{x}'} \cap U_{ji}$ is the image of any point of $V_{\mathbf{x}} \cap U_{ij}$ by ϕ_{ji}, $\forall (i,j) \in \mathcal{L}$ with $i \neq j$, and $\forall \mathbf{x} \in \partial(U_{ij}) \cap U_i$, $\forall \mathbf{x}' \in \partial(U_{ji}) \cap U_j$, where $\partial(U_{ij})$ is the boundary of U_{ij}.

In our proposed approach, centering and normalization map of features used to construct ST spaces is a transition map induced by an *indexing mechanism* that satisfies the conditions defined above. Next, we provide a theorem which associates the measurement spaces defined by RFs to D dimensional manifolds.

Theorem 2.1 ([15]). *For every set $\mathfrak{RF}$ defined in (2) there is a D dimensional manifold whose transition maps are $\{\phi_{ji}\}_{(i,j)\in\mathcal{L}}$, where $\mathcal{L}$ is an index set. More precisely, there exists a manifold $\mathfrak{M} = (\mathfrak{RF}, \{\psi\}_{i\in\mathfrak{I}})$, where $\psi_i : U_i \to \mathbb{R}^D$ is defined as*

$$\psi_i \triangleq \psi_j \circ \phi_{ji}, \forall (i,j) \in \mathcal{L}.$$

Corollary 1. *$\mathcal{P}$ is a manifold whose co-domain and transition maps are defined by $\mathfrak{RF}$ with measurements on two dimensional images.*

The formal proof of the **Corollary 1** follows the proof idea of the **Theorem 1** which is given in the Supplemental Material. We will provide two crucial ideas of the proof in order to support our motivation of employing quotient maps and indexing mechanisms below.

Proof idea of the Corollary 1. The proof idea is first i) forming the disjoint union of sets in the co-domain, and then identifying overlapping sets with transition maps which are ii) defined by quotient maps, and iii) traced by indexing.

Disjoint union of sets is defined by

$$\mathbb{U} \triangleq \coprod_{i\in\mathfrak{I}} U_i = \bigcup_{i\in\mathfrak{I}} \mathcal{U}_i, \tag{3}$$

where $\mathcal{U}_i = U_i \times \{i\}$, $\forall i \in \mathfrak{I}$ [11, 33]. Disjoint union of sets is a collection of sets where the indices of the elements are recorded by the *indexing mechanism*. In other words, a disjoint topology requires the employment of a set of indices of the elements using an indexing mechanism. Moreover, the mathematical structure of indexing mechanism is crucial for the proof. Let's start by defining the orbit of X as

$$\mathfrak{o}(X) = [X] = \{X_{er} : X_{er} \in \mathcal{S}_{er}\},$$

where $\mathcal{S}_{er}$ is defined in (1). Then, we obtain $\mathfrak{o} : \mathbb{U} \to \mathcal{P}$. Following the definition of the orbit by quotients, we can define an equivalence relation as $X \sim X'$ iff $X' \in \mathfrak{o}(X)$. We can employ this equivalence relation on $\mathbb{U}$ stating for all $X, X' \in \mathbb{U}$ that $X \sim X'$ iff $\exists (i,j) \in \mathcal{L}$, such that $X' = \phi_{ij}(X)$ for $X \in U_{ij}, X' \in U_{ji}$. Therefore, we have $\mathcal{P} = \mathbb{U}/\sim$.

Next, we define a natural injection $\eta_i : U_i \hookrightarrow \mathbb{U}, \forall i \in \mathfrak{I}$ which will be used for indexing and defining

$$t_i = \mathfrak{o} \circ \eta_i : U_i \to \mathcal{P}.$$

Since $A_i = t_i(U_i)$ and $\phi_i = t_i^{-1}$, we observe that (A_i, ϕ_i) is a chart of $\mathcal{P}$ and their collection form $\mathcal{P}$, $\forall i \in \mathfrak{I}$.

In order to prove that $\mathcal{P}$ is Hausdorff, we choose two orbits $[X] \in \mathcal{P}$ and $[X'] \in \mathcal{P}$ of any two elements $X \in U_i$ and $X' \in U_j$, $\forall i,j \in \mathfrak{I}$. Then, these measurements can be represented by either disjoint or overlapping models in $\mathcal{P}$. Since there are countably many charts defined by the finite number of RFs, $\mathcal{P}$ is second countable. Additionally, ϕ_{ij} are the maps of $\mathcal{P}$ since $\phi_i = \tau_i^{-1}$ and $\phi_{ji} = \phi_j \circ \phi_i^{-1}$, $\forall i,j \in \mathfrak{I}$, $\forall (i,j) \in \mathcal{L}$. □

In **Corollary 1**, we show that a part $\mathcal{P}$ is a manifold. Next, we propose that their composition is a manifold.

Theorem 2.2. *Given a collection of parts $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$, where $\mathfrak{N}$ is a set of indices, a composition $\mathcal{C}$ of parts defined as*

$$\mathcal{C} = \coprod_{n\in\mathfrak{N}} \mathcal{P}_n \tag{4}$$

is a manifold.

The formal proof of the **Theorem 2.2** is given in the Supplemental Material.

2.2.3 Statistical Learning of Representations

Let Q be a probability distribution on $\mathcal{C}$, and $\{\Omega_n\}_{n\in\mathfrak{N}}$ are random variables associated to $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$ and drawn independent and identically distributed (i.i.d.) from Q. In the statistical learning model, we aim to compute the joint observation probability of two parts with different part indices or ids in locally defined open sets, which are clusters. In other words, we compute a set of disjoint clusters $\Lambda_k \subset \mathcal{C}$, $\forall k = 1, 2, \ldots, K$, such that $\bigcup_{k=1}^{K} \Lambda_k = \mathcal{C}$. Moreover, we require that if $\Lambda_k \subseteq \Lambda_l$, then $\sigma(\Lambda_k) \leq \sigma(\Lambda_l)$, where $\sigma(\cdot)$ is a cost function used to avoid obtaining subsets of clusters. This condition is defined and used as the Monotone Clustering Property (MCP) for Intracluster criterion-based clustering in [9]. We define the cost function of an instance of a cluster as

$$\sigma(\lambda_k, \omega_n, \omega_m) = \Pr(\lambda_k \in \Lambda_k | \omega_n \in \Omega_n, \omega_m \in \Omega_m) \tag{5}$$

which is the probability of observing two given part instances $\omega_n \in \Omega_n$ and $\omega_m \in \Omega_m$ in a cluster instance $\lambda_k \in \Lambda_k$, $\forall k$. Computation of a collection of clusters which maximize (5) is considered as a minimum entropy clustering problem [9, 31] and defined as

$$\mathcal{K} = \arg\min_{\mathfrak{K}} \sum_{\omega_n \in \Omega_n} \sum_{\omega_m \in \Omega_m} \sum_{\lambda_k \in \Lambda_k} \Pr{}^{\alpha}(\lambda_k | \omega_n, \omega_m) - 1, \quad (6)$$

where $\mathfrak{K}$ is the set of all possible partitions of the dataset among K clusters. We choose $0 < \alpha < 1$ unlike the clustering problem proposed in [2] to satisfy the MCP criterion. In a clustering algorithm [31] which computes (6), a part $\mathcal{P}_m$ that resides in the neighborhood of a given part $\mathcal{P}_n$ is selected to minimize the conditional entropy associated with the conditional probability defined in (5). In the computation of the neighborhood of $\mathcal{P}_n$, we use different distance functions by employing extrinsic analysis on the shape manifold $\mathcal{C}$. We used $\mathcal{S}_{est}$, $\mathcal{S}_{er}$, the Reflection Space $\mathcal{S}_r$ and the Affine Space $\mathcal{S}_a$ [5] in the experiments.

2.2.4 Compositional Formation of Hierarchy of Shape Manifolds

In the proposed approach, we construct parts $\{\mathcal{P}_n\}_{n \in \mathfrak{N}}$, and the shape manifold $\mathcal{C}$ employing a disjoint union operation [11]. In addition, we integrate a set of shape manifolds $\{\mathcal{C}_m\}_{m \in \mathfrak{M}}$ to construct a shape vocabulary $\Upsilon = \coprod_{m \in \mathfrak{M}} \mathcal{C}_m$, where $\mathfrak{M}$ is a nonempty countable set of indices. Therefore, we consider a compositional approach for the formation of the vocabulary Υ.

We represent a shape manifold $\mathcal{C}_m^l$ constructed at the l^{th} layer by a graph $\mathcal{G}_m^l = (\mathcal{V}_m^l, \mathcal{E}_m^l)$, $\forall m \in \mathfrak{M}$. $\mathcal{V}_m^l$ is the set of nodes where each node $v_i^{l+1} = n$ represents an index label or id $n \in \mathfrak{N}$ of a subpart $\mathcal{P}_n^l$, $\forall m$.

At $l = 1$, we have disconnected parts, therefore $e_{ij} = \emptyset$, $\forall e_{ij} \in \mathcal{E}_m^1$, $\forall m$. For $l > 1$, if two parts $\mathcal{P}_n^l$ and $\mathcal{P}_{n'}^l$ co-occur in a cluster λ_k with probability $\sigma(\lambda_k, \omega_n, \omega_{n'}) > 0$, $v_i^l = n$ and $v_j^l = n'$, then $e_{ij} = k$. Following the recursive definition of the vocabulary, a shape manifold $\mathcal{C}_m^l$ is considered as a part of another shape manifold $\mathcal{C}^{l+1}$ which is referred as a *composition* constructed at the $(l+1)^{st}$ layer. At each layer l, we can compare two graph representations $\mathcal{G}_m^l$ and $\mathcal{G}_{m'}^l$ considering the manifold structure of shape spaces. Given two graphs $\mathcal{G}_m^l$ and $\mathcal{G}_{m'}^l$, deciding whether they are isomorphic is defined as a graph isomorphism problem [36, 43].

At the first layer $l = 1$ in the vocabulary, we assure the existence of a bijection $\mathfrak{b} : \mathcal{V}_n^l \to \mathcal{V}_{n'}^l$ between two graph representations $\mathcal{G}_n^l$ and $\mathcal{G}_{n'}^l$ of two parts $\mathcal{P}_n^l$ and $\mathcal{P}_{n'}^l$ in a shape manifold $\mathcal{C}_m^{l+1}$, respectively. In other words, $\mathcal{G}_n^l \sim \mathcal{G}_{n'}^l$ iff $\mathcal{G}_{n'}^l \in \mathfrak{b}(\mathcal{G}_n^l)$, $\forall n, n' \in \mathfrak{N}$ following the compositional formation of parts as described in the previous section.

For $l > 1$, we search and select the graphs $\mathcal{G}_n^l$ of the parts $\mathcal{P}_n^l$ that are subgraphs and isomorphic to the graphs $\mathcal{G}_m^{l+1}$ of compositions $\mathcal{C}_m^{l+1}$, $\forall m \in \mathfrak{M}$. This problem has been defined and studied as a sub-graph isomorphism problem [2, 36, 43]. In this work, we select the subgraphs of compositions $\mathcal{C}_m^{l+1}$, $\forall m \in \mathfrak{M}$ solving the sub-graph isomorphism problem using the implementation provided in [2]. We search the shape manifolds and increment the hierarchy until we obtain a compact connected shape manifold which contains a single component at the highest layer of the hierarchy.

2.3. Inference of Shape Representations on Test Data

In the inference algorithm, the indexing mechanism proposed in the previous section is employed to induce a disjoint union topology in the inference tree $\mathcal{T}$. We first extract a set of landmarks or features from a test image. Then, we represent the test features as parts $\{\check{\mathcal{P}}_n^l\}_{n \in \mathfrak{N}}$ and shape manifolds $\{\check{\mathcal{C}}_m^l\}_{m \in \mathfrak{M}}$ at the first layer $l = 1$ of an inference tree $\mathcal{T}$ such that $\check{\mathcal{C}}_m^1 = \coprod_{n \in \mathfrak{N}} \check{\mathcal{P}}_n^1$, $\forall m \in \mathfrak{M}$.

At each consecutive layer $l > 1$, we search for a set of graphs $\{\check{\mathcal{G}}_u^l : u \in \mathfrak{U}\}$ whose members are graph isomorphic to the shape manifolds $\mathcal{G}_n^l$, $\forall n \in \mathfrak{N}$ in the vocabulary Υ. More precisely, we first compute a set of all graphs $iso(\mathcal{G}_n^l)$ which are isomorphic to $\mathcal{G}_n^l$ $\forall n$, and then search the set of parts $\check{\mathcal{G}}_u^l \in iso(\mathcal{G}_n^l)$, $\forall u \in \mathfrak{U}$, where $\mathfrak{U}$ is a nonempty countable set of indices. In order to obtain an approximate solution to the graph isomorphism problem, we used the inexact graph matching implementation provided in [2].

3. Shape Classification using Multi-component Shape Representations

In the shape classification problem, we consider *contribution* of each part, composition and shape manifold in the vocabulary to the representation of classes using a top-down approach. For this purpose, we employ the topological structure of the vocabulary Υ which is induced by the indexing mechanism, such as the recursive representation of parts and compositions at each layer $l > 1$ of Υ.

After shape manifolds are formed in a hierarchical vocabulary, we first compute training and test inference trees $\mathfrak{T} = \{\mathcal{T}_n : n = 1, 2, \ldots, N\}$ and $\check{\mathcal{T}}$ using a set of N training images and a given test image, respectively. Let $F \in \mathbb{R}^{N \times H}$ denote an input matrix whose row vector $\mathbf{f}_n$ represents the detection of a feature by a part or a composition in Υ on the n^{th} image provided by an inference tree $\mathcal{T}_n$. We encode the class labels using 1-of-W coding of classes using a matrix $Y \in \mathbb{R}^{N \times W}$, where $y_{nw} = 1$ if the n^{th} image belongs to the w^{th} class, and $y_{nw} = 0$, otherwise. In addition, we assume a linear class model for each class variable as

$$\mathbf{y}_w = F\mathbf{z}_w + \epsilon_w, \forall w = 1, 2, \ldots, W,$$

where $\mathbf{z}_w \in \mathbb{R}^H$ is a vector of model coefficients, and ϵ_w is a vector that represents model error. Then the class model estimation problem is defined as [25]

$$\hat{Z} = \arg\min \Big(\sum_{w=1}^{W} \big(\mathbf{y}_w - F\mathbf{z}_w\big)^T \big(\mathbf{y}_w - F\mathbf{z}_w\big) + \mathfrak{L} \Big), \quad (7)$$

where $\hat{Z}$ is the matrix of estimated model variables, and $\mathfrak{L}$ is a structural loss function. In order to employ the structure induced by the vocabulary in the classification models, we define the loss function $\mathfrak{L}$ as [25]

$$\mathfrak{L} \triangleq \xi \sum_{h=1}^{H} \sum_{\upsilon_i \in \Lambda_{int}} \mu_{\upsilon_i} \|\mathbf{z}^h_{\Phi_{\upsilon_i}}\|_2 + \xi \sum_{h=1}^{H} \sum_{\upsilon_j \in \Lambda_{leaf}} \mu_{\upsilon_j} \|\mathbf{z}^h_{\Phi_{\upsilon_j}}\|_2, \quad (8)$$

where ξ is a regularization parameter. Λ_{leaf} is the number of leaf nodes in an inference tree which represent the detections by parts in Υ. Λ_{int} is the number of internal nodes in an inference tree which represent the detections by either parts or compositions in Υ according to the topology induced by the indexing mechanism. In other words, Λ_{int} consists of shared parts and compositions among different objects and different layers, and defines the overlapping groups. μ_{υ_i} and μ_{υ_j} are the regularization weights associated to each node $\upsilon_i \in \Lambda_{int}$ and $\upsilon_j \in \Lambda_{leaf}$.

Φ_{υ_i} and Φ_{υ_j} are group structures whose members are the indices of leaf nodes of the subtrees rooted at the nodes $\upsilon_i \in \Lambda_{int}$ and $\upsilon_j \in \Lambda_{leaf}$. Using the group structures encoded by the indexing mechanism, values of features detected at the corresponding parts and compositions are encoded in $\mathbf{z}^h_{\Phi_{\upsilon_i}}$ and $\mathbf{z}^h_{\Phi_{\upsilon_j}}$, which are the vectors of variables $z^h_{\varrho_i}, \forall \varrho_i \in \Phi_{\upsilon_i}, \forall \upsilon_i \in \Lambda_{int}$ and $z^h_{\varrho_j}, \forall \varrho_j \in \Phi_{\upsilon_j}, \forall \upsilon_j \in \Lambda_{int}$ [25]. The problem (7) is defined as a tree guided group lasso problem in [25]. In this work, we solve the problem for an ensemble of trees induced by the learned shape vocabulary.

4. Experiments

Landmark points are defined by features which are extracted from the images using Gabor Filters whose parameters are selected using the Information Diagram method as suggested in [14]. We have used an implementation of the MEC algorithm given in [31] for statistical learning. We used the algorithms and implementation provided by [2] to solve Graph and Subgraph Isomorphism problems. Shape vocabularies are learned in different shape spaces, and classification models are learned using the TL algorithm [25] whose parameters are selected as suggested in [25]. The results are given for four different cases each of which is denoted by **MC-Z**, where $\mathbf{Z} \in \{1,2,3,4\}$ represents the employment of statistical learning algorithms in the shape spaces $\mathcal{S}_{est}$, $\mathcal{S}_{er}$, $\mathcal{S}_r$ and $\mathcal{S}_a$, respectively.

Table 1: Classification performance (%) of the algorithms on the benchmark shape datasets.

Algorithm	CS	M-R	M-RBI
DBN-3 [28]	81.37	87.70	71.49
DAE2 [8]	81.56	88.06	55.08
CAE2 [8]	80.70	86.38	51.75
mDAE2 [8]	81.90	89.64	53.88
MC-1	82.90	89.76	79.58
MC-2	84.29	88.22	80.53
MC-3	82.62	89.30	65.60
MC-4	84.84	86.14	68.63

4.1. Experimental Analysis on Artificial Datasets

In this section, we examine the suggested approach and the algorithms on three different benchmark artificial datasets. The datasets are generated such that the shapes consisting of single and/or multiple components are corrupted by rigid body transformations and noise [28].

Convex Shapes (**CS**) dataset [28] consists of 8000 training and 50000 test images each of which contains convex and non-convex regions, and belong to two classes. In this dataset, the classification problem is defined as a binary classification of convex and non-convex regions. Mnist-Rotation (**M-R**) dataset [28] consists of 12000 training and 50000 test images of the Mnist digits which belong to ten classes and are rotated by an angle generated uniformly between 0 and 2π radians. In the Mnist-RBI (**M-RBI**) dataset [28], a random patch from a black and white image is used as the background for the rotated images obtained from the Mnist-Rotation dataset. The size of each image is 28×28. The proposed algorithms are compared with Deep Belief Networks with three layers (**DBN-3**), Denoising (**DAE-2**), Contractive (**CAE-2**) and marginalized Denoising auto-enconder (**mDAE-2**) algorithms with two layers.

The results given in Table 1 show that the proposed algorithms outperform the state-of-the-art algorithms. In addition, we observe that the proposed algorithms can employ invariant properties of shape manifolds on shape representations. For instance, affine invariant shape spaces are used in **MC-4** for encoding information about convexity of shapes. Therefore, the best classification performance is observed in **MC-4** on the **CS** dataset.

In addition, we observe that EST invariant shape manifolds used in **MC-1** provide better representations of rotated shapes obtained from the **M-R** dataset than the other shape spaces. An interesting observation is that the proposed methods which model representations in a more complex shape space $\mathcal{S}_{er}$ in **MC-2** outperform the algorithms in noisy images which contain corrupted shapes belonging to the **M-RBI** dataset. The reason of performance boost of the algorithms which use $\mathcal{S}_{er}$ is that the symmetry patterns observed in the background images are recognized and removed from the shape representations by the proposed inhibition and graph compression algorithms.

Table 2: Classification performance (% Mean ± Variance) of the proposed algorithms and the shape descriptors on the Animals Dataset.

CSS [40]	**IDSC** [32]	**CSSP** [3]	**NN-SVM** [4]
69.7	73.6	78.4	84.30 ± 1.0
MC-1	**MC-2**	**MC-3**	**MC-4**
85.1 ± 2.3	86.4 ± 2.1	77.9 ± 2.3	81.1 ±2.3

Table 3: Classification performance (% Mean ± Variance) of algorithms using the Butterfly Dataset.

S-kFP [19]	**S-kFPG** [20]	**MKL** [19]	**SM** [19]
57.75 ± 2.0	60.37 ± 1.6	60.84 ± 2.0	63.98 ± 1.6
MC-1	**MC-2**	**MC-3**	**MC-4**
64.81 ± 1.8	67.13 ± 1.4	61.34 ± 2.2	61.57 ± 2.1

4.2. Classification Results on Real-world Datasets

In this section, we provide classification results of the algorithms using two benchmark real-world datasets. The Animals Dataset [3] consists of 2000 shapes belonging to 20 categories. We have used this dataset instead of the MPEG-7 dataset since the Animals dataset contains more images that are deformed with more articulated and non-rigid transformations, and larger class-variation than the MPEG-7 dataset [4]. Following the data setup suggested in [4], 50 images per class were randomly chosen for training, and the rest of the samples were used for testing. The procedure is repeated 10 times for the proposed algorithms and the results are compared with the performances of the state-of-the-art shape descriptors.

The results given in Table 2 show that the average performance values of the proposed algorithms are greater and the variance of values are less than the values of the other algorithms. Note that similar variance values are obtained for the proposed algorithms. In addition, we report the performances of the proposed algorithms over 10 sets of experiments unlike 100 sets of experiments reported for the **NN-SVM** [4]. Therefore, the higher variance can be attributed to the smaller variance of the data generated in the experimental setup.

In the second set of the experiments, we examined the classification performance of the proposed algorithms with the state-of-the-art shape classification algorithms that are implemented on shape manifolds using the Butterfly Dataset [44]. The dataset contains 823 images belonging to 10 classes. Following the data setup suggested in [19], we randomly selected 40 shapes from each class for training and used the rest for testing. The procedure was repeated 10 times, and mean and variance of the classification performance values are given in Table 3.

In the real world datasets which contain shapes that are deformed by articulation and non-rigid body transformations, the algorithms need to model higher variance of the geometric patterns distributed among different measurements, objects and classes. Therefore, the algorithms which use shape spaces with *richer* geometric properties perform better than the other algorithms on these datasets.

For instance, **S-kFP** and **MKL** employ full Procrustes (FP) on the Kendall's shape space [19]. Similarly, we have used the FP distance for statistical learning on the $\mathcal{S}_{est}$ which corresponds to the Kendall's shape in **MC-1**. Meanwhile, Grassmann manifolds are used in **SM**. Since our proposed method considers the contribution of each part and composition to the classification model, we obtain better classification performances. Moreover, we extended the $\mathcal{S}_{est}$ with manifold involution in **MC-2** obtaining a *richer* shape space. Therefore, we obtained *more complex* representations and better classification performances. Additional experimental results are given in the Supplemental Material.

5. Conclusion

Multi-component Shape Representation problem has been proposed and addressed for design and construction of shape representation models on manifolds. For this purpose, shape representations are modeled with topologies induced by the statistical properties of the data and geometric properties of shape spaces that we want to achieve. The relationship between the data and the shape spaces has been reformulated using RFs, and a method for designing RFs is proposed. Then, a inhibition method has been employed to remove the redundant measurements and degenerate representations in models.

In order to employ and encode the geometric information obtained from measurements in the shape models, we have suggested a mathematical framework of an indexing mechanism. We have modeled formation of shape models in a vocabulary using a disjoint union topology of shape manifolds induced by the indexing structure of the vocabulary.

The proposed shape representation methods have been used for structural shape classification. In the experiments, we have constructed different shape models by designing various shape manifolds with different invariant properties for shape classification. The proposed algorithms have performed better than the state-of-the-art classification algorithms on various benchmark datasets.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] U. Aktas, M. Ozay, A. Leonardis, and J. Wyatt. A graph theoretic approach for object shape representation in compositional hierarchies using a hybrid generative-descriptive model. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Proceedings of the 2014 European Conference on Computer Vision*, ECCV '14, pages 566–581. Springer, 2014.

[3] X. Bai, W. Liu, and Z. Tu. Integrating contour and skeleton for shape classification. In *Proceedings of the 2009 IEEE International Conference on Computer Vision*, ICCV '09, pages 360–367, Sept 2009.

[4] X. Bai, C. Rao, and X. Wang. Shape vocabulary: A robust and efficient shape representation for shape matching. *IEEE Trans. Image Processing*, 23(9):3935–3949, Sept 2014.

[5] A. Bhattacharya and R. Bhattacharya. *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*. Cambridge University Press, New York, NY, USA, 1st edition, 2012.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan 2011.

[7] D. Bryner, E. Klassen, H. Le, and A. Srivastava. 2d affine and projective shape analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(5):998–1011, May 2014.

[8] M. Chen, K. Q. Weinberger, F. Sha, and Y. Bengio. Marginalized denoising auto-encoders for nonlinear representations. In T. Jebara and E. P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1476–1484. JMLR Workshop and Conference Proceedings, 2014.

[9] I. Christou. Coordination of cluster ensembles via exact methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):279–293, Feb 2011.

[10] A. Coates and A. Y. Ng. Selecting receptive fields in deep networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2528–2536. Curran Associates, Inc., 2011.

[11] M. Crossley. *Essential Topology*. Springer Undergraduate Mathematics Series. Springer, 2006.

[12] S. de Medrano. *Involutions on Manifolds*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, 1971.

[13] I. Dryden and K. Mardia. *Statistical shape analysis*. Wiley series in probability and statistics. Wiley, 1998.

[14] H. G. Feichtinger and T. Strohmer, editors. *Gabor Analysis and Algorithms: Theory and Applications*. Birkhauser Boston, 1st edition, 1997.

[15] J. Gallier, D. Xu, and M. Siqueira. Parametric pseudo-manifolds. *Differential Geometry and its Applications*, 30(6):702–736, 2012.

[16] G. H. Golub, V. C. Klema, and G. W. Stewart. Rank degeneracy and least squares problems. Technical report, Stanford, CA, USA, 1976.

[17] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[18] J. J. Hunt, P. Dayan, and G. J. Goodhill. Sparse coding can predict primary visual cortex receptive field changes induced by abnormal visual input. *PLoS Comput Biol*, 9(5):e1003005, 05 2013.

[19] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Optimizing over radial kernels on compact manifolds. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 3802–3809, June 2014.

[20] S. Jayasumana, M. Salzmann, H. Li, and M. Harandi. A framework for shape analysis via hilbert space embedding. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 1249–1256, Washington, DC, USA, 2013. IEEE Computer Society.

[21] Y. Jia, C. Huang, , and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '12, pages 3370–3377, Los Alamitos, CA, USA, 2012. IEEE Computer Society.

[22] D. Kendall, D. Barden, T. Carne, and H. Le. *Shape and Shape Theory*. Wiley Series in Probability and Statistics. Wiley, 2009.

[23] D. G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

[24] D. G. Kendall. A survey of the statistical theory of shape. *Statist. Sci.*, 4(2):87–99, 05 1989.

[25] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proc. of the 27th International Conference on Machine Learning*, ICML '10, pages 543–550, 2010.

[26] S. Kurtek, E. Klassen, J. Gore, Z. Ding, and A. Srivastava. Elastic geodesic paths in shape space of parameterized surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1717–1730, Sept 2012.

[27] Y. lan Boureau, J. Ponce, and Y. Lecun. A theoretical analysis of feature pooling in visual recognition. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, ICML-10, pages 111–118. Omnipress, 2010.

[28] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 473–480, New York, NY, USA, 2007. ACM.

[29] J. M. Lee. *Introduction to smooth manifolds*. Graduate texts in mathematics. Springer, New York, Berlin, Heidelberg, 2nd edition, 2013.

[30] J. Lewi, R. Butera, D. M. Schneider, S. Woolley, and L. Paninski. Designing neurophysiology experiments to opti-

mally constrain receptive field models along parametric submanifolds. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 945–952. Curran Associates, Inc., 2009.

[31] H. Li, K. Zhang, and T. Jiang. Minimum entropy clustering and applications to gene expression analysis. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, CSB '04, pages 142–151, Washington, DC, USA, 2004. IEEE Computer Society.

[32] H. Ling and D. W. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):286–299, Feb 2007.

[33] J. Munkres. *Topology*. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2nd edition, 2000.

[34] N. Murray and F. Perronnin. Generalized max pooling. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 2473–2480. IEEE Computer Society, June 2014.

[35] G. Peyré. Manifold models for signals and images. *Comput. Vis. Image Underst.*, 113(2):249–260, Feb 2009.

[36] D. Raviv, R. Kimmel, and A. M. Bruckstein. Graph isomorphisms and automorphisms via spectral signatures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1985–1993, Aug 2013.

[37] C. Samir, A. Srivastava, M. Daoudi, and E. Klassen. An intrinsic framework for analysis of facial surfaces. *Int. J. Comput. Vision*, 82(1):80–95, Apr 2009.

[38] C. Small. *The Statistical Theory of Shape*. Springer Series in Statistics. Springer New York, 1996.

[39] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1415–1428, July 2011.

[40] K. B. Sun and B. J. Super. Classification of contour shapes using class segment sets. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '05, pages 727–733, Washington, DC, USA, 2005. IEEE Computer Society.

[41] T. P. Trappenberg, E. T. Rolls, and S. M. Stringer. Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes. In *Advances in Neural Information Processing Systems*, pages 293–300, Vancouver, British Columbia, Canada, Dec 2001. MIT Press.

[42] L. Tu. *An Introduction to Manifolds*. Springer, 1st edition, 2007.

[43] J. R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, 23(1):31–42, Jan 1976.

[44] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference*, BMVC '14, pages 1–11. BMVA Press, 2009.

[45] J. Winkeler, B. Manjunath, and S. Chandrasekaran. Subset selection for active object recognition. In *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2 of *CVPR '99*, pages 511–516, Ft. Collins, CO, USA, June 1999.

[46] Q. Xie, I. Jermyn, S. Kurtek, and A. Srivastava. Numerical inversion of srnfs for efficient elastic shape analysis of star-shaped objects. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 485–499. Springer International Publishing, 2014.

[47] C. Xu and N. Vasconcelos. Learning receptive fields for pooling from tensors of feature response. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 835–842. IEEE Computer Society, June 2014.

Supplementary Material
of
Compositional Formation of 2D Shape Models for Multi-component Shape Representation

Mete Ozay, Umit Rusen Aktas, Jeremy L. Wyatt and Aleš Leonardis
School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, United Kingdom.
{m.ozay, u.aktas, j.l.wyatt, a.Leonardis}@cs.bham.ac.uk

1. Introduction

In this document, we provide the supplemental material for the paper "Compositional Formation of 2D Shape Models for Multi-component Shape Representation". In the next section, the proofs of the theorems **Corollary 1** and **Theorem 2.2** proposed in the main text and the required definitions are given. In Section 3, additional experimental results are provided.

2. Proofs of the Theorems

In this section, we provide proofs of **Corollary 1** and **Theorem 2.2** given in the main text. The definitions which are given in the main text are recalled as they are employed in the proofs.

2.1. Proof of Corollary 1

Definition 2.1 (Part). *A part $\mathcal{P}$ is defined as*

$$\mathcal{P} = \{(A_j, \phi_j) | \phi_j : A_j \to U_j \subset \mathbb{R}^D, \forall j \in \mathfrak{I}\},$$

where $\mathfrak{I}$ is a nonempty countable set, ϕ_j is a homeomorphism from an open set A_j onto an open set U_j such that the composition of the function ϕ_j with its inverse ϕ_j^{-1} is the identity map id_{U_j}, i.e. $\phi_j \circ \phi_j^{-1} = id_{U_j}$, $\forall j \in \mathfrak{I}$.

Definition 2.2 (Shape Spaces). *Let us denote a matrix representation of the measurements as $X \in \mathbb{R}^{N_j \times D}$, where the i^{th} row of X is $\mathbf{x}_i \in A_j$, and $N_j = |A_j|$, $\forall j \in \mathfrak{I}$. A scaled and translated (ST) shape is defined as*

$$[X]_{st} = \{\beta X + I_{N_j}\gamma^T : \beta \in \mathbb{R}^+, \gamma \in \mathbb{R}^D\},$$

where $\beta \in \mathbb{R}^+$ is a scale variable used for isotropic scaling, I_{N_j} is an N_j dimensional unit vector and $\gamma \in \mathbb{R}^D$ is a translation vector. The ST space $\mathcal{S}_{st}$ is the space of all possible pre-shapes [3, 6].

The EST shape space $\mathcal{S}_{est}$ consists of all possible EST Shapes $[X]_{est}$ which are defined as

$$[X]_{est} = \{X_{st}\Gamma : \Gamma \in SO(D), X_{st} \in \mathcal{S}_{st}\},$$

where $SO(D)$ is the special orthogonal group which is the set of all $D \times D$ rotation matrices Γ with $\Gamma^T\Gamma = \Gamma\Gamma^T = I_D$, I_D is a $D \times D$ identity matrix and $\Gamma \in SO(D)$ is a rotation matrix [6]. A shape space $\mathcal{S}_{er}$ consists of all shapes which are defined as

$$[X]_{er} = \{X_{est}R : R \in O(D), X_{est} \in \mathcal{S}_{est}\}, \quad (1)$$

where $O(D)$ is the group of all $D \times D$ orthogonal matrices.

Definition 2.3 (Transition Maps). *Given any two charts (A_i, ϕ_i) and (A_j, ϕ_j) on $\mathcal{P}$, where $A_i \cap A_j \neq \varnothing$, the transition maps are defined as [4]*

$$\phi_{ji} : \phi_i(A_i \cap A_j) \to \phi_j(A_i \cap A_j),$$

$$\phi_{ij} : \phi_j(A_i \cap A_j) \to \phi_i(A_i \cap A_j),$$

$$\phi_{ji} = \phi_j \circ \phi_i^{-1} \text{ and } \phi_{ij} = \phi_i \circ \phi_j^{-1}, \forall i, j \in \mathfrak{I}.$$

Definition 2.4 (Measurements obtained from Receptive Fields). *A set of measurements which are obtained from Receptive Fields (RFs) that are employed on two dimensional images is defined as*

$$\mathfrak{RF} = \{\{U_i\}_{i\in\mathfrak{I}}, \{U_{ij}\}_{(i,j)\in\mathcal{L}}, \{\phi_{ji}\}_{(i,j)\in\mathcal{L}}\}, \quad (2)$$

where $\mathcal{L} = \{(i,j) \in \mathfrak{I} \times \mathfrak{I} : U_{ij} \neq \varnothing\}$ is an index set and $U_{ij} \triangleq \phi_i(A_i \cap A_j)$ which satisfy the following conditions:

1. *$\phi_{ii} = id_{U_i}$, where id_{U_i} is an identity map of U_i, $\forall i \in \mathfrak{I}$.*
2. *$\phi_{ij} = \phi_{ji}^{-1}$, $\forall (i,j) \in \mathcal{L}$.*

3. *If $U_{ji} \cap U_{jk} \neq \varnothing$, then $\phi_{ij}(U_{ji} \cap U_{jk}) = U_{ij} \cap U_{ik}$, $\forall i,j,k \in \mathfrak{I}$, and $\phi_{ki}(\mathbf{x}) = \phi_{kj} \circ \phi_{ji}(\mathbf{x})$, $\forall \mathbf{x} \in (U_{ij} \cap U_{ik})$.*

4. *There are open balls $V_{\mathbf{x}}$ and $V_{\mathbf{x}'}$ centered at $\mathbf{x}$ and $\mathbf{x}'$ such that no point of $V_{\mathbf{x}'} \cap U_{ji}$ is the image of any point of $V_{\mathbf{x}} \cap U_{ij}$ by ϕ_{ji}, $\forall (i,j) \in \mathcal{L}$ with $i \neq j$, and $\forall \mathbf{x} \in \partial(U_{ij}) \cap U_i$, $\forall \mathbf{x}' \in \partial(U_{ji}) \cap U_j$, where $\partial(U_{ij})$ is the boundary of U_{ij}.*

Definition 2.5 (Disjoint Union of Sets [7])**.** *Disjoint union of sets is defined by*

$$\begin{aligned} \mathbb{U} &\triangleq \coprod_{i \in \mathfrak{I}} U_i \\ &= \bigcup_{i \in \mathfrak{I}} \mathcal{U}_i, \end{aligned} \tag{3}$$

where $\mathcal{U}_i = U_i \times \{i\}$, $\forall i \in \mathfrak{I}$ [2, 9].

Definition 2.6 (Disjoint Union Topology)**.** *Suppose that $\{(U_i, \mathcal{T}_i)\}_{i \in \mathfrak{I}}$ is a family of topological spaces, and $\mathbb{U}$ is a disjoint union of sets U_i with topology $\mathcal{T}_i$, $\forall i \in \mathfrak{I}$. Then the disjoint union topology on $\mathbb{U}$ is defined as*

$$\mathcal{T} = \{V | V \subseteq \mathbb{U}, V \cap U_i \in \mathcal{T}_i, \forall i \in \mathfrak{I}\}.$$

In other words, a subset V of the disjoint union $\mathbb{U}$ is said to be open if and only if its intersection with each set $V \cap U_i$ is open in U_i, $\forall i \in \mathfrak{I}$ [7]. A disjoint union set $\mathbb{U}$ with the topology $\mathcal{T}$ is called a disjoint union space.

Definition 2.7 (Injection in Disjoint Union Topology)**.** *For each $i \in \mathfrak{I}$, there is an injection $\eta_i : U_i \hookrightarrow \mathbb{U}$ defined as $\eta_i(u) = (u, i)$, where $u \in U_i$ [7].*

Corollary 1. *$\mathcal{P}$ is a manifold whose co-domain and transition maps are defined by $\mathfrak{RF}$ with measurements on two dimensional images.*

Proof. Let's start by defining the orbit of X as

$$\begin{aligned} \mathfrak{o}(X) &\triangleq [X] \\ &= \{X_{er} : X_{er} \in \mathcal{S}_{er}\}, \end{aligned} \tag{4}$$

where $\mathcal{S}_{er}$ is defined in (1). Then, we obtain $\mathfrak{o} : \mathbb{U} \to \mathcal{P}$. Following the definition of the orbit by quotients, we can define an equivalence relation as

$$X \sim X' \Longleftrightarrow X' \in \mathfrak{o}(X).$$

We can employ this equivalence relation on $\mathbb{U}$ stating for all $X, X' \in \mathbb{U}$ that $X \sim X'$ iff $\exists (i,j) \in \mathcal{L}$, such that $X' = \phi_{ij}(X)$ for $X \in U_{ij}, X' \in U_{ji}$. Therefore, we have $\mathcal{P} = \mathbb{U}/\sim$.

Next, we employ the composition

$$t_i \triangleq \mathfrak{o} \circ \eta_i : U_i \to \mathcal{P}, \forall i \in \mathfrak{I}.$$

Since $A_i = t_i(U_i)$ and $\phi_i = t_i^{-1}$, we observe that (A_i, ϕ_i) is a chart of $\mathcal{P}$ for each $i \in \mathfrak{I}$ and their collection $\{(A_i, \phi_i)\}_{i \in \mathfrak{I}}$ form $\mathcal{P}$.

In order to prove that $\mathcal{P}$ is Hausdorff, we choose two orbits $[X] \in \mathcal{P}$ and $[X'] \in \mathcal{P}$ of any two elements $X \in U_i$ and $X' \in U_j$. Then, these measurements can be represented by either disjoint or overlapping models in $\mathcal{P}$. In the first case, $\tau_i(U_i) \cap \tau_j(U_j) = \varnothing$. Since τ_i and τ_j are homeomorphisms, $[X]$ and $[X']$ belong to disjoint sets. For $\tau_i(U_i) \cap \tau_j(U_j) \neq \varnothing$, we should consider four subcases;

1. If $i = j$, then X and X' are separated by disjoint open sets $V_{\mathbf{X}}$ and $V_{\mathbf{X}'}$. In addition, $[X]$ and $[X']$ are separated by disjoint open subsets $\tau_i(V_{\mathbf{X}})$ and $\tau_j(V_{\mathbf{X}'})$ since τ_i is a homeomorphism, $\forall i$.

2. If $i \neq j$, $X \in U_i - cl(U_{ij})$ and $X \in U_j - cl(U_{ji})$, where $cl(\cdot)$ is the set closure, then $[X]$ and $[X']$ are separated by the disjoint open sets $\tau_i(U_i - cl(U_{ij}))$ and $\tau_j(U_j - cl(U_{ji}))$.

3. If $i \neq j$, $X \in U_{ij}$, $X' \in U_{ji}$, $[X] \neq [X']$ and $[X'] \sim \phi_{ij}[X']$, then $[X] \neq \phi_{ij}(X')$.

4. If $i \neq j$, $X \in \partial(U_{ij}) \cap U_i$ and $X' \in \partial(U_{ji}) \cap U_j$, then $[X]$ and $[X']$ are separated by $\tau_i(V_{\mathbf{X}})$ and $\tau_j(V_{\mathbf{X}'})$.

Since there are countably many charts defined by the finite number of RFs, $\mathcal{P}$ is second countable. Additionally, ϕ_{ij} are the maps of $\mathcal{P}$ since $\phi_i = \tau_i^{-1}$ and $\phi_{ji} = \phi_j \circ \phi_i^{-1}$, $\forall i,j \in \mathfrak{I}$ and $\forall (i,j) \in \mathcal{L}$.

□

In **Corollary 1**, we show that a part $\mathcal{P}$ is a manifold. Next, we show that the composition of parts is a manifold under the disjoint union topology.

2.2. Proof of Theorem 2.2

Lemma 2.1 (**Proposition 3.42** in [7])**.** *Let $S = \{U_i\}_{i \in \mathfrak{I}}$ be an indexed family of topological spaces, and $\mathbb{U} = \coprod_{i \in \mathfrak{I}} U_i$.*

1. *If each $U_i \in S$ is second countable and the index set A is countable, then $\mathbb{U}$ is second countable.*

2. *If each $U_i \in S$ is Hausdorff, then so is $\mathbb{U}$.*

3. *If each $U_i \in S$ is first countable, then so is $\mathbb{U}$.*

4. *Each injection $\eta_i : U_i \hookrightarrow \mathbb{U}$, $\forall i \in \mathfrak{I}$, is a topological embedding, open and closed map.*

5. *A subset of $\mathbb{U}$ is closed iff its intersection with each U_i, $i \in \mathfrak{I}$, is closed.*

Theorem 2.2. *Given a collection of parts $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$, where $\mathfrak{N}$ is a set of indices, a composition $\mathcal{C}$ of parts defined as*

$$\mathcal{C} = \coprod_{n\in\mathfrak{N}} \mathcal{P}_n \tag{5}$$

is a manifold.

Proof. Following the propositions 1 and 2 in **Lemma 2.1** given above, $\mathcal{C}$ is *second countable* and *Hausdorff* since each $\mathcal{P}_n$ is a manifold by **Corollary 1**.

Since each $\mathcal{P}_n$ is a manifold, $\mathcal{C}$ is *first countable* by 3 in **Lemma 2.1**. In other words, each point $\pi \in \mathcal{P}_n$ has a countable basis of neighborhoods, therefore $\pi \in V$, where V is an open set.

By 4 and 5 in **Lemma 2.1**, there exist homeomorphisms $\psi_n : V_n \to W_n$, $\forall n \in \mathfrak{N}$, where W_n is an open set of $\mathcal{C}$. Since each $\mathcal{P}_n$ is locally Euclidean, $\forall n \in \mathfrak{N}$, and following the results, $\mathcal{C}$ is locally Euclidean.

□

3. Experiments

In this document, we provide additional experimental results of the proposed algorithms introduced in the main text using the Animals Dataset [1]. The properties of the Animals dataset, implementation details and parameter selection methods are given in the main text. In the experiments, the images were resized to 128 × 128 pixels using bicubic interpolation [5]. The results are provided for four different cases each of which is denoted by **MC-Z**, where **Z** $\in \{1, 2, 3, 4\}$ represents the employment of statistical learning algorithms in the shape spaces $\mathcal{S}_{est}$, $\mathcal{S}_{er}$, $\mathcal{S}_r$ and $\mathcal{S}_a$, respectively. The definitions of the shape spaces and the details of the proposed algorithms are given in the main text.

In order to analyze the changes in the performance of the proposed algorithms according to the multi-component structure of the shapes observed in the images, the experiments are performed on noisy images. Shape-based noise was added to the images obtained from the Animals dataset using three different geometric structures as described in the following:

- **Oval-shaped Noise (ON):** Oval-shaped geometric structures were generated using a Cartesian model called *Oval of Descartes* [8]. The parameters of the model were sampled uniformly to generate an oval-shaped structural element (SE) according to the constraint that the whole SE would fit in the image.

- **Polygonal-shaped Noise (PON):** In this dataset, we used Polygonal-shaped geometric structures. The vertex coordinates of the polygonal-shaped SEs were sampled uniformly between 1 and 128.

- **Rectangular-shaped Noise (RAN):** We used rectangular-shaped geometric structures whose height and width values were sampled uniformly between 4 and 64.

After the SEs were created, their interior regions were filled with zero-valued pixels. Then, the SEs were centered on an image at an image coordinate $\mathbf{x} = (x_1, x_2)$ where x_1 and x_2 were sampled uniformly according to the constraint that the SEs would fit in the image. A sample generated by a structure was rejected if the number of pixels in the regions covered by the SEs was less than 16 or greater than 4096. A new structure was generated to corrupt each image in the dataset. Sample images generated using **ON**, **PON** and **RAN** are given in Figure 1, Figure 2 and Figure 3, respectively.

Using the proposed method, we obtain shapes with multiple components. For example, we observe a shape of a bird with two disconnected components depicted with white pixels each of which represents the body and the tail of the bird in Figure 1.a, respectively. In addition, a shape of a flying bird is decomposed into four disconnected components in Figure 1.p, and a shape of a horse is decomposed into three disconnected components in Figure 1.r.

Following the experimental setup explained in the main text, 50 images per class were randomly chosen for training, and the rest of the samples were used for testing. The procedure was repeated 10 times, and the mean values of the classification performances were computed. The performances are analyzed for the cases when $\aleph = 1$ and $\aleph = 2$ structures are employed on the images for each dataset in Table 1. In the first row, the results on the original dataset, which are provided in the main text, are also given. We observe similar classification performances across different noisy image datasets. However, classification performance decreases as the number of structures used for corrupting shapes increases.

Table 1: Classification performance (% Mean) of the algorithms using the images obtained from the Animals Dataset.

Datasets	$\aleph$	**MC-1**	**MC-2**	**MC-3**	**MC-4**
Original Dataset		85.1	86.4	77.9	81.1
ON	1	76.8	78.1	73.0	75.8
	2	67.1	67.3	64.1	65.7
PON	1	77.1	77.6	73.9	76.3
	2	66.5	67.9	64.8	67.8
RAN	1	77.2	78.3	72.4	77.1
	2	67.4	68.2	65.7	68.0

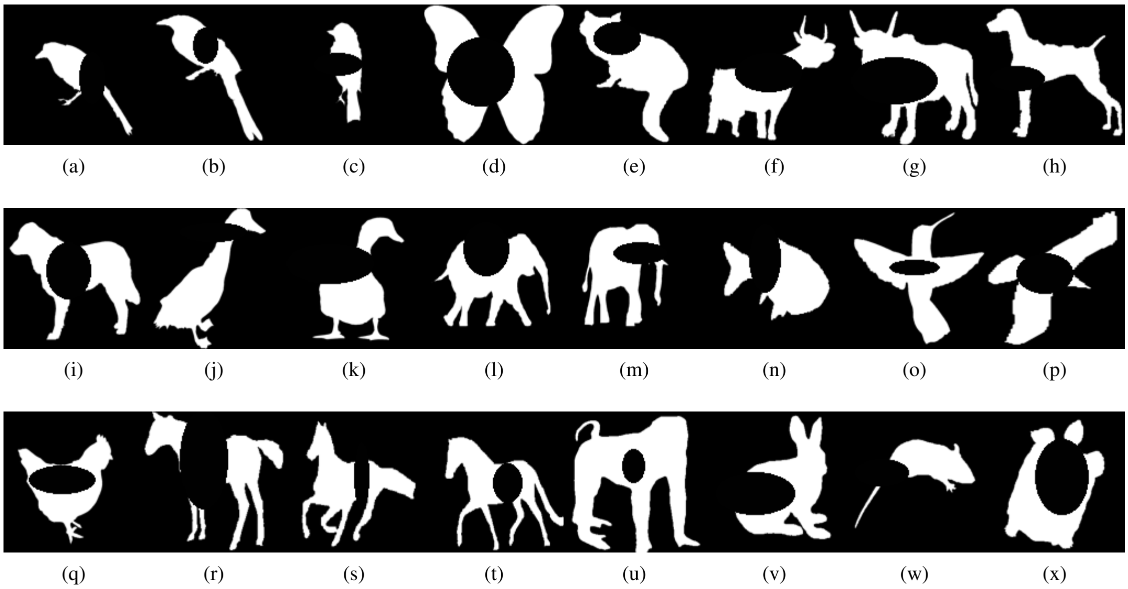


Figure 1: Sample images belonging to the **ON** dataset which contain shapes that are corrupted by randomly placed oval-shaped structural elements.

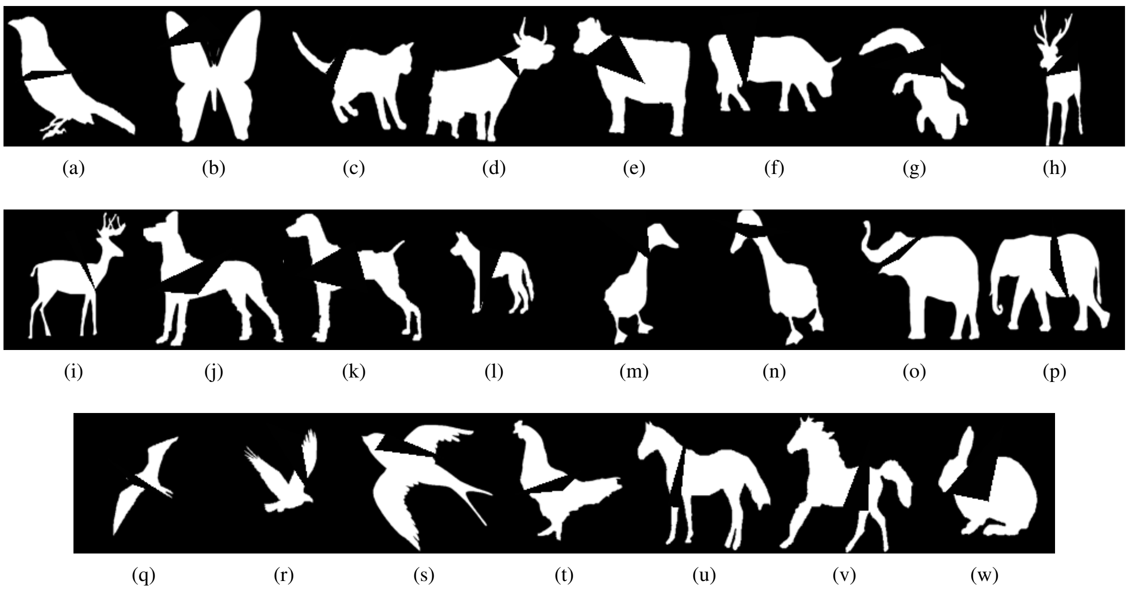


Figure 2: Sample images belonging to the **PON** dataset which contain shapes that are corrupted by randomly placed polygonal-shaped structural elements.

References

[1] X. Bai, W. Liu, and Z. Tu. Integrating contour and skeleton for shape classification. In *Proceedings of the 2009 IEEE International Conference on Computer Vision*, ICCV '09, pages 360–367, Sept 2009. 3

[2] M. Crossley. *Essential Topology*. Springer Undergraduate Mathematics Series. Springer, 2006. 2

[3] I. Dryden and K. Mardia. *Statistical shape analysis*. Wiley series in probability and statistics. Wiley, 1998. 1

[4] J. Gallier, D. Xu, and M. Siqueira. Parametric pseudo-manifolds. *Differential Geometry and its Applications*,

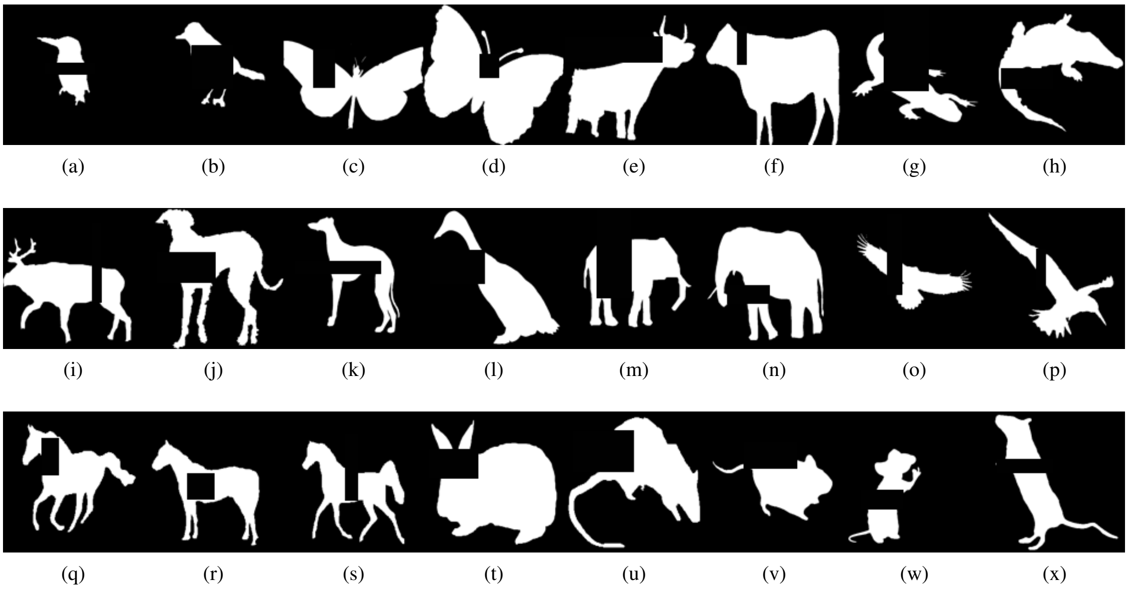


Figure 3: Sample images belonging to the **RAN** dataset which contain shapes that are corrupted by randomly placed rectangular-shaped structural elements.

30(6):702–736, 2012. 1

[5] R. C. Gonzalez and R. E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006. 3

[6] D. Kendall, D. Barden, T. Carne, and H. Le. *Shape and Shape Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. 1

[7] J. M. Lee. *Introduction to topological manifolds*. Graduate texts in mathematics. Springer, New York, Berlin, Heidelberg, 2nd edition, 2011. 2

[8] E. H. Lockwood. *A book of curves*. Cambridge: Cambridge University Press, 2007. 3

[9] J. Munkres. *Topology*. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2nd edition, 2000. 2

Statistical Learning of Geometric Transformations for Modeling Multi-component Shape Representations

Mete Ozay, *Member, IEEE,* U. Rusen Aktas, Jeremy L. Wyatt, and Aleš Leonardis, *Member, IEEE.*

Abstract

We introduce a theoretical framework for statistical learning of geometric transformations for modeling multi-component shape representations in compositional hierarchies. For this purpose, we address the problem of analysis of shapes of objects with multiple components $i)$ in non-manifold measurement spaces at the first $n = 1$ layer of a hierarchy, and $ii)$ in some well-known shape spaces which are not manifolds at some set of points that represent parts and their compositions constructed at the higher layers, $n > 1$, of the hierarchy.

The first problem is considered by reformulating the relationship between data and shape spaces considering the interaction between Receptive Fields (RFs) and shape models in a shape vocabulary. Then, we suggest a method to design RFs considering both the topological structure of the models and statistical properties of the data. Problem $ii)$ is resolved by designing a disjoint union topology using an indexing mechanism for the formation of shape models on shape manifolds in the vocabulary, recursively. In addition, we suggest an approach to design inhibition methods in order to remove redundant measurements avoiding degenerate models in the vocabulary. The proposed approach is used to analyze different invariant properties of the shape spaces such as invariance to Euclidean Similarity Transformations (EST), Reflection(R), EST-R and Affine Transformation. In addition, results on benchmark shape classification outperform state-of-the-art shape representation methods, shape classification algorithms and hierarchical learning methods.

Index Terms

Shape representation, classification, rigid body transformation, affine transformation, compositional hierarchies

✦

• *M. Ozay, U. R. Aktas, J. L. Wyatt and A. Leonardis are with the School of Computer Science, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom.*
E-mail: { m.ozay, xxa334, a.Leonardis, j.l.wyatt } @cs.bham.ac.uk

1 INTRODUCTION

Modeling feature representations which are invariant to geometric transformations has been a challenge and studied in computer vision, pattern recognition and machine learning literature for a long decade [8], [9], [13], [21], [31], [34], [37], [41], [45], [49], [50], [52], [57], [58], [59], [61], [67]. Statistical learning of geometric transformations have been one of the successful approaches for modeling invariant representations [13], [21], [34], [37], [41], [45], [52], [58], [61]. In addition, geometric invariance properties of representations have been used and employed for statistical learning of recognition models using less number of samples, i.e. as $N \to 1$ [3], and in order to boost recognition performance [21], [37], [52].

In the recent works [3], [8], [34], [55], translation and scale invariant hierarchical representations have been modeled using Gabor features as base-layer features, and max-pooling operations. In this work, we analyze reflectance, rigid-body and affine transformation invariant properties of hierarchical representations for shape recognition in a theoretical framework by introducing a hierarchical compositional shape representation approach to the Statistical Shape Theory (SST).

In the SST, shapes are represented in shape spaces which are usually considered as manifolds. This assumption fails if the shape spaces are not manifolds because of the *local singularities* that occur in the *shape spaces* [6], [16], [29], [29], [56]. In addition to this observation reported in the literature [6], [29], [56], we address two cases where the manifold assumption of data and shape spaces fails. First, the singularities may be observed in some components of representations because of the representation of landmarks or configurations using *redundant measurements* obtained from the images. Second, shape spaces may consist of multiple disconnected components, therefore the corresponding *shape manifolds* may be *disconnected* and non-smooth [36]. In this case, the methods which model shape deformations in connected shape manifolds may fail [54].

In the multi-component shape representation problem, we address the aforementioned cases in which the manifold assumptions of the state-of-the-art methods fail because of the multiple disconnected component structures of measurements and shape spaces. In other words, we provide a framework where the manifold assumptions can be relaxed using multi-component approximations in order to model shape representations. In this work, we address the problem of design and construction of shape representation models on manifolds when i) data and shape spaces are not manifolds and/or ii) the data and shape manifolds contain multiple disconnected components.

In order to solve the first problem, the relationship between data and shape spaces is reformulated considering the interaction between Receptive Fields (RFs) and shape models in a shape vocabulary. For this purpose, we first define an RF as a co-domain of a shape manifold that is represented by a shape model in a shape vocabulary. Then, we suggest a method to design structures of RFs considering both the statistical properties of the data, geometric structure of the shape manifolds, and topological structure of the models. We theoretically analyze the conditions that should be satisfied by an indexing mechanism to construct components of shape representations with manifold structures in shape vocabularies using the proposed method, even if the measurements are obtained from overlapping RFs.

A disjoint union topology [14] of components and compositions of representations is designed using an indexing mechanism for the recursive formation of shape models on shape manifolds in the vocabulary. For this purpose, we introduce a mathematical framework to implement the indexing mechanisms for i) designing inhibition methods to remove redundant measurements in order to avoid degenerate models in the vocabulary, and ii) the employment of the vocabulary for structural shape classification. The proposed approach enables us to construct different shape representations with various invariant properties using different distance functions for statistical learning of models, and employ topological shape manifolds in the models.

In the next section, we provide the related work and background. In the Section 3, we first analyze the aforementioned approaches and methods used for Multi-component (MC) shape representation problem in SST. Then, we propose a new approach to resolve the challenges of their assumptions by re-defining the structures used for shape representation and developing mathematical models of inhibition processes, receptive fields, indexing mechanisms and model-based shape representations. In Section 4, we suggest a method to employ a structural classification algorithm using the proposed shape representation approach. In Section 5, the proposed methods are examined on benchmark datasets to perform different tasks in SST such as the construction of shape models equipped with different invariant properties for representation and classification of multi-component shapes observed in images. The results are compared with the state-of-the-art methods. Section 6 concludes the paper.

2 Related Work and Background

Prior works in SST used landmarks or configurations to represent geometric information of *local* structures of shapes [16], [27], [29], [56]. In this approach, *shape* of an object is defined as all the geometrical information that remains when location, scale and rotational attributes are filtered out from the object [16], [28]. Mathematically speaking, shape of an object is a representation which is invariant under the Euclidean Similarity Transformations (EST) [16].

In order to construct EST-invariant representations, first a set of landmarks or primitive features is extracted from the measurements on objects, e.g. images [27]. Then, a scale and translation invariant representation, called *pre-shape*, is obtained using a set of centered unit-size landmarks. The complete set of all such pre-shapes is called *pre-shape space*, and its members lie on a unit sphere S_D^N [29]. Finally, an equivalence class (orbit) of the landmarks is computed under action of the special orthogonal group $SO(D)$, which is the set of all $D \times D$ rotation matrices Γ with $\Gamma^T \Gamma = \Gamma \Gamma^T = I_D$, where I_D is a $D \times D$ identity matrix [29]. The elements of the orbits reside in a space Σ_D^N called *shape space*. For $D = 2$,

Σ_2^N is a compact smooth manifold, or *shape manifold*, and a shape is represented as a point in $\mathbb{C}P^{N-2}$ or a complex line in $\mathbb{C}^{N-1}$. Since the geometry of shape manifolds is non-trivial, development of recognition and statistical inference algorithms on shape manifolds is a challenge [24].

The methods which employ an *explicit* analysis on shape manifolds first embed the manifolds into some Euclidean spaces [6], [29]. Then, statistical learning and inference algorithms are implemented on the embedded submanifolds [6]. Kernel methods were employed in the implicitly mapped and embedded submanifolds [23], [24]. However, computation of the mappings is challenging because of the kernel parameter and type estimation problems. Statistical learning and optimization methods are employed on shape manifolds for *intrinsic* analysis of shape representations [1], [6], [29]. Statistical analysis of manifolds has been performed by (i) first mapping shape structures to tangent spaces, (ii) then computing the statistical properties on tangent spaces, and finally (iii) mapping the computed statistical models into shape manifolds [9], [31], [59]. However, the mappings may not exist, and the assumption of the algorithms to construct the models using the mappings may fail. Moreover, the mappings may not be invertible or unique [67]. The algorithms proposed for computing the mappings have several limitations such that the suggested representations and metrics may not be invariant to translations, and they lack of physical interpretations [31].

Note that, the manifold assumption of the aforementioned algorithms regarding the structure of Σ_D^N may fail in cases where Σ_D^N is not a manifold. For instance, multiple length minimizing geodesics between the *redundant* shape representations are observed on Σ_3^N [6], [29]. In order to solve the singularity problem, Riemannian submersion is employed on the non-singular part of S_D^N [6], [29], [56]. Although removing the *redundant* representations from pre-shape spaces and constructing shape spaces using the non-redundant representations provide a Riemannian manifold, it is not complete. Therefore, intrinsic analysis cannot be employed on the obtained shape manifold, and the methods which use explicit analysis are implemented to provide approximate solutions [6], [23], [24].

Shape representation algorithms which employ statistical analysis of manifolds on data obtained from images assume that the data lie on a manifold [51]. In order to obtain *manifold-like* data spaces, measurements are made on images using data structures that are approximations to local and open subsets such as local patches, sliding windows, or Receptive Fields (RFs) [51]. Therefore, the measurements obtained from RFs are considered as manifolds, and shape representation algorithms are employed on the data based on this assumption. However, the manifold interpretation of RFs may fail in some applications [38]. This problem has been defined as the determination of the structure of RFs and selection of their parameters [12], [22], [26], [32], [48], [62], [68].

3 MULTI-COMPONENT SHAPE REPRESENTATION

Multi-component (MC) shape representation problem consists of three sub-problems, namely i) measurement problem, ii) statistical learning of shape distributions, and iii) construction of topological shape manifolds. The methods which solve the MC problem are considered in a compositional shape formation framework.

3.1 Measurement Problem

We will consider the design and utilization of RFs for the formation of shape models considering the statistical properties of the data, and topological properties of the shape manifolds. For instance, we consider a scenario where data measurements are obtained from an RF depicted with blue square region to model a manifold representation of a shape of the number 8 given in Fig. 1.a. However, the data space of the shape is not a manifold. If we make measurements from another RF with different size and structure depicted by the red circular region to obtain the data manifold in Fig. 1.b, we also observe that the red cross $\times$ is not a manifold. The reason is: the subspace topology is not

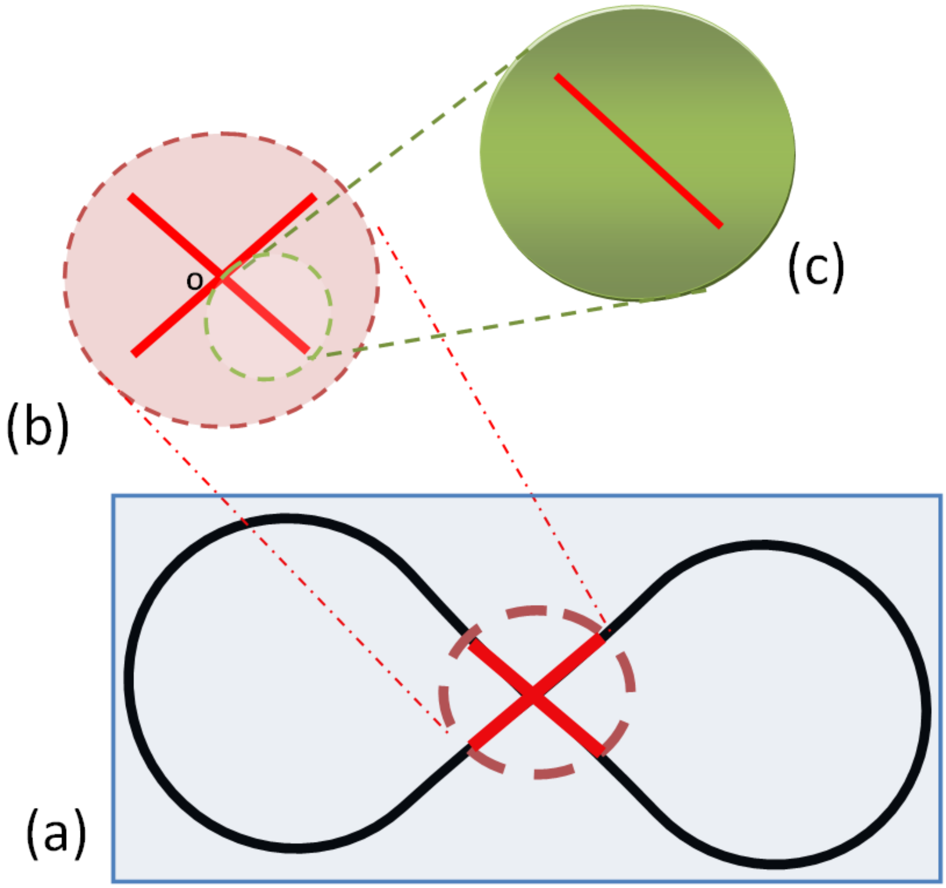


Fig. 1: Measurement of shapes in Receptive Fields depicted by blue, red and green regions. The 8 shape given in (a) is not a manifold, and the cross (X) in (b) is not a manifold. However, the line in (c) is a one dimensional manifold.

locally Euclidean at the intersection point $\mathbf{O}$. In other words, if we consider a neighborhood U of $\mathbf{O}$, then we observe that $U - \{\mathbf{O}\}$ has four connected components which prevent the existence of a homeomorphism from U to an open ball $\mathcal{B}(\mathbf{x}, r) = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < r\} \subset \mathbb{R}^D$, where $d(\cdot, \cdot)$ is a metric and $r \in \mathbb{R}^D$ [63]. On the other hand, each component of Fig. 1.b, one of which is a line depicted in Fig. 1.c, is a one dimensional manifold.

As we observed in this example, we cannot assure the representation of shapes on shape manifolds by using mappings from data spaces obtained from different RFs with different structures and sizes, since the mappings may not be homeomorphism or their parametrization may not exist. We address this problem by considering a shape space as a manifold, and the measurements obtained from an RF as a co-domain of the manifold.

In our proposed approach, we assume that a set of measurements are obtained from an open ball $\mathcal{B}(\mathbf{x}_c, r) = \{\mathbf{x}' : \|\mathbf{x}_c - \mathbf{x}'\|_2 < r, r \in \mathbb{R}, \mathbf{x}' \in A\}$ which is represented by a Receptive Field centered at $\mathbf{x}_c \in \mathbb{R}^D$, where $A \in \mathbb{R}^D$ is an open subset. In this work, we use two-dimensional images, therefore $D = 2$ will be considered in the rest of the paper, if not stated explicitly. Additionally, we assume that we extract landmarks as Gabor features extracted from images at each image point $\mathbf{x} \in \mathbb{R}^2$ using a two-dimensional Gabor kernel $G(\mathbf{x}, \mathcal{R})$, where $\mathcal{R}$ is a set of parameters [17]. Assuming that the parameters of the Gabor functions are fixed, we propose a method to select the size or structure of the RFs as follows. We define the intensity value of the pixel at $\mathbf{x}$ by $I(\mathbf{x})$. The Gabor filter $G(\mathbf{x})$ and $I(\mathbf{x})$ are convolved using the measurements obtained from an RF centered at $\mathbf{x}_c$ as

$$\mathfrak{C}(\mathbf{x}_c, r) = \mathfrak{F}^{-1}\{\mathfrak{F}(I(\mathbf{x})) \odot \mathfrak{F}(G(\mathbf{x}))\}, \forall \mathbf{x}_c,$$

where $\mathfrak{F}(I(\mathbf{x}))$ is the Fourier transform of $I(\mathbf{x})$ and $\odot$ is the point-wise multiplication. Then, we estimate an optimal radius $\hat{r}$ of the ball $\mathcal{B}(\mathbf{x}_c, r)$ as

$$\hat{r} = \arg\max_{r}\{\frac{1}{2r} \arg\max_{\mathbf{x}_c} \mathfrak{C}(\mathbf{x}_c, r)\}, \forall \mathbf{x}_c,$$

where $\frac{1}{2r}$ is used following the circular structure of the RF. This method was previously suggested for learning receptive fields using three-dimensional tensors in images [68]. In our proposed approach, we employ the method for designing open ball shaped RFs in two-dimensional images. Due to the co-domain interpretation of the proposed RFs, our implementation is faster by order of number of scales since the RFs are used to obtain scale invariant manifold representations which is explained in the next section.

Selection of the RFs using this approach provides us the features with maximum response in a neighborhood or ball $\mathcal{B}(\mathbf{x}_c, r)$, $\forall \mathbf{x}_c$. However, redundant measurements may be obtained from the regions in which the RFs are overlapped. In order to remove the redundancy, pooling [26], [32], weighted selection [48] or clustering [68] methods have been employed on the set of measurements. These methods aim to remove the redundancy according to the statistical properties [32], [12] of the data independent of the requirements of the representation models due to the nonlinearity of the relationship between the models and the measurements.

3.2 Multi-component Shape Representation Model

We consider a Receptive Field (RF) as a co-domain from which the measurements are obtained within open sets using open balls. The co-domain U determined by an RF is defined together with a manifold

$$\mathcal{P} = \{(A_j, \phi_j) | \phi_j : A_j \to U_j \subset \mathbb{R}^D, \forall j \in \mathfrak{I}\},$$

where $\mathfrak{I}$ is a nonempty countable set. We call the manifold $\mathcal{P}$ a *part*, and (A_j, ϕ_j) is a chart of the $\mathcal{P}$, $\forall j \in \mathfrak{I}$. In SST, a part $\mathcal{P}$ can be considered as an element of a shape space, which is also designed to be a manifold. Therefore, we should first show that $\mathcal{P}$ is a manifold which is a non-degenerate representation model. Following the shape space construction approach used in SST and discussed in the previous section, parts will be constructed using equivalence relations between shape representation models through quotient maps of data and shape spaces.

3.2.1 *Geometry of Shape Representation Models*

In order to employ group operations on the measurements, we denote a matrix representation of the measurements as $X \in \mathbb{R}^{N_j \times D}$, where the i^{th} row of X is $\mathbf{x}_i \in A_j$, and $N_j = |A_j|$, $\forall j \in \mathfrak{I}$. Then, we extend the Euclidean Similarity Transformation (EST) model used in SST [16] to obtain EST and reflection invariant (EST-R) representations. In addition, we suggest an inhibition method to define non-degenerate shape models in different shape spaces.

For this purpose, we first define a scaled and translated (ST) shape as

$$[X]_{st} = \{\beta X + I_D \gamma^T : \beta \in \mathbb{R}, \gamma \in \mathbb{R}^D\},$$

where $\beta \in \mathbb{R}^+$ is a scale variable used for isotropic scaling, and $\gamma \in \mathbb{R}^D$ is a translation vector. Note that the ST space $\mathcal{S}_{st}$, which is the space of all possible pre-shapes, is scale and translation invariant [16], [29].

Inhibition: In order to assure that the ST space $\mathcal{S}_{st}$ is a manifold, we should satisfy the requirement that $rank(X_c) \geq D$, where $X_c \in \mathcal{S}_{st}$. Selection of κ number of independent columns (rows) of a matrix has been studied as a subset selection problem [19], [20], [66]. In a recent work, a sparse encoding algorithm is proposed to provide an approximate solution to the problem [48]. In this work, first the generalized pooling method of [48] is extended with the constraint $D \leq \kappa \leq N$. Then the problem is solved using a feature split distributed optimization algorithm [7] considering the recursive structural relationship between co-domains of parts, parts and their compositions.

In order to obtain EST invariant representations, we employ actions from the rotation group as

$$[X]_{est} = \{X_{st}\Gamma : \Gamma \in SO(D), X_{st} \in \mathcal{S}_{st}\},$$

where $\Gamma \in SO(D)$ is a rotation matrix. The EST shape space consisting of all possible EST Shapes $[X]_{est}$ is denoted by $\mathcal{S}_{est}$.

Note that, if $D > 2$, then the shape space consisting of elements of $\mathcal{S}_{est}$ is degenerate. In order to remove the degeneracy, the parts which have rank-deficient measurements are removed from the space [6]. Although this operation provides a Riemannian manifold, the elements are disconnected and the space is separated into multiple components [6]. Since we consider $D = 2$ in this work, we assure that the space of $\mathcal{S}_{est}$ is non-degenerate. In order to solve this problem, the action of manifold involution is employed on the $\mathcal{S}_{est}$ [6], [15], [29]. Then, we obtain a reflection shape space $\mathcal{S}_{er}$ consisting of all reflection shapes which are defined as

$$[X]_{er} = \{X_{est}R : R \in O(D), X_{est} \in \mathcal{S}_{est}\}, \tag{1}$$

where $O(D)$ is the group of all $D \times D$ orthogonal matrices. In order to detect and remove the degenerate representations in the $\mathcal{S}_{er}$ space, we should check the rank deficiency property of the measurements $X_{er} \in \mathcal{S}_{er}$. For this purpose we employ the proposed inhibition method.

Once the degenerate representations are removed, we obtain a compact differentiable manifold, and can employ extrinsic and intrinsic analysis on the manifold for statistical learning of shape representations [6]. The distance functions of representations that will be employed in the learning algorithms are designed according to the geometric and topological structure of the spaces. Therefore, we will analyze the geometric and topological properties of shape spaces in the following subsection before introducing the methods proposed for statistical learning of representations.

3.2.2 Topological Properties of Shape Manifolds

We have shown that a part $\mathcal{P}$ is a manifold using quotient maps to construct EST and reflection invariant representations. Now, we will analyze the conditions which enable us to model the manifold structure of a part $\mathcal{P}$ when there is an overlap among the charts of the $\mathcal{P}$, and among the corresponding RFs in the measurement spaces. Given any two charts (A_i, ϕ_i) and (A_j, ϕ_j) on $\mathcal{P}$, where $A_i \cap A_j \neq \varnothing$, the transition maps are defined as [18]

$$\phi_{ji} : \phi_i(A_i \cap A_j) \to \phi_j(A_i \cap A_j),$$

$$\phi_{ij} : \phi_j(A_i \cap A_j) \to \phi_i(A_i \cap A_j),$$

$$\phi_{ji} = \phi_j \circ \phi_i^{-1} \text{ and } \phi_{ij} = \phi_i \circ \phi_j^{-1}, \forall i, j \in \mathfrak{I}.$$

Defining $U_{ij} = \phi_i(A_i \cap A_j)$ and $U_{ji} = \phi_j(A_i \cap A_j)$, we compute a map $\phi_{ji} : U_{ij} \to U_{ji}$ which is called a *gluing map* in [18]. Then, a set of gluing measurements, which is a superset of RFs used on the images, is defined as [18]

$$\mathfrak{RF} = \{\{U_i\}_{i \in \mathfrak{I}}, \{U_{ij}\}_{(i,j) \in \mathcal{L}}, \{\phi_{ji}\}_{(i,j) \in \mathcal{L}}\}, \tag{2}$$

where $\mathcal{L} = \{(i,j) \in \mathfrak{I} \times \mathfrak{I} : U_{ij} \neq \varnothing\}$ is an *index set*. In order to assure that there is a manifold with co-domain and maps defined by $\mathfrak{RF}$, the maps should satisfy the following conditions [18]:

1) $\phi_{ii} = id_{U_i}$, where id_{U_i} is an identity map of U_i, $\forall i \in \mathfrak{I}$.
2) $\phi_{ij} = \phi_{ji}^{-1}$, $\forall (i,j) \in \mathcal{L}$.
3) If $U_{ji} \cap U_{jk} \neq \varnothing$, then $\phi_{ij}(U_{ji} \cap U_{jk}) = U_{ij} \cap U_{ik}$, $\forall i,j,k \in \mathfrak{I}$, and $\phi_{ki}(\mathbf{x}) = \phi_{kj} \circ \phi_{ji}(\mathbf{x})$, $\forall \mathbf{x} \in (U_{ij} \cap U_{ik})$.
4) There are open balls $V_{\mathbf{x}}$ and $V_{\mathbf{x}'}$ centered at $\mathbf{x}$ and $\mathbf{x}'$ such that no point of $V_{\mathbf{x}'} \cap U_{ji}$ is the image of any point of $V_{\mathbf{x}} \cap U_{ij}$ by ϕ_{ji}, $\forall (i,j) \in \mathcal{L}$ with $i \neq j$, and $\forall \mathbf{x} \in \partial(U_{ij}) \cap U_i$, $\forall \mathbf{x}' \in \partial(U_{ji}) \cap U_j$, where $\partial(U_{ij})$ is the boundary of U_{ij}.

In our proposed approach, centering and normalization map of features used to construct ST spaces is a transition map induced by an *indexing mechanism* that satisfies the conditions defined above. Next, we provide a theorem which associates the measurement spaces defined by RFs to D dimensional manifolds.

Theorem 3.1 **([18]).** *For every set $\mathfrak{RF}$ defined in (11) there is a D dimensional manifold whose transition maps are $\{\phi_{ji}\}_{(i,j) \in \mathcal{L}}$, where $\mathcal{L}$ is an index set. More precisely, there exists a manifold $\mathfrak{M} = (\mathfrak{RF}, \{\psi\}_{i \in \mathfrak{I}})$, where $\psi_i : U_i \to \mathbb{R}^D$ is defined as*

$$\psi_i \triangleq \psi_j \circ \phi_{ji}, \forall (i,j) \in \mathcal{L}.$$

Corollary 1. *$\mathcal{P}$ is a manifold whose co-domain and transition maps are defined by $\mathfrak{RF}$ with measurements on two dimensional images.*

The formal proof of the **Corollary 1** follows the proof idea of the **Theorem 1** which is given in the Appendix. We will provide two crucial ideas of the proof in order to support our motivation of employing quotient maps and indexing mechanisms below.

Proof idea of the Corollary 1: The proof idea is first i) forming the disjoint union of sets in the co-domain, and then identifying overlapping sets with transition maps which are ii) defined by quotient maps, and iii) traced by indexing.

Disjoint union of sets is defined by

$$\mathbb{U} \triangleq \coprod_{i \in \mathfrak{I}} U_i = \bigcup_{i \in \mathfrak{I}} \mathcal{U}_i, \tag{3}$$

where $\mathcal{U}_i = U_i \times \{i\}$, $\forall i \in \mathfrak{I}$ [14], [47]. Disjoint union of sets is a collection of sets where the indices of the elements are recorded by the *indexing mechanism*. In other words, a disjoint topology requires the employment of a set of indices of the elements using an indexing mechanism. Moreover, the mathematical structure of indexing mechanism is crucial for the proof. Let's start by defining the orbit of X as

$$\mathfrak{o}(X) = [X] = \{X_{er} : X_{er} \in \mathcal{S}_{er}\},$$

where $\mathcal{S}_{er}$ is defined in (10). Then, we obtain $\mathfrak{o} : \mathbb{U} \to \mathcal{P}$. Following the definition of the orbit by quotients, we can define an equivalence relation as $X \sim X'$ iff $X' \in \mathfrak{o}(X)$. We can employ this equivalence relation on $\mathbb{U}$ stating for all $X, X' \in \mathbb{U}$ that $X \sim X'$ iff $\exists (i,j) \in \mathcal{L}$, such that $X' = \phi_{ij}(X)$ for $X \in U_{ij}, X' \in U_{ji}$. Therefore, we have $\mathcal{P} = \mathbb{U}/\sim$.

Next, we define a natural injection $\eta_i : U_i \hookrightarrow \mathbb{U}, \forall i \in \mathfrak{I}$ which will be used for indexing and defining

$$t_i = \mathfrak{o} \circ \eta_i : U_i \to \mathcal{P}.$$

Since $A_i = t_i(U_i)$ and $\phi_i = t_i^{-1}$, we observe that (A_i, ϕ_i) is a chart of $\mathcal{P}$ and their collection form $\mathcal{P}$, $\forall i \in \mathfrak{I}$.

In order to prove that $\mathcal{P}$ is Hausdorff, we choose two orbits $[X] \in \mathcal{P}$ and $[X'] \in \mathcal{P}$ of any two elements $X \in U_i$ and $X' \in U_j$, $\forall i, j \in \mathfrak{I}$. Then, these measurements can be represented by either disjoint or overlapping models in $\mathcal{P}$. Since there are countably many charts defined by the finite number of RFs, $\mathcal{P}$ is second countable. Additionally, ϕ_{ij} are the maps of $\mathcal{P}$ since $\phi_i = \tau_i^{-1}$ and $\phi_{ji} = \phi_j \circ \phi_i^{-1}$, $\forall i, j \in \mathfrak{I}$, $\forall (i,j) \in \mathcal{L}$. □

In **Corollary 1**, we show that a part $\mathcal{P}$ is a manifold. Next, we propose that their composition is a manifold.

Theorem 3.2. *Given a collection of parts $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$, where $\mathfrak{N}$ is a set of indices, a composition $\mathcal{C}$ of parts defined as*

$$\mathcal{C} = \coprod_{n\in\mathfrak{N}} \mathcal{P}_n \tag{4}$$

is a manifold.

The formal proof of the **Theorem 2.2** is given in the appendix in Section 7.2.

3.2.3 Statistical Learning of Representations

Let Q be a probability distribution on $\mathcal{C}$, and $\{\Omega_n\}_{n\in\mathfrak{N}}$ are random variables associated to $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$ and drawn independent and identically distributed (i.i.d.) from Q. In the statistical learning model, we aim to compute the joint observation probability of two parts with different part indices or ids in locally defined open sets, which are clusters. In other words, we compute a set of disjoint clusters $\Lambda_k \subset \mathcal{C}$, $\forall k = 1, 2, \dots, K$, such that $\bigcup_{k=1}^{K} \Lambda_k = \mathcal{C}$. Moreover, we require that if $\Lambda_k \subseteq \Lambda_l$, then $\sigma(\Lambda_k) \leq \sigma(\Lambda_l)$, where $\sigma(\cdot)$ is a cost function used to avoid obtaining subsets of clusters. This condition is defined and used as the Monotone Clustering Property (MCP) for Intracluster criterion-based clustering in [11]. We define the cost function of an instance of a cluster as

$$\sigma(\lambda_k, \omega_n, \omega_m) = \Pr(\lambda_k \in \Lambda_k | \omega_n \in \Omega_n, \omega_m \in \Omega_m) \tag{5}$$

which is the probability of observing two given part instances $\omega_n \in \Omega_n$ and $\omega_m \in \Omega_m$ in a cluster instance $\lambda_k \in \Lambda_k$, $\forall k$. Computation of a collection of clusters which maximize (5) is considered as a minimum entropy clustering problem [11], [39] and defined as

$$\mathcal{K} = \arg\min_{\mathfrak{K}} \sum_{\omega_n\in\Omega_n} \sum_{\omega_m\in\Omega_m} \sum_{\lambda_k\in\Lambda_k} \Pr{}^{\alpha}(\lambda_k | \omega_n, \omega_m) - 1, \tag{6}$$

where $\mathfrak{K}$ is the set of all possible partitions of the dataset among K clusters. We choose $0 < \alpha < 1$ unlike the clustering problem proposed in [2] to satisfy the MCP criterion. In a clustering algorithm [39] which computes (6), a part $\mathcal{P}_m$ that resides in the neighborhood of a given part $\mathcal{P}_n$ is selected to minimize the conditional entropy associated with the conditional probability defined in (5). In the computation of the neighborhood of $\mathcal{P}_n$, we use different distance functions by employing extrinsic analysis on the shape manifold $\mathcal{C}$. We used $\mathcal{S}_{est}$, $\mathcal{S}_{er}$, the Reflection Space $\mathcal{S}_r$ and the Affine Space $\mathcal{S}_a$ [6] in the experiments.

3.2.4 Compositional Formation of Hierarchy of Shape Manifolds

In the proposed approach, we construct parts $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$, and the shape manifold $\mathcal{C}$ employing a disjoint union operation [14]. In addition, we integrate a set of shape manifolds $\{\mathcal{C}_m\}_{m\in\mathfrak{M}}$ to construct a shape vocabulary $\Upsilon = \coprod_{m\in\mathfrak{M}} \mathcal{C}_m$, where $\mathfrak{M}$ is a nonempty countable set of indices. Therefore, we consider a compositional approach for the formation of the vocabulary Υ.

We represent a shape manifold $\mathcal{C}^l_m$ constructed at the l^{th} layer by a graph $\mathcal{G}^l_m = (\mathcal{V}^l_m, \mathcal{E}^l_m)$, $\forall m \in \mathfrak{M}$. $\mathcal{V}^l_m$ is the set of nodes where each node $v^{l+1}_i = n$ represents an index label or id $n \in \mathfrak{N}$ of a subpart $\mathcal{P}^l_n$, $\forall m$.

At $l = 1$, we have disconnected parts, therefore $e_{ij} = \varnothing$, $\forall e_{ij} \in \mathcal{E}^1_m$, $\forall m$. For $l > 1$, if two parts $\mathcal{P}^l_n$ and $\mathcal{P}^l_{n'}$ co-occur in a cluster λ_k with probability $\sigma(\lambda_k, \omega_n, \omega_{n'}) > 0$, $v^l_i = n$ and $v^l_j = n'$, then $e_{ij} = k$. Following the recursive definition of the vocabulary, a shape manifold $\mathcal{C}^l_m$ is considered as a part of another shape manifold $\mathcal{C}^{l+1}$ which is referred as a *composition* constructed at the $(l+1)^{st}$ layer. At each layer l, we can compare two graph representations $\mathcal{G}^l_m$ and $\mathcal{G}^l_{m'}$ considering the manifold structure of shape spaces. Given two graphs $\mathcal{G}^l_m$ and $\mathcal{G}^l_{m'}$, deciding whether they are isomorphic is defined as a graph isomorphism problem [53], [64].

At the first layer $l = 1$ in the vocabulary, we assure the existence of a bijection $\mathfrak{b} : \mathcal{V}^l_n \to \mathcal{V}^l_{n'}$ between two graph representations $\mathcal{G}^l_n$ and $\mathcal{G}^l_{n'}$ of two parts $\mathcal{P}^l_n$ and $\mathcal{P}^l_{n'}$ in a shape manifold $\mathcal{C}^{l+1}_m$, respectively. In other words, $\mathcal{G}^l_n \sim \mathcal{G}^l_{n'}$ iff $\mathcal{G}^l_{n'} \in \mathfrak{b}(\mathcal{G}^l_n)$, $\forall n, n' \in \mathfrak{N}$ following the compositional formation of parts as described in the previous section.

For $l > 1$, we search and select the graphs $\mathcal{G}^l_n$ of the parts $\mathcal{P}^l_n$ that are subgraphs and isomorphic to the graphs $\mathcal{G}^{l+1}_m$ of compositions $\mathcal{C}^{l+1}_m$, $\forall m \in \mathfrak{M}$. This problem has been defined and studied as a sub-graph isomorphism problem [2], [53], [64]. In this work, we select the subgraphs of compositions $\mathcal{C}_m^{l+1}$, $\forall m \in \mathfrak{M}$ solving the sub-graph isomorphism problem using the implementation provided in [2]. We search the shape manifolds and increment the hierarchy until we obtain a compact connected shape manifold which contains a single component at the highest layer of the hierarchy.

3.3 Inference of Shape Representations on Test Data

In the inference algorithm, the indexing mechanism proposed in the previous section is employed to induce a disjoint union topology in the inference tree $\mathcal{T}$. We first extract a set of landmarks or features from a test image. Then, we represent the test features as parts $\{\check{\mathcal{P}}_n^l\}_{n \in \mathfrak{N}}$ and shape manifolds $\{\check{\mathcal{C}}_m^l\}_{m \in \mathfrak{M}}$ at the first layer $l = 1$ of an inference tree $\mathcal{T}$ such that $\check{\mathcal{C}}_m^1 = \coprod_{n \in \mathfrak{N}} \check{\mathcal{P}}_n^1$, $\forall m \in \mathfrak{M}$.

At each consecutive layer $l > 1$, we search for a set of graphs $\{\check{\mathcal{G}}_u^l : u \in \mathfrak{U}\}$ whose members are graph isomorphic to the shape manifolds $\mathcal{G}_n^l$, $\forall n \in \mathfrak{N}$ in the vocabulary Υ. More precisely, we first compute a set of all graphs $iso(\mathcal{G}_n^l)$ which are isomorphic to $\mathcal{G}_n^l$ $\forall n$, and then search the set of parts $\check{\mathcal{G}}_u^l \in iso(\mathcal{G}_n^l)$, $\forall u \in \mathfrak{U}$, where $\mathfrak{U}$ is a nonempty countable set of indices. In order to obtain an approximate solution to the graph isomorphism problem, we used the inexact graph matching implementation provided in [2].

4 Shape Classification using Multi-component Shape Representations

In the shape classification problem, we consider *contribution* of each part, composition and shape manifold in the vocabulary to the representation of classes using a top-down approach. For this purpose, we employ the topological structure of the vocabulary Υ which is induced by the indexing mechanism, such as the recursive representation of parts and compositions at each layer $l > 1$ of Υ.

After shape manifolds are formed in a hierarchical vocabulary, we first compute training and test inference trees $\mathfrak{T} = \{\mathcal{T}_n : n = 1, 2, \ldots, N\}$ and $\check{\mathcal{T}}$ using a set of N training images and a given test image, respectively. Let $F \in \mathbb{R}^{N \times H}$ denote an input matrix whose row vector $\mathbf{f}_n$ represents the detection of a feature by a part or a composition in Υ on the n^{th} image provided by an inference tree $\mathcal{T}_n$. We encode the class labels using 1-of-W coding of classes using a matrix $Y \in \mathbb{R}^{N \times W}$, where $y_{nw} = 1$ if the n^{th} image belongs to the w^{th} class, and $y_{nw} = 0$, otherwise. In addition, we assume a linear class model for each class variable as

$$\mathbf{y}_w = F\mathbf{z}_w + \epsilon_w, \forall w = 1, 2, \ldots, W,$$

where $\mathbf{z}_w \in \mathbb{R}^H$ is a vector of model coefficients, and ϵ_w is a vector that represents model error. Then the class model estimation problem is defined as

$$\hat{Z} = \arg\min \Big(\sum_{w=1}^{W} \big(\mathbf{y}_w - F\mathbf{z}_w\big)^T \big(\mathbf{y}_w - F\mathbf{z}_w\big) + \mathfrak{L} \Big), \tag{7}$$

where $\hat{Z}$ is the matrix of estimated model variables, and $\mathfrak{L}$ is a structural loss function. We use two methods to solve (7)

4.1 Structural Shape Classification using Tree Lasso

The structural loss function $\mathfrak{L}$ is defined as

$$\mathfrak{L} \triangleq \xi \sum_{h=1}^{H} \sum_{\upsilon_i \in \Lambda_{int}} \mu_{\upsilon_i} \|\mathbf{z}^h_{\Phi_{\upsilon_i}}\|_2 + \xi \sum_{h=1}^{H} \sum_{\upsilon_j \in \Lambda_{leaf}} \mu_{\upsilon_j} \|\mathbf{z}^h_{\Phi_{\upsilon_j}}\|_2, \tag{8}$$

where ξ is a regularization parameter. Λ_{leaf} is the number of leaf nodes in an inference tree which represent the detections by parts in Υ. Λ_{int} is the number of internal nodes in an inference tree which represent the detections by either parts or compositions in Υ according to the topology induced by the indexing mechanism. In other words, Λ_{int} consists of shared parts and compositions among different objects and different layers, and defines the overlapping groups. μ_{υ_i} and μ_{υ_j} are the regularization weights associated to each node $\upsilon_i \in \Lambda_{int}$ and $\upsilon_j \in \Lambda_{leaf}$.

Φ_{υ_i} and Φ_{υ_j} are group structures whose members are the indices of leaf nodes of the subtrees rooted at the nodes $\upsilon_i \in \Lambda_{int}$ and $\upsilon_j \in \Lambda_{leaf}$. Using the group structures encoded by the indexing mechanism, values of features detected at the corresponding parts and compositions are encoded in $\mathbf{z}^h_{\Phi_{\upsilon_i}}$ and $\mathbf{z}^h_{\Phi_{\upsilon_j}}$, which are the vectors of variables $z^h_{\varrho_i}, \forall \varrho_i \in \Phi_{\upsilon_i}$, $\forall \upsilon_i \in \Lambda_{int}$ and $z^h_{\varrho_j}, \forall \varrho_j \in \Phi_{\upsilon_j}$, $\forall \upsilon_j \in \Lambda_{int}$ [30]. The problem (7) is defined as a tree guided group lasso problem in [30]. In this work, we solve the problem for an ensemble of trees induced by the learned shape vocabulary.

In the experiments, we used the implementation of a Tree-Guided Group Lasso (TL) algorithm to solve the recursively defined problem (7) using (8) provided in [30].

4.2 Structural Shape Classification using Overlapping Lasso

The TL algorithm considers the contribution of each composition and part to the classification model. In order to increase the convergence rate of the algorithm, we relax the lasso penalty on the leaf nodes using the following loss function

$$\mathfrak{L}_o \triangleq \xi_1 \|\bar{\mathbf{z}}\|_1 + \xi_2 \sum_{h=1}^{H} \sum_{\upsilon \in \Lambda} \mu_\upsilon \|\mathbf{z}^h_{\Phi_\upsilon}\|_2, \tag{9}$$

where $\|\cdot\|_1$ is the ℓ_1 norm, ξ_1 and ξ_2 are regularization parameters on the ℓ_1 and ℓ_2 norms, $\bar{\mathbf{z}}$ is the class model vector whose variables are $\mathbf{z}^h_{\Phi_\upsilon}$, $h = 1, 2, \ldots, H$ and $\upsilon \in \Lambda$. Then, the problem (7) with $\mathfrak{L}_o$ is solved using the Fast Overlapping Lasso Algorithm (FOL) given in [69].

Note that TL employs the recursive structure of the vocabulary, such that a composition represented by an internal node can be considered as a root of the group defined at a layer $l > 1$, or a member (part) of another group whose root is at another layer $l' > l$. However, FOL employs the group structures considering only the root nodes, i.e. the compositions constructed at the top layer of the tree.

5 EXPERIMENTS

Landmark points are defined by features which are extracted from the images using Gabor Filters whose parameters are selected using the Information Diagram method as suggested in [17], [46]. We have used an implementation of the MEC algorithm given in [39] for statistical learning. We used the algorithms and implementation provided by [2] to solve Graph and Subgraph Isomorphism problems. Shape vocabularies are learned in different shape spaces, and classification models are learned using FOL and TL algorithms. We used the implementations of the FOL and TL algorithms given in [69] and [30], respectively. The algorithm parameters are selected as suggested in [69] and [30].

In the tables, the results are given for eight different cases and each case is denoted by **MC-$\beth$-∇**. $\beth \in \{\mathrm{O}, \mathrm{T}\}$, where O and T represents the implementation of the FOL and TL algorithm, respectively. $\nabla \in \{1, 2, 3, 4\}$ represents the statistical learning in the shape spaces $\mathcal{S}_{est}$, $\mathcal{S}_{er}$, $\mathcal{S}_r$ and $\mathcal{S}_a$, respectively.

TABLE 1: Classification error (%) of the proposed and hierarchical shape classification algorithms on the CS Dataset.

SR [33]	**SP [33]**	**DBN-1 [33]**	**SAA-3 [33]**	**DBN-3 [33]**	**DAE2 [10]**	**CAE2 [10]**	**mDAE2 [10]**
19.13	19.82	19.92	18.41	18.63	18.44	19.30	18.10
MC-O-1	**MC-O-2**	**MC-O-3**	**MC-O-4**	**MC-T-1**	**MC-T-2**	**MC-T-3**	**MC-T-4**
18.19	16.56	17.42	16.92	17.10	15.71	17.38	15.16

5.1 Experimental Analysis on Artificial Datasets

In this section, we examine the suggested approach and the algorithms on three different benchmark artificial datasets. The datasets are generated such that the shapes consisting of single and/or multiple components are corrupted by rigid body transformations and noise [33].

TABLE 2: Classification error (%) of the proposed and hierarchical shape classification algorithms on the M-R Dataset.

SR [33]	**SP** [33]	**DBN-1** [33]	**SAA-3** [33]	**DBN-3** [33]	**DAE2** [10]	**CAE2** [10]	**mDAE2** [10]
10.38	13.61	12.11	11.43	12.3	11.94	13.62	10.36
MC-O-1	**MC-O-2**	**MC-O-3**	**MC-O-4**	**MC-T-1**	**MC-T-2**	**MC-T-3**	**MC-T-4**
9.18	10.04	9.64	10.58	10.24	11.78	10.70	13.86

TABLE 3: Classification error (%) of the proposed and hierarchical shape classification algorithms on the M-RBI Dataset.

SR [33]	**SP** [33]	**DBN-1** [33]	**SAA-3** [33]	**DBN-3** [33]	**DAE2** [10]	**CAE2** [10]	**mDAE2** [10]
32.62	37.59	31.84	24.09	28.51	44.92	48.25	46.12
MC-O-1	**MC-O-2**	**MC-O-3**	**MC-O-4**	**MC-T-1**	**MC-T-2**	**MC-T-3**	**MC-T-4**
22.04	21.25	30.69	32.75	20.42	19.47	34.40	31.37

Convex Shapes (**CS**) dataset [33] consists of 8000 training and 50000 test images each of which contains convex and non-convex regions, and belong to two classes. In this dataset, the classification problem is defined as a binary classification of convex and non-convex regions. Mnist-Rotation (**M-R**) dataset [33] consists of 12000 training and 50000 test images of the Mnist digits which belong to ten classes and are rotated by an angle generated uniformly between 0 and 2π radians. In the Mnist-RBI (**M-RBI**) dataset [33], a random patch from a black and white image is used as the background for the rotated images obtained from the Mnist-Rotation dataset. The size of each image is 28×28. The proposed algorithms are compared with Deep Belief Networks with three layers (**DBN-3**), Denoising (**DAE-2**), Contractive (**CAE-2**) and marginalized Denoising auto-enconder (**mDAE-2**) algorithms with two layers.

The results given in Table **??** show that the proposed algorithms outperform the state-of-the-art algorithms. In addition, we observe that the proposed algorithms can employ invariant properties of shape manifolds on shape representations. For instance, affine invariant shape spaces are used in **MC-4** for encoding information about convexity of shapes. Therefore, the best classification performance is observed in **MC-4** on the **CS** dataset.

In addition, we observe that EST invariant shape manifolds used in **MC-1** provide better representations of rotated shapes obtained from the **M-R** dataset than the other shape spaces. An interesting observation is that the proposed methods which model representations in a more complex shape space $\mathcal{S}_{er}$ in **MC-2** outperform the algorithms in noisy images which contain corrupted shapes belonging to the **M-RBI** dataset. The reason of performance boost of the algorithms which use $\mathcal{S}_{er}$ is that the symmetry patterns observed in the background images are recognized and removed from the shape representations by the proposed inhibition and graph compression algorithms.

In order to analyze the changes in the performance of the proposed algorithms according to the multi-component structure of the shapes observed in the images, the experiments are performed on noisy images. Shape-based noise was added to the images obtained from the Animals dataset using three different geometric structures as described in the following:

- **Oval-shaped Noise (ON):** Oval-shaped geometric structures were generated using a Cartesian model called *Oval of Descartes* [44]. The parameters of the model were sampled uniformly to generate an oval-shaped structural element (SE) according to the constraint that the whole SE would fit in the image.
- **Polygonal-shaped Noise (PON):** In this dataset, we used Polygonal-shaped geometric structures. The vertex coordinates of the polygonal-shaped SEs were sampled uniformly between 1 and 128.
- **Rectangular-shaped Noise (RAN):** We used rectangular-shaped geometric structures whose height and width values were sampled uniformly between 4 and 64.

After the SEs were created, their interior regions were filled with zero-valued pixels. Then, the SEs were centered on an image at an image coordinate $\mathbf{x} = (x_1, x_2)$ where x_1 and x_2 were sampled uniformly according to the constraint that the SEs would fit in the image. A sample generated by a structure was rejected if the number of pixels in the regions covered by the SEs was less than 16 or greater than 4096. A new structure was generated to corrupt each image in the dataset. Sample images generated using **ON**, **PON** and **RAN** are given in Figure 2, Figure 3 and Figure 4, respectively.

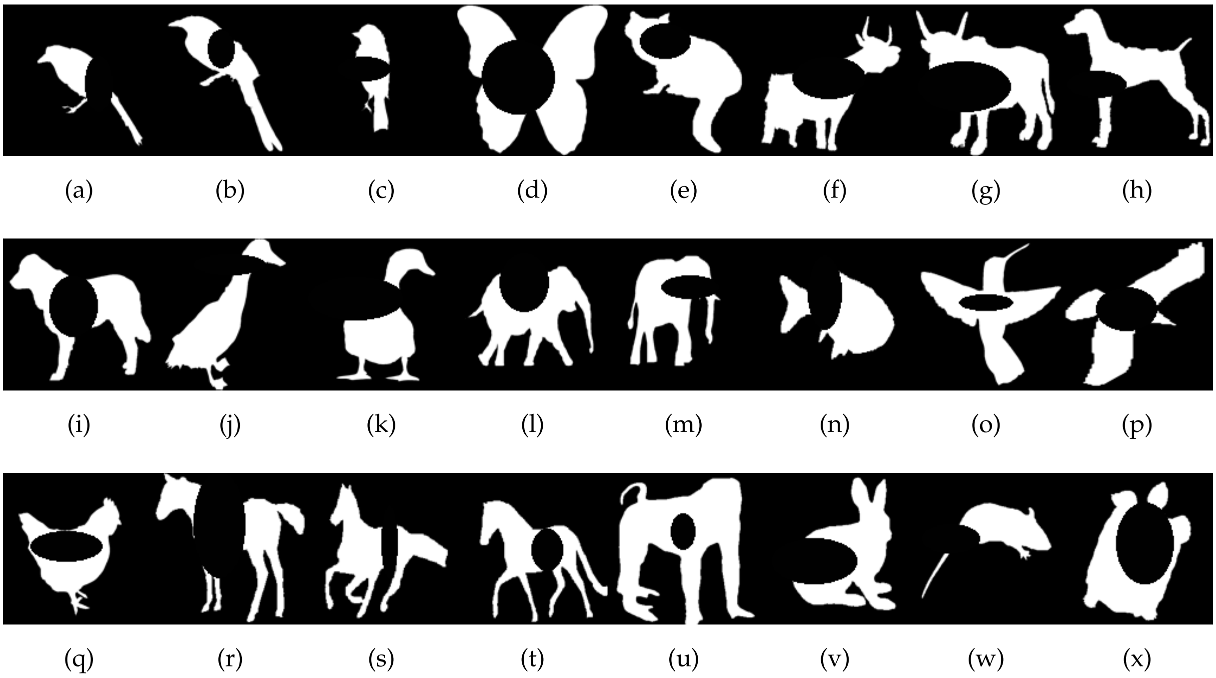


Fig. 2: Sample images belonging to the **ON** dataset which contain shapes that are corrupted by randomly placed oval-shaped structural elements.

Using the proposed method, we obtain shapes with multiple components. For example, we observe a shape of a bird with two disconnected components depicted with white pixels each of which represents the body and the tail of the bird in Figure 2.a, respectively. In addition, a shape of a flying bird is decomposed into four disconnected components in Figure 2.p, and a shape of a horse is decomposed into three disconnected components in Figure 2.r.

Following the experimental setup explained in the main text, 50 images per class were randomly chosen for training, and the rest of the samples were used for testing. The procedure was repeated 10 times, and the mean values of the classification performances were computed. The performances are analyzed for the cases when $\aleph = 1$ and $\aleph = 2$ structures are employed on the images for each dataset in Table 5. We observe similar classification performances across different noisy image datasets. However, classification performance decreases as the number of structures used for corrupting shapes increases.

TABLE 4: Classification performance (% Mean) of the algorithms using the images obtained from the Animals Dataset.

Datasets	$\aleph$	MC-1	MC-2	MC-3	MC-4
Original Dataset		85.1	86.4	77.9	81.1
ON	1	76.8	78.1	73.0	75.8
	2	67.1	67.3	64.1	65.7
PON	1	77.1	77.6	73.9	76.3
	2	66.5	67.9	64.8	67.8
RAN	1	77.2	78.3	72.4	77.1
	2	67.4	68.2	65.7	68.0

5.2 Classification Results on Real-world Datasets

In this section, we provide classification results of the algorithms using two benchmark real-world datasets. The Animals Dataset [4] consists of 2000 shapes belonging to 20 categories. We have used this dataset instead of the MPEG-7 dataset since the Animals dataset contains more images that are deformed with more articulated and non-rigid transformations, and larger class-variation than the MPEG-7 dataset [5]. Following the data setup suggested in [5], 50 images

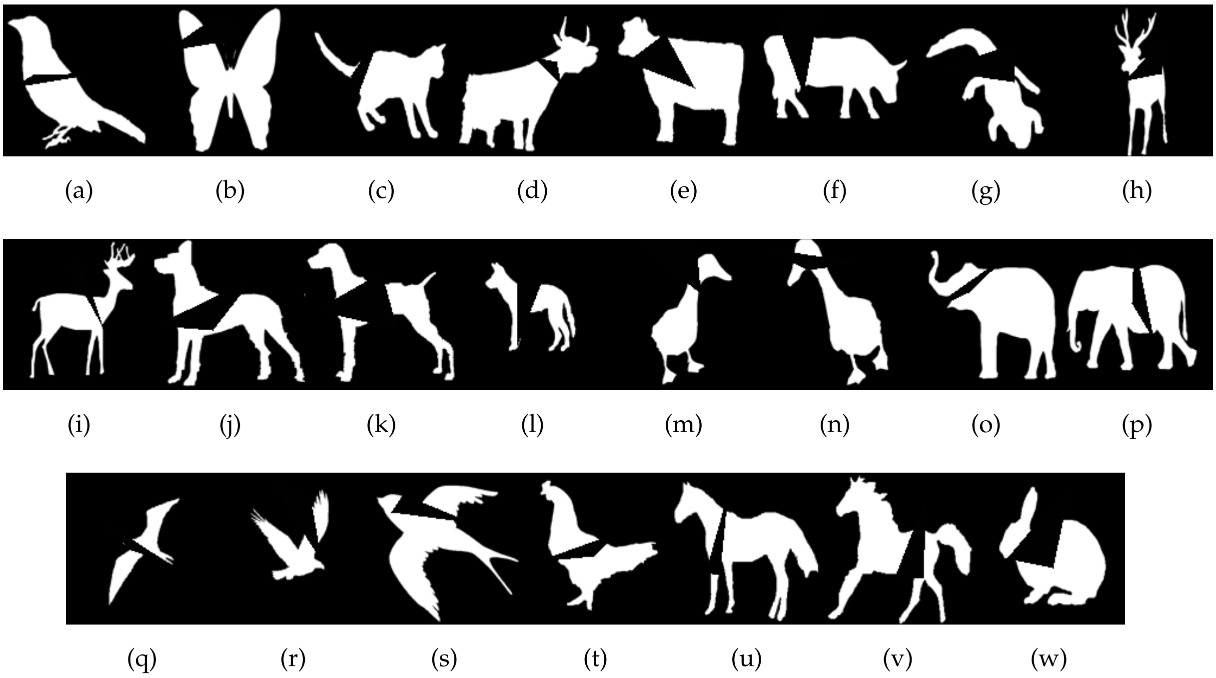


Fig. 3: Sample images belonging to the **PON** dataset which contain shapes that are corrupted by randomly placed polygonal-shaped structural elements.

per class were randomly chosen for training, and the rest of the samples were used for testing. The procedure is repeated 10 times for the proposed algorithms and the results are compared with the performances of the state-of-the-art shape descriptors.

The results given in Table 5 show that the average performance values of the proposed algorithms are greater and the variance of values are less than the values of the other algorithms. Note that similar variance values are obtained for the proposed algorithms. In addition, we report the performances of the proposed algorithms over 10 sets of experiments unlike 100 sets of experiments reported for the **NN-SVM** [5]. Therefore, the higher variance can be attributed to the smaller variance of the data generated in the experimental setup.

In the second set of the experiments, we examined the classification performance of the proposed algorithms with the state-of-the-art shape classification

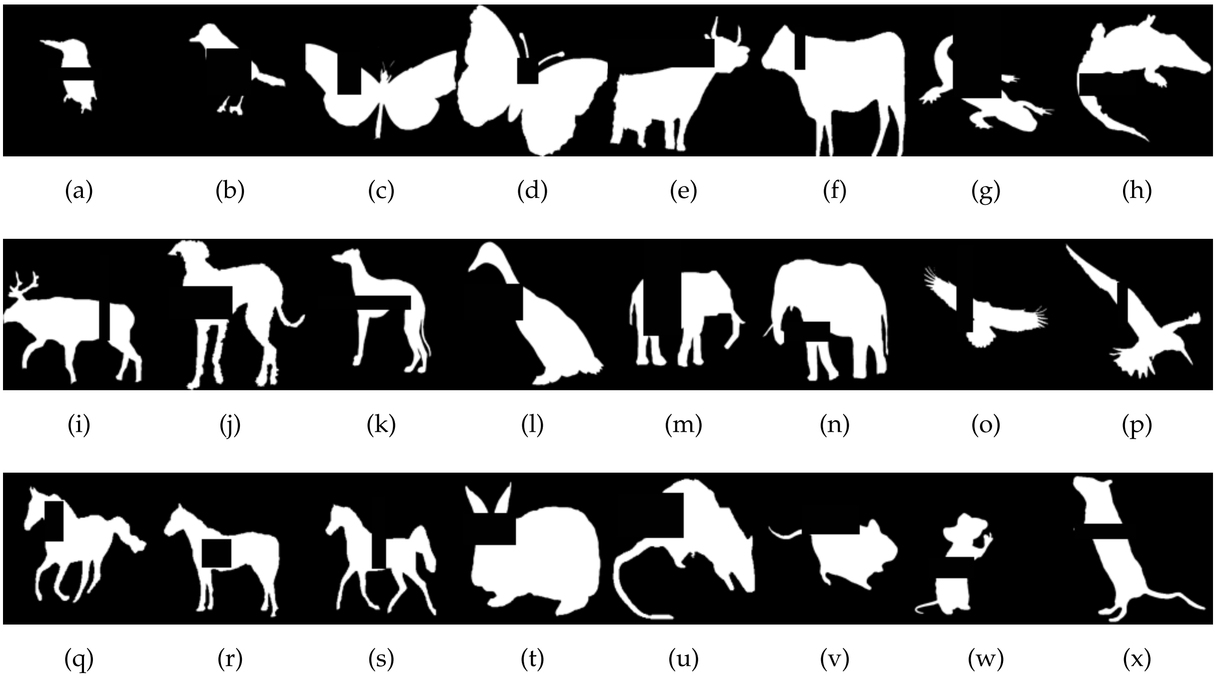


Fig. 4: Sample images belonging to the **RAN** dataset which contain shapes that are corrupted by randomly placed rectangular-shaped structural elements.

algorithms that are implemented on shape manifolds using the Butterfly Dataset [65]. The dataset contains 823 images belonging to 10 classes. Following the data setup suggested in [23], we randomly selected 40 shapes from each class for training and used the rest for testing. The procedure was repeated 10 times, and mean and variance of the classification performance values are given in Table 6.

In the real world datasets which contain shapes that are deformed by articulation and non-rigid body transformations, the algorithms need to model higher variance of the geometric patterns distributed among different measurements, objects and classes. Therefore, the algorithms which use shape spaces with *richer* geometric properties perform better than the other algorithms on these datasets.

For instance, **S-kFP** and **MKL** employ full Procrustes (FP) on the Kendall's shape space [23]. Similarly, we have used the FP distance for statistical learning on the $\mathcal{S}_{est}$ which corresponds to the Kendall's shape in **MC-1**. Meanwhile, Grassmann manifolds are used in **SM**. Since our proposed method considers the contribution of each part and composition to the classification model, we obtain better classification performances. Moreover, we extended the $\mathcal{S}_{est}$ with manifold involution in **MC-2** obtaining a *richer* shape space. Therefore, we obtained *more complex* representations and better classification performances.

TABLE 5: Classification performance (% Mean ± Variance) of proposed algorithms and the shape descriptors on the Animals Dataset.

CSS [60]	**IDSC** [43]	**CS** [4]	**CSSP** [4]	**HST** [40]	**Lim** [42]	**NN-TIP** [5]	**NN-SVM** [5]
69.7	73.6	71.7	78.4	80	80.4	77.42 ± 1.2	84.30 ± 1.0
MC-T-1	**MC-T-2**	**MC-T-3**	**MC-T-4**	**MC-O-1**	**MC-O-2**	**MC-O-3**	**MC-O-4**
85.1 ± 2.3	86.4 ± 2.1	77.9 ± 2.3	81.1 ± 2.3	83.3 ± 1.8	83.6 ± 1.7	74.1 ± 1.8	79.9 ± 1.8

TABLE 6: Classification performance (% Mean ± Variance) of the state of the art algorithms employed on shape manifolds, and the proposed algorithms using the Butterfly Dataset.

S-kFP [23]	**S-kFPG** [25]	**MKL** [23]	**SM** [23]
57.75 ± 2.0	60.37 ± 1.6	60.84 ± 2.0	63.98 ± 1.6
MC-O-1	**MC-O-2**	**MC-O-3**	**MC-O-4**
64.12 ± 2.1	65.05 ± 1.8	60.65 ± 2.5	60.19 ± 27
MC-T-1	**MC-T-2**	**MC-T-3**	**MC-T-4**
64.81 ± 1.8	67.13 ± 1.4	61.34 ± 2.2	61.57 ± 2.1

6 Conclusion

We has been proposed a new approach for design and construction of representations on manifolds in order to model shapes with multiple disconnected components. For this purpose, we first reformulated the relationship between the data and the shape spaces using Receptive Fields (RFs). Then, we proposed a method for designing RFs in order model the geometric properties of shapes using statistical analysis on data spaces. In order to remove the redundant measurements and degenerate representations in models, we employed an inhibition method.

In order to employ and encode the geometric information obtained from measurements in the shape models, we have suggested a mathematical framework of an indexing mechanism. We have modeled formation of shape models in a vocabulary using a disjoint union topology of shape manifolds induced by the indexing structure of the vocabulary.

The proposed shape representation methods have been used for structural shape classification. In the experiments, we have constructed different shape models by designing various shape manifolds with four different invariance properties, namely invariance to Euclidean Similarity Transformations (EST), Reflection(R), EST-R and Affine Transformation. We have suggested two approaches for hierarchical shape classification using hierarchical compositional models of shapes. The results of the proposed methods on benchmark shape classification outperform state-of-the-art i) shape representation methods, such as Contour Segments & Skeleton Paths, Inner-distance Shape Context, Contour Shapes with Class Segment Sets, ii) shape classification algorithms such as Support Vector Machines (SVM) with Integrated Shape Descriptors, SVM with Procrustes Kernel, SVM with the projection Gaussian kernel, Multiple Kernel Learning with the projection Gaussian kernel, Shape Manifold Classifiers, and iii) hierarchical learning methods such as Denoising, Contractive and marginalized Denoising Auto-enconders, and Deep Belief Networks.

7 APPENDIX: PROOFS OF THE THEOREMS

In this section, we provide proofs of **Corollary 1** and **Theorem 2.2** given in the main text. The definitions which are given in the main text are recalled as they are employed in the proofs.

7.1 Proof of Corollary 1

Definition 7.1 (Part). *A part $\mathcal{P}$ is defined as*

$$\mathcal{P} = \{(A_j, \phi_j) | \phi_j : A_j \to U_j \subset \mathbb{R}^D, \forall j \in \mathfrak{J}\},$$

where $\mathfrak{J}$ is a nonempty countable set, ϕ_j is a homeomorphism from an open set A_j onto an open set U_j such that the composition of the function ϕ_j with its inverse ϕ_j^{-1} is the identity map id_{U_j}, i.e. $\phi_j \circ \phi_j^{-1} = id_{U_j}$, $\forall j \in \mathfrak{J}$.

Definition 7.2 (Shape Spaces). *Let us denote a matrix representation of the measurements as $X \in \mathbb{R}^{N_j \times D}$, where the i^{th} row of X is $\mathbf{x}_i \in A_j$, and $N_j = |A_j|$, $\forall j \in \mathfrak{J}$. A scaled and translated (ST) shape is defined as*

$$[X]_{st} = \{\beta X + I_{N_j} \gamma^T : \beta \in \mathbb{R}^+, \gamma \in \mathbb{R}^D\},$$

where $\beta \in \mathbb{R}^+$ is a scale variable used for isotropic scaling, I_{N_j} is an N_j dimensional unit vector and $\gamma \in \mathbb{R}^D$ is a translation vector. The ST space $\mathcal{S}_{st}$ is the space of all possible pre-shapes [16], [29].

The EST shape space $\mathcal{S}_{est}$ consists of all possible EST Shapes $[X]_{est}$ which are defined as

$$[X]_{est} = \{X_{st} \Gamma : \Gamma \in SO(D), X_{st} \in \mathcal{S}_{st}\},$$

where $SO(D)$ is the special orthogonal group which is the set of all $D \times D$ rotation matrices Γ with $\Gamma^T \Gamma = \Gamma \Gamma^T = I_D$, I_D is a $D \times D$ identity matrix and $\Gamma \in SO(D)$ is a rotation matrix [29]. A shape space $\mathcal{S}_{er}$ consists of all shapes which are defined as

$$[X]_{er} = \{X_{est} R : R \in O(D), X_{est} \in \mathcal{S}_{est}\}, \tag{10}$$

where $O(D)$ is the group of all $D \times D$ orthogonal matrices.

Definition 7.3 (Transition Maps). *Given any two charts* (A_i, ϕ_i) *and* (A_j, ϕ_j) *on* $\mathcal{P}$*, where* $A_i \cap A_j \neq \varnothing$*, the transition maps are defined as [18]*

$$\phi_{ji} : \phi_i(A_i \cap A_j) \to \phi_j(A_i \cap A_j),$$

$$\phi_{ij} : \phi_j(A_i \cap A_j) \to \phi_i(A_i \cap A_j),$$

$$\phi_{ji} = \phi_j \circ \phi_i^{-1} \text{ and } \phi_{ij} = \phi_i \circ \phi_j^{-1}, \forall i, j \in \mathfrak{I}.$$

Definition 7.4 (Measurements obtained from Receptive Fields). *A set of measurements which are obtained from Receptive Fields (RFs) that are employed on two dimensional images is defined as*

$$\mathfrak{RF} = \{\{U_i\}_{i\in\mathfrak{I}}, \{U_{ij}\}_{(i,j)\in\mathcal{L}}, \{\phi_{ji}\}_{(i,j)\in\mathcal{L}}\}, \tag{11}$$

where $\mathcal{L} = \{(i,j) \in \mathfrak{I} \times \mathfrak{I} : U_{ij} \neq \varnothing\}$ *is an index set and* $U_{ij} \triangleq \phi_i(A_i \cap A_j)$ *which satisfy the following conditions:*

1) $\phi_{ii} = id_{U_i}$*, where* id_{U_i} *is an identity map of* U_i*,* $\forall i \in \mathfrak{I}$.
2) $\phi_{ij} = \phi_{ji}^{-1}$*,* $\forall (i,j) \in \mathcal{L}$.
3) *If* $U_{ji} \cap U_{jk} \neq \varnothing$*, then* $\phi_{ij}(U_{ji} \cap U_{jk}) = U_{ij} \cap U_{ik}$*,* $\forall i, j, k \in \mathfrak{I}$*, and* $\phi_{ki}(\mathbf{x}) = \phi_{kj} \circ \phi_{ji}(\mathbf{x})$*,* $\forall \mathbf{x} \in (U_{ij} \cap U_{ik})$.
4) *There are open balls* $V_{\mathbf{x}}$ *and* $V_{\mathbf{x}'}$ *centered at* $\mathbf{x}$ *and* $\mathbf{x}'$ *such that no point of* $V_{\mathbf{x}'} \cap U_{ji}$ *is the image of any point of* $V_{\mathbf{x}} \cap U_{ij}$ *by* ϕ_{ji}*,* $\forall (i,j) \in \mathcal{L}$ *with* $i \neq j$*, and* $\forall \mathbf{x} \in \partial(U_{ij}) \cap U_i$*,* $\forall \mathbf{x}' \in \partial(U_{ji}) \cap U_j$*, where* $\partial(U_{ij})$ *is the boundary of* U_{ij}.

Definition 7.5 (Disjoint Union of Sets [35]). *Disjoint union of sets is defined by*

$$\begin{aligned} \mathbb{U} &\triangleq \coprod_{i\in\mathfrak{I}} U_i \\ &= \bigcup_{i\in\mathfrak{I}} \mathcal{U}_i, \end{aligned} \tag{12}$$

where $\mathcal{U}_i = U_i \times \{i\}$*,* $\forall i \in \mathfrak{I}$ *[14], [47].*

Definition 7.6 (Disjoint Union Topology). *Suppose that* $\{(U_i, \mathcal{T}_i)\}_{i\in\mathfrak{I}}$ *is a family of topological spaces, and* $\mathbb{U}$ *is a disjoint union of sets* U_i *with topology* $\mathcal{T}_i$*,* $\forall i \in \mathfrak{I}$*. Then* *the disjoint union topology on* $\mathbb{U}$ *is defined as*

$$\mathcal{T} = \{V | V \subseteq \mathbb{U}, V \cap U_i \in \mathcal{T}_i, \forall i \in \mathfrak{I}\}.$$

In other words, a subset V *of the disjoint union* $\mathbb{U}$ *is said to be open if and only if its intersection with each set* $V \cap U_i$ *is open in* U_i, $\forall i \in \mathfrak{I}$ *[35]. A disjoint union set* $\mathbb{U}$ *with the topology* $\mathcal{T}$ *is called a disjoint union space.*

Definition 7.7 (Injection in Disjoint Union Topology)**.** *For each* $i \in \mathfrak{I}$*, there is an injection* $\eta_i : U_i \hookrightarrow \mathbb{U}$ *defined as* $\eta_i(u) = (u, i)$*, where* $u \in U_i$ *[35].*

Corollary 2. $\mathcal{P}$ *is a manifold whose co-domain and transition maps are defined by* $\mathfrak{RF}$ *with measurements on two dimensional images.*

Proof:

Let's start by defining the orbit of X as

$$\begin{aligned} \mathfrak{o}(X) &\triangleq [X] \\ &= \{X_{er} : X_{er} \in \mathcal{S}_{er}\}, \end{aligned} \tag{13}$$

where $\mathcal{S}_{er}$ is defined in (10). Then, we obtain $\mathfrak{o} : \mathbb{U} \to \mathcal{P}$. Following the definition of the orbit by quotients, we can define an equivalence relation as

$$X \sim X' \iff X' \in \mathfrak{o}(X).$$

We can employ this equivalence relation on $\mathbb{U}$ stating for all $X, X' \in \mathbb{U}$ that $X \sim X'$ iff $\exists (i,j) \in \mathcal{L}$, such that $X' = \phi_{ij}(X)$ for $X \in U_{ij}, X' \in U_{ji}$. Therefore, we have $\mathcal{P} = \mathbb{U}/\sim$.

Next, we employ the composition

$$t_i \triangleq \mathfrak{o} \circ \eta_i : U_i \to \mathcal{P}, \forall i \in \mathfrak{I}.$$

Since $A_i = t_i(U_i)$ and $\phi_i = t_i^{-1}$, we observe that (A_i, ϕ_i) is a chart of $\mathcal{P}$ for each $i \in \mathfrak{I}$ and their collection $\{(A_i, \phi_i)\}_{i \in \mathfrak{I}}$ form $\mathcal{P}$.

In order to prove that $\mathcal{P}$ is Hausdorff, we choose two orbits $[X] \in \mathcal{P}$ and $[X'] \in \mathcal{P}$ of any two elements $X \in U_i$ and $X' \in U_j$. Then, these measurements can be represented by either disjoint or overlapping models in $\mathcal{P}$. In the first case, $\tau_i(U_i) \cap \tau_j(U_j) = \varnothing$. Since τ_i and τ_j are homeomorphisms, $[X]$ and $[X']$ belong to disjoint sets. For $\tau_i(U_i) \cap \tau_j(U_j) \neq \varnothing$, we should consider four subcases;

1) If $i = j$, then X and X' are separated by disjoint open sets $V_{\mathbf{X}}$ and $V_{\mathbf{X}'}$. In addition, $[X]$ and $[X']$ are separated by disjoint open subsets $\tau_i(V_{\mathbf{X}})$ and $\tau_j(V_{\mathbf{X}'})$ since τ_i is a homeomorphism, $\forall i$.
2) If $i \neq j$, $X \in U_i - cl(U_{ij})$ and $X \in U_j - cl(U_{ji})$, where $cl(\cdot)$ is the set closure, then $[X]$ and $[X']$ are separated by the disjoint open sets $\tau_i(U_i - cl(U_{ij}))$ and $\tau_j(U_j - cl(U_{ji}))$.
3) If $i \neq j$, $X \in U_{ij}$, $X' \in U_{ji}$, $[X] \neq [X']$ and $[X'] \sim \phi_{ij}[X']$, then $[X] \neq \phi_{ij}(X')$.
4) If $i \neq j$, $X \in \partial(U_{ij}) \cap U_i$ and $X' \in \partial(U_{ji}) \cap U_j$, then $[X]$ and $[X']$ are separated by $\tau_i(V_{\mathbf{X}})$ and $\tau_j(V_{\mathbf{X}'})$.

Since there are countably many charts defined by the finite number of RFs, $\mathcal{P}$ is second countable. Additionally, ϕ_{ij} are the maps of $\mathcal{P}$ since $\phi_i = \tau_i^{-1}$ and $\phi_{ji} = \phi_j \circ \phi_i^{-1}$, $\forall i, j \in \mathfrak{I}$ and $\forall (i,j) \in \mathcal{L}$.

□

In **Corollary 1**, we show that a part $\mathcal{P}$ is a manifold. Next, we show that the composition of parts is a manifold under the disjoint union topology.

7.2 Proof of Theorem 2.2

Lemma 7.1 (**Proposition 3.42** in [35]). *Let $S = \{U_i\}_{i \in \mathfrak{I}}$ be an indexed family of topological spaces, and $\mathbb{U} = \coprod_{i \in \mathfrak{I}} U_i$.*

1) *If each $U_i \in S$ is second countable and the index set A is countable, then $\mathbb{U}$ is second countable.*
2) *If each $U_i \in S$ is Hausdorff, then so is $\mathbb{U}$.*
3) *If each $U_i \in S$ is first countable, then so is $\mathbb{U}$.*
4) *Each injection $\eta_i : U_i \hookrightarrow \mathbb{U}$, $\forall i \in \mathfrak{I}$, is a topological embedding, open and closed map.*
5) *A subset of $\mathbb{U}$ is closed iff its intersection with each U_i, $i \in \mathfrak{I}$, is closed.*

Theorem 7.2. *Given a collection of parts $\{\mathcal{P}_n\}_{n\in\mathfrak{N}}$, where $\mathfrak{N}$ is a set of indices, a composition $\mathcal{C}$ of parts defined as*

$$\mathcal{C} = \coprod_{n\in\mathfrak{N}} \mathcal{P}_n \tag{14}$$

is a manifold.

Proof:

Following the propositions 1 and 2 in **Lemma 2.1** given above, $\mathcal{C}$ is *second countable* and *Hausdorff* since each $\mathcal{P}_n$ is a manifold by **Corollary 1**.

Since each $\mathcal{P}_n$ is a manifold, $\mathcal{C}$ is *first countable* by 3 in **Lemma 2.1**. In other words, each point $\pi \in \mathcal{P}_n$ has a countable basis of neighborhoods, therefore $\pi \in V$, where V is an open set.

By the propositions 4 and 5 in **Lemma 2.1**, there exist homeomorphisms $\psi_n : V_n \to W_n$, $\forall n \in \mathfrak{N}$, where W_n is an open set of $\mathcal{C}$. Since each $\mathcal{P}_n$ is locally Euclidean,$\forall n \in \mathfrak{N}$, and following the results, $\mathcal{C}$ is locally Euclidean.

□

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008. 6

[2] U. Aktas, M. Ozay, A. Leonardis, and J. Wyatt, "A graph theoretic approach for object shape representation in compositional hierarchies using a hybrid generative-descriptive model," in *Proceedings of the 2014 European Conference on Computer Vision*, ser. ECCV '14, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, 2014, pp. 566–581. 15, 17, 20

[3] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations in hierarchical architectures," *CoRR*, vol. abs/1311.4158, 2013. 3

[4] X. Bai, W. Liu, and Z. Tu, "Integrating contour and skeleton for shape classification," in *Proceedings of the 2009 IEEE International Conference on Computer Vision*, ser. ICCV '09, Sept 2009, pp. 360–367. 24, 27

[5] X. Bai, C. Rao, and X. Wang, "Shape vocabulary: A robust and efficient shape representation for shape matching," *IEEE Trans. Image Processing*, vol. 23, no. 9, pp. 3935–3949, Sept 2014. 24, 25, 27

[6] A. Bhattacharya and R. Bhattacharya, *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*, 1st ed. New York, NY, USA: Cambridge University Press, 2012. 3, 6, 7, 11, 12, 16

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan 2011. 11

[8] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, Aug 2013. 3

[9] D. Bryner, E. Klassen, H. Le, and A. Srivastava, “2d affine and projective shape analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 998–1011, May 2014. 3, 6

[10] M. Chen, K. Q. Weinberger, F. Sha, and Y. Bengio, “Marginalized denoising auto-encoders for nonlinear representations,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, T. Jebara and E. P. Xing, Eds. JMLR Workshop and Conference Proceedings, 2014, pp. 1476–1484. 20, 21

[11] I. Christou, “Coordination of cluster ensembles via exact methods,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 279–293, Feb 2011. 15

[12] A. Coates and A. Y. Ng, “Selecting receptive fields in deep networks,” in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2528–2536. 7, 10

[13] C. Conrad and R. Mester, “Learning geometrical transforms between multi camera views using canonical correlation analysis,” in *British Machine Vision Conference BMVC*, Surrey, UK, Sep. 2012, pp. 1–12. 3

[14] M. Crossley, *Essential Topology*, ser. Springer Undergraduate Mathematics Series. Springer, 2006. 4, 14, 16, 30

[15] S. de Medrano, *Involutions on Manifolds*, ser. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, 1971. 11

[16] I. Dryden and K. Mardia, *Statistical shape analysis*, ser. Wiley series in probability and statistics. Wiley, 1998. 3, 5, 10, 11, 29

[17] H. G. Feichtinger and T. Strohmer, Eds., *Gabor Analysis and Algorithms: Theory and Applications*, 1st ed. Birkhauser Boston, 1997. 9, 20

[18] J. Gallier, D. Xu, and M. Siqueira, “Parametric pseudo-manifolds,” *Differential Geometry and its Applications*, vol. 30, no. 6, pp. 702–736, 2012. 12, 13, 30

[19] G. H. Golub, V. C. Klema, and G. W. Stewart, “Rank degeneracy and least squares problems,” Stanford, CA, USA, Tech. Rep., 1976. 11

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996. 11

[21] G. Huang and E. Learned-Miller, “Learning class-specific image transformations with higher-order boltzmann machines,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2010, pp. 25–32. 3

[22] J. J. Hunt, P. Dayan, and G. J. Goodhill, “Sparse coding can predict primary visual cortex receptive field changes induced by abnormal visual input,” *PLoS Comput Biol*, vol. 9, no. 5, p. e1003005, 05 2013. 7

[23] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Optimizing over radial kernels on compact manifolds," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14, June 2014, pp. 3802–3809. 6, 7, 26, 27

[24] S. Jayasumana, M. Salzmann, H. Li, and M. Harandi, "A framework for shape analysis via hilbert space embedding," *Proceedings of the 2013 IEEE International Conference on Computer Vision*, pp. 1249–1256, 2013. 6, 7

[25] ——, "A framework for shape analysis via hilbert space embedding," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ser. ICCV '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1249–1256. 27

[26] Y. Jia, C. Huang, , and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '12. Los Alamitos, CA, USA: IEEE Computer Society, 2012, pp. 3370–3377. 7, 10

[27] D. G. Kendall, "Shape manifolds, procrustean metrics, and complex projective spaces," *Bulletin of the London Mathematical Society*, vol. 16, no. 2, pp. 81–121, 1984. 5

[28] ——, "A survey of the statistical theory of shape," *Statist. Sci.*, vol. 4, no. 2, pp. 87–99, 05 1989. 5

[29] D. Kendall, D. Barden, T. Carne, and H. Le, *Shape and Shape Theory*, ser. Wiley Series in Probability and Statistics. Wiley, 2009. 3, 5, 6, 11, 29

[30] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proc. of the 27th International Conference on Machine Learning*, ser. ICML '10, 2010, pp. 543–550. 19, 20

[31] S. Kurtek, E. Klassen, J. Gore, Z. Ding, and A. Srivastava, "Elastic geodesic paths in shape space of parameterized surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1717–1730, Sept 2012. 3, 6

[32] Y. lan Boureau, J. Ponce, and Y. Lecun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th International Conference on Machine Learning*, ser. ICML-10, J. Frnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 111–118. 7, 10

[33] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 473–480. 20, 21

[34] Y. LeCun, "Learning invariant feature hierarchies," in *Proceedings of the 12th International Conference on Computer Vision - Volume Part I*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 496–505. 3

[35] J. M. Lee, *Introduction to topological manifolds*, 2nd ed., ser. Graduate texts in mathematics. New York, Berlin, Heidelberg: Springer, 2011. 30, 31, 32

[36] ——, *Introduction to smooth manifolds*, 2nd ed., ser. Graduate texts in mathematics. New York, Berlin, Heidelberg: Springer, 2013. 3

[37] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," *CoRR*, vol. abs/1411.5908, 2014. 3

[38] J. Lewi, R. Butera, D. M. Schneider, S. Woolley, and L. Paninski, "Designing neurophysiology experiments to optimally constrain receptive field models along parametric submanifolds," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 945–952. 7

[39] H. Li, K. Zhang, and T. Jiang, "Minimum entropy clustering and applications to gene expression analysis," in *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, ser. CSB '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 142–151. 15, 20

[40] Y. Li, J. Zhu, and F. Li, "A hierarchical shape tree for shape classification," in *Proceedings of the 2010 IEEE Conference Image and Vision Computing New Zealand*, ser. IVCNZ '10, Nov 2010, pp. 1–6. 27

[41] Q. Liao, J. Z. Leibo, and T. Poggio, "Learning invariant representations and applications to face verification," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 3057–3065. 3

[42] K.-L. Lim and H. K. Galoogahi, "Shape classification using local and global features," in *Proceedings of the 2010 IEEE Fourth Pacific-Rim Symposium on Image and Video Technology*, ser. PSIVT '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 115–120. 27

[43] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb 2007. 27

[44] E. H. Lockwood, *A book of curves*. Cambridge: Cambridge University Press, 2007. 22

[45] R. Memisevic, "Learning to relate images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1829–1846, Aug 2013. 3

[46] P. Moreno, A. Bernardino, and J. Santos-Victor, "Gabor parameter selection for local feature detection," in *Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science, J. Marques, N. Prez de la Blanca, and P. Pina, Eds. Springer Berlin Heidelberg, 2005, vol. 3522, pp. 11–19. 20

[47] J. Munkres, *Topology*, 2nd ed., ser. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000. 14, 30

[48] N. Murray and F. Perronnin, "Generalized max pooling," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. IEEE Computer Society, June 2014, pp. 2473–2480. 7, 10, 11

[49] X. Peng, B. Sun, K. Ali, and K. Saenko, "Exploring invariances in deep convolutional neural networks using synthetic images," *Under review in ICLR 2015*, vol. abs/1412.7122, 2014. 3

[50] A. Peter and A. Rangarajan, "Information geometry for landmark shape analysis: Unifying shape representation and deformation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 337–350, Feb 2009. 3

[51] G. Peyré, "Manifold models for signals and images," *Comput. Vis. Image Underst.*, vol. 113, no. 2, pp. 249–260, Feb 2009. 7

[52] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8. 3

[53] D. Raviv, R. Kimmel, and A. M. Bruckstein, "Graph isomorphisms and automorphisms via spectral signatures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1985–1993, Aug 2013. 16, 17

[54] C. Samir, A. Srivastava, M. Daoudi, and E. Klassen, "An intrinsic framework for analysis of facial surfaces," *Int. J. Comput. Vision*, vol. 82, no. 1, pp. 80–95, Apr 2009. 3

[55] H. Schulz-Mirbach, "Constructing invariant features by averaging techniques," in *International Conference on Pattern Recognition*, vol. 2, Oct 1994, pp. 387–390. 3

[56] C. Small, *The Statistical Theory of Shape*, ser. Springer Series in Statistics. Springer New York, 1996. 3, 5, 6

[57] S. Soatto, "Visual scene representations: Sufficiency, minimality, invariance and approximations," *CoRR*, vol. abs/1411.7676, 2014. 3

[58] K. Sohn and H. Lee, "Learning invariant representations with local transformations," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, J. Langford and J. Pineau, Eds. New York, NY, USA: ACM, 2012, pp. 1311–1318. 3

[59] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1415–1428, July 2011. 3, 6

[60] K. B. Sun and B. J. Super, "Classification of contour shapes using class segment sets," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 727–733. 27

[61] J. Susskind, R. Memisevic, G. Hinton, and M. Pollefeys, "Modeling the joint density of two images under a variety of transformations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 2793–2800. 3

[62] T. P. Trappenberg, E. T. Rolls, and S. M. Stringer, "Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes," in *Advances in Neural Information Processing Systems*. Vancouver, British Columbia, Canada: MIT Press, Dec 2001, pp. 293–300. 7

[63] L. Tu, *An Introduction to Manifolds*, 1st ed. Springer, 2007. 8

[64] J. R. Ullmann, "An algorithm for subgraph isomorphism," *J. ACM*, vol. 23, no. 1, pp. 31–42, Jan 1976. 16, 17

[65] J. Wang, K. Markert, and M. Everingham, "Learning models for object recognition from natural language descriptions," in *Proceedings of the British Machine Vision Conference*, ser. BMVC '14. BMVA Press, 2009, pp. 1–11. 26

[66] J. Winkeler, B. Manjunath, and S. Chandrasekaran, "Subset selection for active object recognition," in *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, ser. CVPR '99, vol. 2, Ft. Collins, CO, USA, June 1999, pp. 511–516. 11

[67] Q. Xie, I. Jermyn, S. Kurtek, and A. Srivastava, "Numerical inversion of srnfs for efficient elastic shape analysis of star-shaped objects," in *Computer Vision ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8693, pp. 485–499. 3, 6

[68] C. Xu and N. Vasconcelos, "Learning receptive fields for pooling from tensors of feature response,"

in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. IEEE Computer Society, June 2014, pp. 835–842. 7, 9, 10

[69] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group lasso," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2104–2116, 2013. 19, 20

Categorisation of 3D Objects in Range Images Using Compositional Hierarchies of Parts Based on MDL and Entropy Selection Criteria

Vladislav Kramarev, Krzysztof Walas, and Aleš Leonardis

Intelligent Robotics Laboratory, School of Computer Science,
University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
{vvk201,walask,a.leonardis}@birmingham.ac.uk
http://www.cs.bham.ac.uk/research/groupings/robotics/

Abstract. This paper presents a novel approach to object categorisation for range images by using a hierarchical compositional representation. Three dimensional objects are described using surfaces of growing size and complexity. The atomic elements at the bottom layer of the hierarchy are small depth differences which are combined to form the second layer parts. Subsequent layers are formed in the recursive manner, each higher layer is statistically learnt on the layer below via a growing receptive field. Applying compositional hierarchies in this manner allows us to deal with partial occlusions – which are challenging for purely global shape descriptors – and to deal with intra-category variations of shape – which are hard to model using purely local shape descriptors. The compositional hierarchy learns and in a compositional way combines both descriptors in one structure.
Moreover, in this paper we focus on the part selection problem which concerns the choice of the optimisation criterion which provides the information on which parts should be promoted to the higher layer of the hierarchy. Namely two methods based on Minimum Description Length and category based entropy are introduced. The proposed approach was extensively tested on two widely available datasets for object categorisation with the results which are of the same quality as the best results achieved for those datasets.

Keywords: range images, object categorisation, compositional hierarchies, shape parts

1 Introduction

In recent years, processing of range images has experienced a renaissance in the computer vision community. The topic was popular in 1980s and early 1990s [3, 6], but attention waned until the advent of affordable range cameras (e.g. Microsoft Kinect). This has led to two approaches: i) ideas from the 1990s have been re-tested on new hardware, and ii) more recent techniques for intensity images have been adapted to range images. The recent emphasis has been on categorisation and detection tasks.

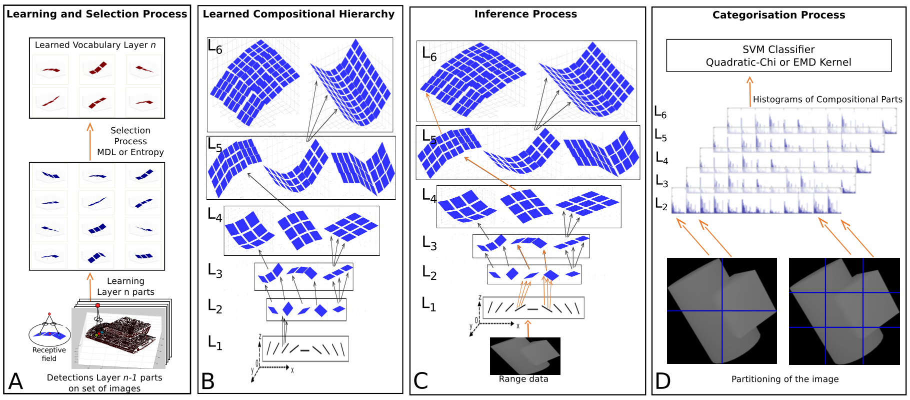


Fig. 1. Overview of the approach. In the learning process (A) a set of depth images is fed into the system. Detections of possible compositions for the next layer, through the whole learning set, are found. The rich set of candidate parts goes through a selection process which provides the constrained vocabulary for the selected layer. Based on the learnt compositional hierarchy (B) an inference process is performed (C). The parts inferred for the specific object are fed into Histograms of Compositional Parts (D). Based on the discretised distribution of parts in partitioned images, categorisation of the object is performed.

This work aims at improved representations of 3D shape for the purposes of categorisation. Based on the successful implementation of the compositional hierarchies [7, 19] and their use in categorisation tasks on intensity images, a proper hierarchical representation is sought for range data. However, it is not a trivial task as intensity and depth data are quite different in their properties. For example the notion of scale variance for intensity images is associated with a projection from the 3D world onto a 2D image plane. Whereas, range images provide metric data and the notion of scale variance is thus associated with support size – the area for which descriptor is computed. The overview of the information flow in the system is shown in Figure 1. The approach presented in this paper is based on computing small depth differences at the bottom layer. These atomic parts are then combined, using statistical learning (Figure 1B), to form compositions at the second layer of the hierarchy. The compositions at each subsequent layer of the hierarchy are learnt from the previous layer. This representation is inherently surface based. In the learning process (Figure 1A) the set of depth images is fed into the system, then the detection of possible compositions are found. The rich set of candidate parts goes through a selection process which provides the constrained vocabulary for the selected layer. Compositions are promoted to the next layer if they fulfil the selected optimisation criterion. This work provides a study of the two selection criteria and their influence on the object categorisation. For the Learnt Compositional Hierarchy inference process, where specific nodes of the hierarchy activate, is performed for an example object (Figure 1C). Active compositions are subsequently fed into to Histogram of Compositional Parts (Figure 1D). Based on the discretised distribution of parts in the partitioned images, categorisation of the object is performed.

The paper begins with the description of the related work in the field of categorisation and detection using range images and compositional representations. In section 2, the proposed hierarchical compositional representation for range images is presented. Section 3 provides a study of the part selection process. The theoretical description is then followed by an evaluation based on two widely available datasets [8, 16]. Finally, concluding remarks followed by future work plans are given.

1.1 State of the Art

Most vision categorisation systems, either for intensity or depth images, are based on feature extraction and have three main sub-systems. First, the keypoints, which are local extrema of a saliency measure, are found. The keypoints are then encoded using a descriptor. Finally, different feature matching algorithms are applied to perform the actual object recognition process. This survey examines all the aforementioned aspects of feature based categorisation systems.

Keypoint Detection Over the years the computer vision community has devised several keypoint detectors for depth data. An extensive survey of such keypoint detection algorithms is presented in [10]. The authors divide these techniques into two main categories: Fixed-Scale and Adaptive-Scale detectors. An extensive evaluation of different keypoint detection algorithms is presented in [27]. The results on the repeatability of keypoints (detecting the same salient points under slight object transformations) were reported mainly for the object detection scenario.

Local Shape Descriptors For detected keypoints to be valuable in the feature matching process they should be equipped with a good descriptor, which maintains a high level of distinctiveness and separability between descriptors so as to make the matches between keypoints unique.

A recent survey of local surface descriptors by Guo et al. [10], divides approaches into: signature, histogram, and transform based methods. The authors list 38 different descriptor types. Apart from the approaches mentioned in the survey it is worth mentioning an application of Deformable Part Models (DPM) in 2D supported by 3D data. Depth data enable pruning of the hypotheses from 2D, based on the metrical relations in 3D [13]. Such a technique could also be used for hierarchical compositional representation in 2D and 3D.

Global Shape Descriptors in contrast to local shape descriptors, are designed to describe a whole object with a single descriptor, see [2] for a survey. The descriptors are of four main categories: histogram, transform, 2D view and graph based. The latest approach reported in [25] uses covariance descriptors. Another method is to use a Hough Transform for 3D Shape Recognition [30]. It is also possible to use aspect graphs for fast 3D object recognition [29]. Another potential measure of shape similarity is proposed in [21] which enables scale-invariant 3D shape recognition.

Compositional Hierarchies Local shape descriptors are suitable for detection tasks because they are robust to clutter and viewpoint changes. On the other hand global shape descriptors are better for shape retrieval tasks where intra-class variability is the main challenge. One concept to link these two worlds is that of compositional hierarchy. It is a bottom up learnt structure, where each layer is learnt using parts from the previous layer. Each higher layer reduces the spatial resolution of the image while at the same time retaining the structure of the object. Lower layers of the hierarchy are similar to local descriptors, whereas higher layers are regarded as global shape descriptors.

The idea of using compositional hierarchies for 3D data is inspired by the already successful application of this structure to 2D intensity images [7, 19]. Similar ideas can also be found in the literature regarding 3D data, starting with the notion of 3D visual phrases in outdoor settings [11], and indoor scenes [5]. Obtaining 3D object templates by hierarchical quantisation [18] is one way of building the compositional hierarchy for 3D data using a volumetric representation. A different approach to learning a hierarchical representation of the data is to use a Deep Neural Network (DNN) with the set of auto-encoders which learn in an unsupervised manner the representation of the provided range data [23]. An approach using depth images and building the compositional representation from surface patches is described in our previous work [14]. In this paper we go beyond these presented approaches to 3D compositional hierarchies and provide a solution to an unsolved problem, namely the mechanisms of establishing part importance which allows to obtain better classification rate.

Part Importance The problem of part selection has not attracted enough attention in the field of compositional hierarchies. In our research we would like to focus on part selection process, which is understood here as an optimisation technique for selecting appropriate parts which would be combined in the subsequent layer. Part selection process is strongly dependent on the actual task, therefore various part importance criteria may be applied. In shape retrieval process the parts selected ought to be more generic. Whereas, when dealing with pose estimation the parts are required to be discriminative to estimate the pose unambiguously. The problem of part importance for 2D recognition and reconstruction task is considered in [9], as well as the more general problem of finding interesting, salient points [1] or ranking 3D features [28]. Additionally, selected parts – keypoints in general should be repeatable. The problem of how to efficiently learn to detect repeatable parts in depth data was addressed in [12]. In context of 3D compositional hierarchies our previous work [14] used a frequency-based part selection strategy, i.e. the chances of a candidate part to be included in the vocabulary strongly depends on frequency of a part's observation in training data.

Histogram Similarity Measures. In our approach object categories are represented as stacked histograms of compositional parts (e.g. shown in Figure 1D).

Histograms have widely been used for image classification, as shown in study [32]. χ^2 and histogram intersection are very popular histograms similarity measures. However, these measures don't take into account cross-bin similarities, which are very important for compositional hierarchical approaches, since different compositional parts may represent similar surface types.

Earth Mover's Distance (EMD) is a minimal cost that must be paid to transform one histogram into the other. Zamolotskikh et al. [31] show three ways how EMD kernels may be used with SVM. The Quadratic-Chi histogram distance measure [20] not only takes into account cross-bin relations, but also performs normalisation such that difference between small bins becomes as important as difference of large ones.

1.2 Contributions

The main contribution lies in a successful conception, design and implementation of a learnt compositional hierarchy of surface parts, which are used for 3D object categorisation. The representation optimises two different two different part selection criteria which results in state-of-the-art categorisation performance on two publicly available datasets. We also tested three different affinity measures to produce kernel function for SVM classification.

2 Hierarchical Compositional Representation of Surfaces

This section outlines the compositional hierarchical representation for depth data using surface patches. The description is divided into two parts. First, the learning process is described, subsequently the inference procedure is specified. The coordinate system for the reminder of this paper is established as follows: x and y axes span the image plane and the z-axis encodes depth information.

The general overview of the proposed compositional hierarchy is shown in Figure 2d. The representation is defined as a hierarchy of layers L_n where n denotes the layer number. First layer L_1 consists of pre-defined parts. In our case these are quantised differences of depth between pixels at the fixed distance from each other, computed in the x-axis direction. The number of parts at the first layer is selected depending on the required precision of the representation. An example of the vocabulary of the first layer parts is depicted in Figure 2a. In this case it comprises nine different orientations.

In general, the vocabulary of each higher layer of the hierarchy L_n, $\forall n > 1$ is learnt using statistical learning of co-occurrence of parts from the layer below. The composition P_i^n comprise the central part $P_{central}^{n-1}$ and other sub-parts P_j^{n-1} located in the local neighbourhood of $P_{central}^{n-1}$, called the *receptive field*:

$$P_i^n \equiv (P_{central}^{n-1}, \{P_j^{n-1}, \mu_j, \Sigma j\}), \quad (1)$$

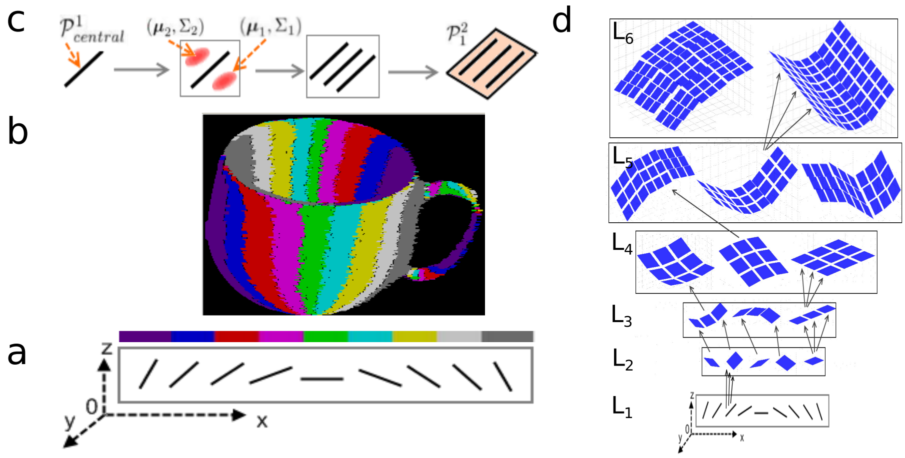


Fig. 2. Process of forming L_2 parts from L_1 sub-parts. Quantisation of depth differences into nine bins (a). Realisation of L_1 parts on the example object – colour coded depth differences (b). Three adjacent depth differences form a small oriented surface patch (c).

where $\mu_j = (x_j, y_j, z_j)$ is the mean relative position of the sub-part P_j^{n-1} w.r.t. the central part, and Σj is the covariance matrix representing variability of possible relative positions.

The process of constructing second layer parts is depicted in Figure 2. Parts at L_3 are assembled from triplets of parts from L_2 which are adjacent in x-direction. Whereas, parts at L_4 are formed from L_3 parts aligned in y-direction.

2.1 Learning Vocabulary of Parts

The aforementioned hierarchical representation is learnt using the procedure the main steps of which are outlined below:

1. Represent the training data in terms of parts of layer L_{n-1}.
2. In the local neighbourhood of each part perform *local inhibition*.
3. Compute statistical maps, for training data, characterising the 3D spatial relation between parts of the layer L_{n-1}.
4. Based on the statistical maps build compositions and list them as *candidate parts*.
5. Perform an optimisation process to form the vocabulary for layer L_n. This is a subset of parts that satisfies some optimality criterion called in this paper *selection criteria*, explained in details in section 3.

2.2 Inference

The learnt compositional hierarchy of parts is used for general computer vision tasks such as categorisation or detection of objects from range images. The

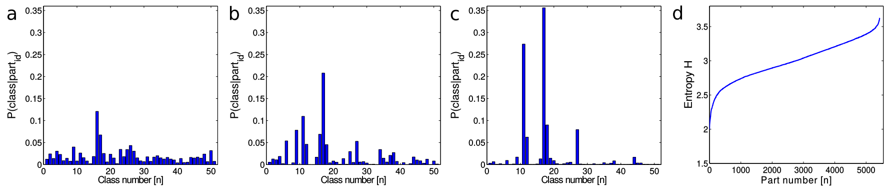


Fig. 3. Entropy based on the probability of the class given individual part $P(class|part_{id})$. Histogram for part with: high entropy value 3.6 (a), mean entropy value 3.0 (b), low entropy value 2.0 (c). Sorted entropy values for the candidate parts at layer three (d).

inference process which generates features from a range image using a given vocabulary is performed as follows:

1. Convolve the range image with an oriented Gaussian-derivative filter (aligned along the x-axis).
2. Quantise the filter responses at each pixel and assign them to the bin that corresponds to the closest first layer part type.
3. Build a subset of the set of potential parts $S_{sel} \subset S_{pot}$ selecting the parts which optimise selected optimisation criterion.
4. Perform the inference at subsequent layers L_n, $\forall n > 1$ as follows.
5. Each part P_i^{n-1} from Layer L_{n-1} is treated as a *central part*.
6. In the local neighbourhood of the $P_{central}^{n-1}$ extract parts present in the actual data together with their relative position w.r.t. $P_{central}^{n-1}$.
7. Spatial configurations of $P_{central}^{n-1}$ and neighbouring parts (extracted from the depth data) form a certain spatial configuration which is matched against parts from the previously learnt vocabulary.
8. The parts obtained in the inference process are strongly overlapped while the detections are performed at each pixel. The optimisation procedure selected in step 3 is performed again.
9. The process runs recursively from point 4 for each subsequent layer. At the end of the procedure the whole depth image is described in terms of parts from the learnt vocabulary for each layer.

3 Part Selection Criteria

In this paper two part selection strategies for learning process are studied. They are based on parts coverage and histogram entropy criteria.

Minimum Description Length MDL based part selection strategy tries to find the minimal subset of candidate parts which is required to represent most of the input range data in terms of the shape vocabulary. We implemented it as a quadratic binary optimisation problem and solved it in greedy fashion very similar to the solution proposed in [17].

Entropy Entropy based part selection is promoting to the higher layer of the hierarchy the compositions with low entropy. The class probability given individual part $P(class|part_{id})$ is approximated using histograms computed for all the parts through the whole learning dataset. The entropy of each individual part was computed using:

$$H = -\sum_{i=1}^{n} p_i \times log(p_i). \tag{2}$$

The $P(class|part_{id})$ for part with high, mean and entropy is shown in Figure 3 a,b,c respectively. The graph presenting the sorted entropy H computed for all the the compositions detected at layer L_3 is shown in Figure 3 d. Minimum and maximum entropy values for layer 3 are equal 2.0 and 3.6 respectively. However, for the layer 4 these values are equal to 1.7 and 3.8 respectively.

4 Experiments

Datasets In order to test performance of our hierarchical compositional representation we perform extensive tests on two publicly available datasets. We focus our attention on the task of object categorization from range images. The first dataset (RGB-D Washington Dataset [16]) contains real world data registered with a Kinect-like device, namely 300 objects that are split into 51 categories. For our experiments we used only 20% of the images of each object instance (depth channel only). To split data into train/test sets we used the exact list of object instances left out provided on a web site of the dataset.

Another dataset we used is SHREC07 [8] containing 400 meshes representing objects that are split into 20 categories. We rotate these mesh models about z-axes with step 18 degrees, thus producing $400 \times 20 = 8000$ depth images. Then we evaluate performance using one leave out cross validation.

Learning of Hierarchical Shape Vocabulary We learn layers 2-6 of compositional hierarchical shape vocabulary on each dataset separately. In our first experiment we learn a compositional hierarchy using the MDL principle. In the second experiment we additionally include the most category-specific parts to the vocabulary of each layer (parts with low entropy). All the experiments are done in a way to avoid data snooping i.e. we extract category-specific parts from the training data only.

Classification procedure starts with the inference of parts of layers 1-6 from range images. Secondly, following [14] we build stacked histograms of parts as shown in Figure 4, i.e we build partition each object image into 14 sub-regions (1 for the whole object, plus 2×2 and 3×3 partitioning). The distribution of parts of layers 2-6 within each sub-region is represented as a histogram; histograms of each sub-region are then concatenated to form a HOCP (Histogram of Compositional Parts) surface descriptor. These descriptors are fed into SVM classifier.

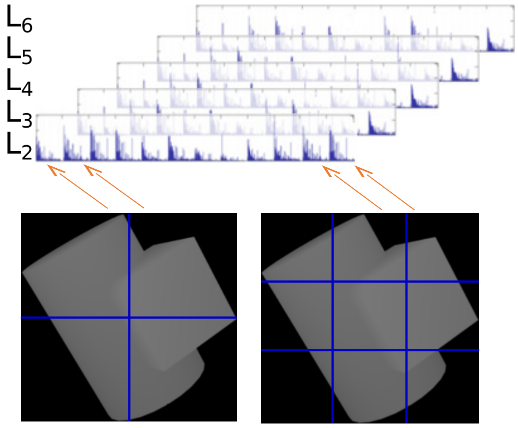


Fig. 4. Histogram of Compositional Parts – detections for each range image sub-region is computed. The histograms from all sub-regions and all layers are stacked together to form the category signature.

Experiments with kernels for SVM Three different types of kernels for SVM were tested in our experiments. χ^2 kernels are implemented in standard packages for SVM. The EMD measure does not produce a positive semi-definite kernel matrix, therefore some transformations are required to modify this metric [31]. We apply the following transformation: $K(S,Q) = exp(-EMD(S,Q))$, where S and Q are histograms. Another task is to define a ground distance metric D_{ij} which denote distance between bins i and j of the histogram. We define this matrix based on geometric similarity of parts, i.e. bins representing parts with similar geometric properties have smaller ground distance. Please note, that we consider part similarities within each sub-region, and do not take into account cross sub-region relations. For Quadratic-Chi measure, it is required to define a bin-similarity matrix A_{ij}. We do it in the same way as in case of EMD kernels, i.e. only similarities of parts within each sub-region are taken into consideration.

5 Results and Discussion

The results achieved for the for the Washington dataset are shown in Figure 5. Comparison of using different part selection criteria with the same kernel for classification is shown in Figure 5a. Our experiments show that combined MDL and entropy based selection perform better than MDL alone, especially for higher layers where parts become more specific to categories. Additionally the comparison of performance of different kernels used in the classification procedure is shown in Figure 5b. Our observation is that the best results are achieved with the Quadratic-Chi kernel, which takes into account cross-bin relations and performs normalisation as well, such that difference between small bins becomes as important as difference of large ones. This turns out to be a substantial property: all parts in the histograms are distributed non-uniformly and some non-frequent but discriminative parts may have larger impact for category detection. This

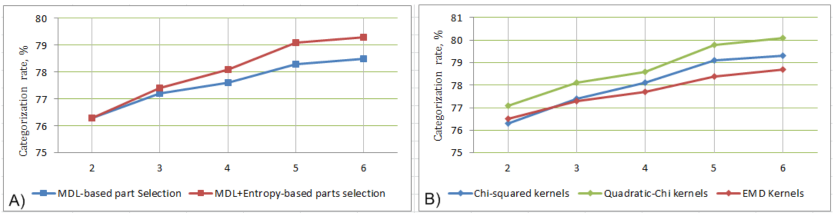


Fig. 5. Categorisation accuracy achieved on Washington RGBD dataset. A) Comparison of part selection strategies (SVM with χ^2 kernels is used for both experiments). B) Comparison of different types of kernels for SVM

observation might also explain the poorer performance of the EMD kernel as it was pointed out in [20]. Comparison of our results to other approaches is given in the Table 1. As it could be observed our approach is almost equal to the best results achieved for this dataset.

Additionally we have also provided the results for the artificial data which are given in the dataset from Shape Retrieval Contest SHREC07 [8]. As it could be observed in the Table 1, our approach is also almost reaching the top score.

6 Conclusion

The research presented in this paper is focused on part selection problem and its influence on the performance of the hierarchical compositional representation in the object categorisation task. Presented results show that the appropriate selection criteria can boost the results. Comparing to our previous approach the gain for RGB-D dataset is about 4%. Additional tests with histogram similarity measure show that additional one percent could be added to the final result. However, the use of the compositional hierarchies, being generative models, not only could provide boost in the performance, but also have additional properties like better scalability and robustness to partial occlusions. Moreover, it enables performing detection task without using sliding windows. A thorough investigation of all those aspects of the compositional hierarchies is envisioned as a future work.

Washington dateset		SHREC'07 dataset	
Method	Depth data only	Method	Depth data
Random Forest [16]	66.8±2.5	ISM 1-NN [24]	79.0
3D SPMK [22]	69.4	BoW [26]	87.3
HCR [14]	75.6	HCR [14]	95.6
HMP3D [15]	76.5±2.2	ISM 2-NN [24]	100.0
CNN-RNN [23]	78.9±3.8		
SP+HMP [4]	81.2±2.3		
Ours	80,1±2.2	Ours	96.4

Table 1. Results for Washington RGB-D dataset [16] and on SHREC'07 dataset [8].

References

1. Aanæs, H., Dahl, A., Steenstrup Pedersen, K.: Interesting Interest Points. Int. J. of Computer Vision 97(1), 18–35 (2012)
2. Akgul, C., Sankur, B., Yemez, Y., Schmitt, F.: 3D Model Retrieval Using Probability Density-Based Shape Descriptors. Pattern Analysis and Machine Intell., IEEE Trans. on 31(6), 1117–1133 (2009)
3. Besl, P.J.: Surfaces in Range Image Understanding. Springer Series in Perception Engineering, Springer-Verlag (1988)
4. Bo, L., Ren, X., Fox, D.: Unsupervised Feature Learning for RGB-D Based Object Recognition. In: Desai, J.P., Dudek, G., Khatib, O., Kumar, V. (eds.) Experimental Robotics, Springer Tracts in Advanced Robotics, vol. 88, pp. 387–402. Springer Int. Publishing (2013)
5. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding Indoor Scenes Using 3D Geometric Phrases. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conf. on. pp. 33–40 (2013)
6. Faugeras, O.: Three-Dimensional Computer Vision. A geometric view point. MIT Press (1993)
7. Fidler, S., Leonardis, A.: Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conf. on. pp. 1–8 (2007)
8. Giorgi, D., Biasotti, S., Paraboschi, L.: SHape REtrieval Contest 2007: Watertight Models Track
9. Guo, G., Wang, Y., Jiang, T., Yuille, A., Fang, F., Gao, W.: A Shape Reconstructability Measure of Object Part Importance with Applications to Object Detection and Localization. Int. J. of Computer Vision 108(3), 241–258 (2014)
10. Guo, Y., Bennamoun, M., Sohel, F., Lu, M., Wan, J.: 3D Object Recognition in Cluttered Scenes with Local Surface Features: A Survey. Pattern Analysis and Machine Intell., IEEE Trans. on 36(11), 2270–2287 (2014)
11. Hao, Q., Cai, R., Li, Z., Zhang, L., Pang, Y., Wu, F.: 3D visual phrases for landmark recognition. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conf. on. pp. 3594–3601 (2012)
12. Holzer, S., Shotton, J., Kohli, P.: Learning to Efficiently Detect Repeatable Interest Points in Depth Data. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012, Lecture Notes in Computer Science, vol. 7572, pp. 200–213. Springer Berlin Heidelberg (2012)
13. Kim, B.s., Xu, S., Savarese, S.: Accurate Localization of 3D Objects from RGB-D Data Using Segmentation Hypotheses. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conf. on. pp. 3182–3189 (2013)
14. Kramarev, V., Zurek, S., Wyatt, J.L., Leonardis, A.: Object Categorization from Range Images using a Hierarchical Compositional Representation. In: Pattern Recognition (ICPR), 2014 22nd Int. Conf. on (2014)
15. Lai, K., Bo, L., Fox, D.: Unsupervised feature learning for 3D scene labeling. In: Robotics and Automation (ICRA), 2014 IEEE Int. Conf. on. pp. 3050–3057 (2014)
16. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: Robotics and Automation (ICRA), 2011 IEEE Int. Conf. on. pp. 1817–1824 (2011)
17. Leonardis, A., Bischof, H., Maver, J.: Multiple eigenspaces. Pattern Recognition 35(11), 2613–2627 (2002)

18. Li, B., Wu, T., Zhu, S.C.: Integrating Context and Occlusion for Car Detection by Hierarchical And-Or Model. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol. 8694, pp. 652–667. Springer Int. Publishing (2014)
19. Ommer, B., Sauter, M., Buhmann, J.M.: Learning Top-Down Grouping of Compositional Hierarchies for Recognition. In: Proceedings of the 2006 Conf. on Computer Vision and Pattern Recognition Workshop. pp. 194–211. CVPRW '06, IEEE Computer Society, Washington, DC, USA (2006)
20. Pele, O., Werman, M.: The Quadratic-Chi Histogram Distance Family. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol. 6312, pp. 749–762. Springer Berlin Heidelberg (2010)
21. Pham, M.T., Woodford, O., Perbet, F., Maki, A., Stenger, B., Cipolla, R.: A new distance for scale-invariant 3D shape recognition and registration. In: Computer Vision (ICCV), 2011 IEEE Int. Conf. on. pp. 145–152 (2011)
22. Redondo-Cabrera, C., López-Sastre, R.J., Acevedo-Rodríguez, J., Maldonado-Bascón, S.: Recognizing in the depth: Selective 3D Spatial Pyramid Matching Kernel for object and scene categorization. Image and Vision Computing 32(12), 965–978 (2014)
23. Richard Socher and Brody Huval and Bharath Bhat and Christopher D. Manning and Andrew Y. Ng: Convolutional-Recursive Deep Learning for 3D Object Classification. In: Advances in Neural Information Processing Systems 25 (2012)
24. Salti, S., Tombari, F., Di Stefano, L.: On the Use of Implicit Shape Models for Recognition of Object Categories in 3D Data. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Computer Vision – ACCV 2010, Lecture Notes in Computer Science, vol. 6494, pp. 653–666. Springer Berlin Heidelberg (2011)
25. Tabia, H., Laga, H., Picard, D., Gosselin, P.H.: Covariance Descriptors for 3D Shape Matching and Retrieval. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on. pp. 4185–4192 (2014)
26. Toldo, R., Castellani, U., Fusiello, A.: A bag of words approach for 3d object categorization. In: Computer Vision/Computer Graphics Collaboration Techniques, pp. 116–127. Springer (2009)
27. Tombari, F., Salti, S., Di Stefano, L.: Performance Evaluation of 3D Keypoint Detectors. Int. J. of Computer Vision 102(1-3), 198–220 (2013)
28. Tuzel, O., Liu, M.Y., Taguchi, Y., Raghunathan, A.: Learning to Rank 3D Features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol. 8689, pp. 520–535. Springer Int. Publishing (2014)
29. Ulrich, M., Wiedemann, C., Steger, C.: Combining Scale-Space and Similarity-Based Aspect Graphs for Fast 3D Object Recognition. Pattern Analysis and Machine Intell., IEEE Trans. on 34(10), 1902–1914 (2012)
30. Woodford, O., Pham, M.T., Maki, A., Perbet, F., Stenger, B.: Demisting the Hough Transform for 3D Shape Recognition and Registration. Int. J. of Computer Vision 106(3), 332–341 (2014)
31. Zamolotskikh, A., Cunningham, P.: An Assessment of Alternative Strategies for Constructing EMD-Based Kernel Functions for Use in an SVM for Image Classification. In: Content-Based Multimedia Indexing, 2007. CBMI '07. Int. Workshop on. pp. 11–17 (2007)
32. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. Int. J. of Computer Vision 73(2), 213 238 (2007)

SHREC2015 - Retrieval of Objects from Highly Cluttered Scene Captured with Depth-Sensing Camera – Evaluation Procedure

Krzysztof Walas

February 11, 2015

1 Measures for Multi-label Classification

The proposed set of measures is described in [4]. The following notation is used:

L – a set of disjoint labels;
$|L|$ – cardinality of the set of labels;
$Y \subseteq L$ – set of labels for current example (scene);
D – multi-label dataset consisting of $|D|$ multi-label examples (x_i, Y_i); $i = 1 \ldots |D|$

Label Cardinality:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i| \tag{1}$$

Label Density:

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \tag{2}$$

Hamming Loss:

$$HammingLoss(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \tag{3}$$

where:
Δ – symmetric difference of two sets;
$Z_i = H(x_i)$ – a set of labels predicted by H for example x_i;
H – multi-label classifier.

Common statistical measures:

$$Accuracy(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Y_i \bigcup Z_i|} \tag{4}$$

$$Precission(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Z_i|} \tag{5}$$

$$Recall(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Y_i|} \tag{6}$$

$$F_1(H, D) = 2 \times \frac{Precission \times Recall}{Precission + Recall} \tag{7}$$

The results for each scene have different weights which depends on the label density. The final result of the recognition task would be the overall F_1 score for the whole testing set.

2 Measures for Objects Detection

In the detection task the Homogeneous Transformation matrices are used:

$$T = \begin{bmatrix} s_x r_{11} & s_y r_{12} & s_z r_{13} & t_x \\ s_x r_{21} & s_y r_{22} & s_z r_{23} & t_y \\ s_x r_{31} & s_y r_{32} & s_z r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

$$T = Trans \cdot Rot \cdot Scale \tag{9}$$

Translation matrix

$$Trans = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

Rotation matrix

$$Rot = \begin{bmatrix} r_{11} & r_{12} & r_{13} & 0 \\ r_{21} & r_{22} & r_{23} & 0 \\ r_{31} & r_{32} & r_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11}$$

Scale matrix

$$Scale = \begin{bmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{12}$$

$$s_x = \sqrt{(s_x r_{11})^2 + (s_x r_{21})^2 + (s_x r_{31})^2};\ s_y = \sqrt{(s_y r_{12})^2 + (s_y r_{22})^2 + (s_y r_{32})^2};\ s_z = \sqrt{(s_z r_{13})^2 + (s_z r_{23})^2 + (s_z r_{33})^2}$$

Bi-invariant Distance metric for SO(3) is proposed as a distance measure for rotations[3]

$$\Phi(R_1, R_2) = \|log(R_1 R)2^T\| \tag{13}$$

where:
$log(R) = \frac{\phi}{2sin\phi}(R - R^T)$; ϕ satisfies $1 + 2cos\phi = Tr(R)$ and $\|logR\|^2 = \phi^2$.

However, from the analysis shown in[1], given two Quaternions, (13) is linearly related to:

$$\Phi(Q_1, Q_2) = arccos(\|Q_1 \cdot Q_2\|), \tag{14}$$

where:
$Q_1 \cdot Q_2 = w_1 w_2 + x_1 x_2 + y_1 y_2 + z_1 z_2$
This solution has a smaller computational cost.
Obtaining quaternions from rotation matrix:

$$Q = \begin{bmatrix} \frac{1}{2}\sqrt{1 + r_{11} + r_{22} + r_{33}} = w \\ \frac{r_{32} - r_{23}}{4w} = x \\ \frac{r_{13} - r_{31}}{4w} = y \\ \frac{r_{21} - r_{12}}{4w} = z \end{bmatrix} \tag{15}$$

The use of this representation leads however to an ambiguity depending on the sign of the quaternion. If the rotation axis is represented by the vector there are two vectors which are collinear with this axis. To alleviate this problem the solution from [2] was used. It could be tested whether $Q_1 \cdot Q_2$ is negative and negate one of the quaternions to obtain its equivalent alternate: $-Q_1 \cdot Q_2 = Q_1 \cdot -Q_2$

3 Scores

3.1 Recognition Task

The participants will provide a text file *scene_number.txt* which contains numbers of classes found in the scene, separated by semicolon. The number of objects for current scene is not given.
Example file: 001.txt $->5;4;6;10;12;15\ldots$

For the set of scenes overall score is computed using F_1 with the precision and recall weighted by the label density for each scene:

$$Precission_w(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Z_i|}\cdot\frac{|Y_i|}{|L|} \tag{16}$$

$$Recall_w(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \bigcap Z_i|}{|Y_i|}\cdot\frac{|Y_i|}{|L|} \tag{17}$$

$$F_1(H,D) = 2\cdot\frac{Precission_w \times Recall_w}{Precission_w + Recall_w} \tag{18}$$

The solution with the highest F_1 score is winning this task.

3.2 Detection Task

For each scene participants will provide a text file *scene_number.txt* which contains class number followed by the Transformation matrix; Matrix entries separated by *semicolon only no additional spaces*; The number of objects for current scene is not given.
Example file: 001.txt $->$:
5
sr_{11};sr_{12};sr_{13};t_x
sr_{21};sr_{22};sr_{23};t_y
sr_{31};sr_{32};sr_{33};t_z
0; 0; 0; 1
4
sr_{11};sr_{12};sr_{13};t_x
sr_{21};sr_{22};sr_{23};t_y
sr_{31};sr_{32};sr_{33};t_z
0; 0; 0; 1
⋮

Scaling is uniform for all the axes.

For the obtained file at first the position of the object is checked. Class and translation are taken into account

$$trans_{dist} = \sqrt{(t_{x1}-t_{x2})^2 + (t_{y1}-t_{y2})^2 + (t_{z1}-t_{z2})^2}. \tag{19}$$

If the distance from the ground truth position is higher then 0.2 the prediction is discarded and gets score 0.2 for translation, rotation is not checked and the prediction gets maximum error π, scale is also not checked and get the maximum error 1. The same applies when the class is missing or superfluous in the predicted set. Change of scale has deep implication due to the curse of dimensionality. If we would like to compare the volume of the cube with edge length $a = 4$. If the object is two times smaller ($s_1 = 0.5$). The difference in volume related to original volume is $(V_o - V_{s_1})/V_o$ i.e. $(2^3 - 1^3)/2^3 = 0.875$ whereas if the object is two times larger ($s_2 = 2$) the difference in volume related to original volume is $(V_o - V_{s_2})/V_o$ i.e. $(2^3 - 4^3)/2^3 = -7$. Hence the error in volume is smaller for downscaling than for up-scaling. For $s_1 = 0.5$ the same error is obtained for $s_2 = 1.2331$. Scale distance is given as:

$$scale_{dist} = \left|\frac{s_1^3 - s_2^3}{s_1^3}\right| \tag{20}$$

Rotation distance is given by 14.

Detection error is computed over all the scenes for the sum of all error factors weighted by the scene density:

$$DetectionError(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{3} \cdot \left(\frac{trans_{dist}}{0.2} + \frac{rot_{dist}}{\pi} + scale_{dist} \right) \cdot \left(1 - \frac{|Y_i|}{|L|} \right) \tag{21}$$

The solution with the lowest $DetectionError(H, D)$ score is winning this task.

References

[1] DuQ. Huynh. Metrics for 3D Rotations: Comparison and Analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009.

[2] J.J. Kuffner. Effective sampling and distance metrics for 3D rigid body path planning. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, volume 4, pages 3993–3998 Vol.4, April 2004.

[3] F. C. Park and Bahram Ravani. Smooth Invariant Interpolation of Rotations. *ACM Trans. Graph.*, 16(3):277–295, July 1997.

[4] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13, 2007.

Analysis of 3D Surface Properties for Modeling 3D Compositional Representations

Technical Report

University of Birmingham

Krzysztof Walas, Mete Ozay, Jeremy L. Wyatt

February 28, 2015

1 Introduction

In this report, we have analyzed geometric properties of range data measured on 3D objects which have been used for modeling compositional representation of 3D shapes of the objects. Understanding of these properties plays a crucial role for modeling compositional representation of 3D shapes of objects. Development of a learned 3D visual vocabulary of a shape of an object is a step towards a new approach for object detection in the scene. In recent years, range data has been popularly used for image understanding by the appearance of the affordable Kinect-like range sensors on the market. There are two approaches for the employment of range images; i) examination of the ideas proposed in early works [1, 2, 3] using new hardware, and ii) adaptation of techniques developed for intensity images for modeling shapes using range images, such as SIFT [4] and SURF [5] descriptors employed for feature extraction from range data.

It is highly desirable to understand geometric properties of a surface of 3D shape of an object, and use appropriate mappings from a space of features extracted from the surface to a Euclidean space of shape models in order to design distance functions of parts of a compositional representation of shapes. The problem of designing mappings from a space for 3D data to a space of 2D data has been studied in order to align a texture pattern on an object. However, the popular techniques used in computer graphics are computationally demanding and their focus is on an exact reconstruction of the shapes [6, 7, 8, 9]. Nonetheless, some ideas regarding surface parametrization might provide us a deeper understanding of flat embedding of 2D and 3D shape features into a space of representations.

On the other hand, computer vision algorithms which have been developed for scene understanding using 2D images and range data, have been optimized for computational efficiency [10, 11, 12]. The aforementioned works have been using local surface patches for the construction of the models. However, they are not taking into account the geometric property of the whole object. Hierarchical compositional representations of 3D models have been proposed in order to bridge the gap between these two approaches [13, 14, 15]. On of the goals of these works have been statistical learning of representations which encompass the description of the object at different levels of abstraction from a level of fine grained details to a level of a coarse representation of the whole shape.

In this report, we provide a detailed analysis of the intrinsic geometry of surfaces of 3D shapes in order to design distance functions of 3D compositional models on the surfaces using surface parametrization techniques. The results will be used for development of hierarchical compositional architectures of 3D models. For this purpose, first we given an overview of the mathematical concepts for the required background. Next, the problem of designing distance functions on a surface is posed. Finally, we analyzed geometric properties of scale space.

2 Background

A parametric form of a general surface $S \subset \mathbb{R}^3$ is defined with respect to a known coordinate system as [1]

$$S = \left\{ \vec{x} \in \mathbb{R}^3 : \begin{bmatrix} x \\ y \\ z \end{bmatrix} \triangleq \begin{bmatrix} d(u,v) \\ e(u,v) \\ f(u,v) \end{bmatrix}, (u,v) \in D \subseteq \mathbb{R}^2 \right\}, \tag{1}$$

where $\vec{x}(u,v)$ is a parametric representation of the surface, and $d(u,v)$, $e(u,v)$ and $f(u,v)$ is the x, y and z component of $\vec{x}(u,v)$, respectively.

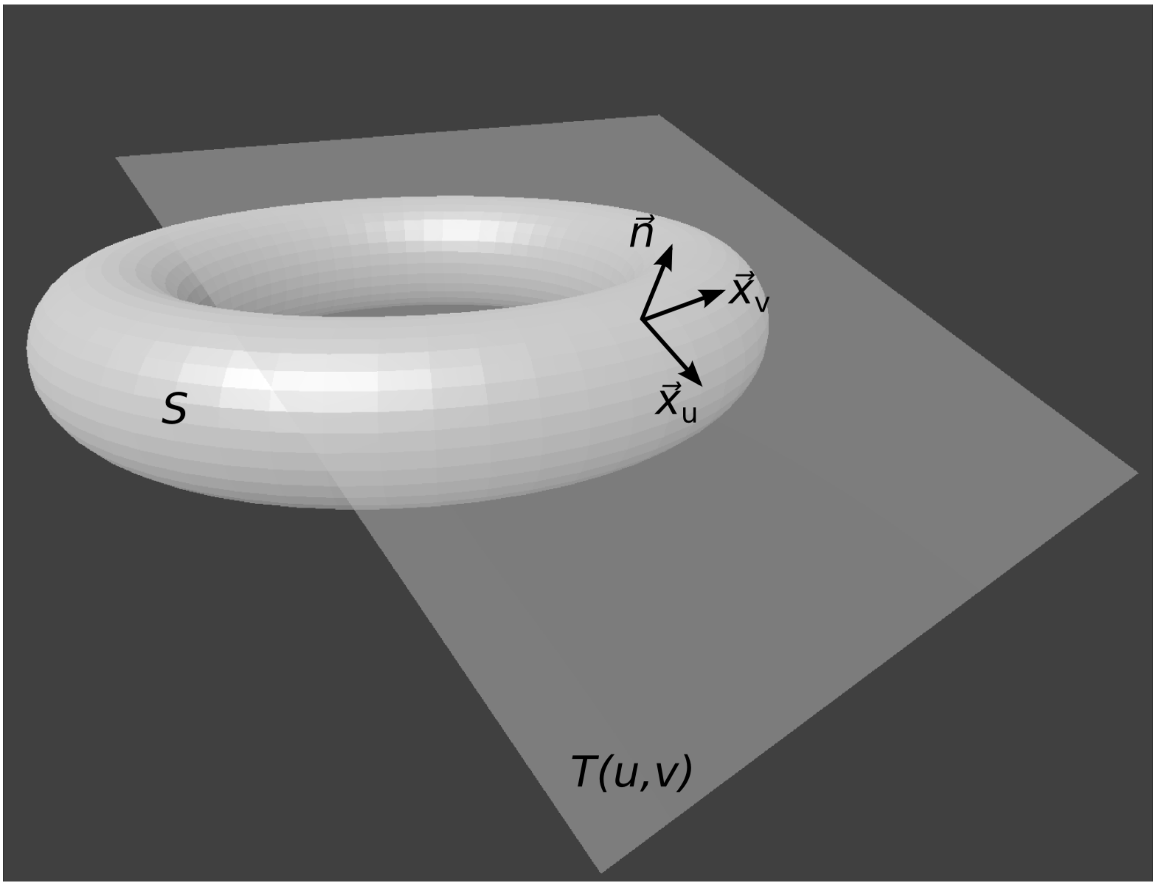


Figure 1: Parametrized surface S together with a tangent plane $T(u, v)$ at a point (u, v).

Then, we define a concept of the *first fundamental form* I of a surface representation $\vec{x}(u, v)$ as

$$\begin{aligned} I(u, v, du, dv) &\triangleq d\vec{x} \cdot d\vec{x} \qquad (2)\\ &= Edu^2 + 2Fdudv + Gdv^2 \\ &= \begin{bmatrix} du & dv \end{bmatrix} \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} du \\ dv \end{bmatrix} \\ &= d\vec{u}^T [g] d\vec{u}, \qquad (3) \end{aligned}$$

where the elements of the matrix are defined as

$$g_{11} \triangleq E = \vec{x}_u(u, v) \cdot \vec{x}_u(u, v)$$

$$g_{22} \triangleq G = \vec{x}_v(u, v) \cdot \vec{x}_v(u, v)$$

$$g_{12} = g_{21} \triangleq F = \vec{x}_u(u, v) \cdot \vec{x}_v(u, v)$$

and the subscripts denotes the partial derivatives $\vec{x}_u(u, v) = \frac{\partial \vec{x}}{\partial u}$ and $\vec{x}_v(u, v) = \frac{\partial \vec{x}}{\partial v}$ which are called u and v tangent vector, respectively. The tangent vectors are depicted in Figure 1. Note that they lie in and form a basis for a tangent plane $T(u, v)$ of the surface at the point $\vec{x}(u, v)$.

In addition, $I(u, v, du, dv)$ is invariant to parametrization, translation and rotation of the surface. The *first fundamental* form does not depend on how the surface is embedded in a 3D space. The dependence on the surface itself, not its ambient space, provides us a set of intrinsic properties of the surface.

The *second fundamental form* of a surface is dependent on the embedding of the surface in a 3D space. It is an extrinsic property of the surface and defined by

$$
\begin{aligned}
II(u,v,du,dv) &\triangleq -d\vec{x}\cdot d\vec{n} && (4)\\
&= Ldu^2 + 2Mdudv + Ndv^2\\
&= \begin{bmatrix} du & dv \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} du \\ dv \end{bmatrix}\\
&= d\vec{u}^T[b]d\vec{u}, && (5)
\end{aligned}
$$

where the elements of the matrix are defined as

$$b_{11} \triangleq L = \vec{x}_{uu}(u,v)\cdot\vec{n}(u,v)$$

$$b_{22} \triangleq N = \vec{x}_{vv}(u,v)\cdot\vec{n}(u,v)$$

$$b_{12} = b_{21} \triangleq M = \vec{x}_{uv}(u,v)\cdot\vec{n}(u,v)$$

where $\vec{n}(u,v) = \frac{\vec{x}_u\times\vec{x}_v}{\|\vec{x}_u\times\vec{x}_v\|}$, $\vec{x}_{uu}(u,v) = \frac{\partial^2\vec{x}}{\partial u^2}$, $\vec{x}_{vv}(u,v) = \frac{\partial^2\vec{x}}{\partial v^2}$ and $\vec{x}_{uv}(u,v) = \vec{x}_{vu}(u,v) = \frac{\partial^2\vec{x}}{\partial u\partial v}$.

$II(u,v,du,dv)$ measures how the normal vector $d\vec{n}$ changes in the presence of the change in the surface position $d\vec{x}$ at a surface point (u,v) as a function of a infinitesimal movement (d_u,d_v) in the parameter space. Since the second fundamental form depends on a normal to the surface, which is an extrinsic concept, it is an extrinsic property of the surface.

2.1 Curvature Estimation

We can analyze the geometric properties of the principal curvatures of a surface using the *first* and the *second fundamental form*.

We first propose a shape operator

$$[\beta] = [g^{-1}][b]. \tag{6}$$

which is used to compute a Gaussian curvature

$$K = det[\beta]] = \det\left(\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}^{-1}\right)\det\left(\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}\right), \tag{7}$$

and a Mean curvature

$$H = \frac{1}{2}tr[\beta] = \frac{1}{2}\left(\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}^{-1}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}\right). \tag{8}$$

Those entities are also represented through principal curvatures κ_1 and κ_2 as

$$K = \kappa_1\kappa_2 \quad \text{and} \quad H = \frac{(\kappa_1+\kappa_2)}{2}. \tag{9}$$

When a plane is rotated through the normal at a point, the curvature is changed as:

$$\kappa^2 - \sum b_i^i\kappa + \det(b_i^j) = 0, \tag{10}$$

and we obtain a quadratic equation:

$$\kappa^2 - 2H\kappa + K = 0. \tag{11}$$

Principal curvatures are computed using the two roots of a quadratic equation as

$$\kappa_{1,2} = H \pm \sqrt{H^2 - K}. \tag{12}$$

The Gaussian and Mean curvatures with principal curvatures are main entities which characterize the surface type. In Table 1, eight primary types of the surfaces are shown. Each type could be established by looking at the sign of the Gaussian and Mean curvature. The content of the table is also known as HK-map.

	K		
H	+	0	-
-	Peak	Ridge	Saddle ridge
0	none	Flat	Minimal surface
+	Pit	Valley	Saddle valley

Table 1: Surface types from the Gaussian and Mean Curvature.

3 Estimation of Surface Normal

Estimation of surface normals depends on the underlying representation of surface. For point cloud data, one of the most popular estimation method is the analysis of the eigenvectors and eigenvalues of a covariance matrix computed from a point in a neighbourhod of a query point. Then, it is posed as a least-square plane (tangent to a surface) fitting estimation problem [16]. The covariance matrix is computed as

$$\mathcal{C} = \frac{1}{k}\sum_{i=1}^{k} \zeta_i (\mathbf{p}_i - \bar{\mathbf{p}}) \cdot (\mathbf{p}_i - \bar{\mathbf{p}})^T. \tag{13}$$

For a symmetric and positive semi-definite matrix $\mathcal{C}$, its eigenvalues are real numbers $\lambda_j \in \mathbb{R}$ and the eigenvectors $\vec{v}_j$ form an orthogonal frame. The eigenvector $\vec{v}_1$ corresponding to the smallest eigenvalue λ_1 is the sought normal to the surface at a given point.

In case of the meshed clouds, where the triangulation is given, the problem of finding the normal at a vertex of the mesh is simpler. A surface normal for a triangle is computed using the cross product of two edges of that polygon. For a triangle (p_1, p_2, p_3), the normal is computed as

$$\vec{N} = \vec{U} \times \vec{V}, \tag{14}$$

where $\vec{U} = p_2 - p_1$ and $\vec{V} = p_3 - p_1$. The order of the vertices affects the direction of the normal.

3.1 Quantisation of Surface Normals

In order to use surface normals as atomic features at the base layer of a compositional hierarchy, first a set of normals has to be quantized. Cardinality of the set is determined according to the task such that a large set is constructed to minimize distortions for reconstruction of shapes, and a small set is constructed for categorization of shapes.

If we gather all the possible normal vectors, then they will form a unit sphere which is called a Gaussian sphere. If we want to quantize normals, the the Gaussian sphere has to be discretized. One of the popular approaches is using spherical coordinates, which may exhibit some deficiencies. For example, the area of the polygons gets significantly smaller as we approach the poles. Additionally, the poles themselves are singularities.

On of the methods which resolves this challenge is the determination of a minimum energy configuration of N equally charged particles that repel each other with a force given by Coulomb's law and are confined to the surface of the sphere [17]. Using an implementation of the algorithm [1], we were able to find a discretization of a sphere with different density values. An example with 900 particles is shown in Figure 2.a, and a case for 450 particles is shown in Figure 2.b. The uniformity level of sampling could be observed in Figure 2.c and Figure 2.d, where the angular distances of all the particles to a selected particle were calculated. Note that the method described above is computationally complex.

In order to properly discretize the normals, a distance function has to be designed. A bi-invariant distance function is proposed in order measure a distance for rotations of $SO(3)$ [18] as

$$\Phi(R_1, R_2) = \|log(R_1 R)2^T\|, \tag{15}$$

where $log(R) = \frac{\phi}{2\sin\phi}(R - R^T)$, ϕ is the solution of $1 + 2\cos\phi = Tr(R)$ and $\|logR\|^2 = \phi^2$. However, given two Quaternions, (15) is linearly related to [19]

$$\Phi(Q_1, Q_2) = \arccos(\|Q_1 \cdot Q_2\|), , \tag{16}$$

where $Q_1 \cdot Q_2 = w_1 w_2 + x_1 x_2 + y_1 y_2 + z_1 z_2$. This solution has a smaller computational cost.

However an ambiguity is observed when this representation is used. If the rotation axis is represented by the vector, then we compute two vectors which are collinear with this axis. To alleviate this problem, the solution suggested in [20] was used. It could be used to test whether $Q_1 \cdot Q_2$ is negative and negate one of the quaternions to obtain its equivalent alternate using $-Q_1 \cdot Q_2 = Q_1 \cdot -Q_2$.

4 Measuring Distances on Surfaces

So far we have addressed a problem of finding normals for the range data. However, there exists a problem of measuring the distance between parts and compositions of shapes on surfaces. When a surface has a curvature different from zero, the Riemannian metric (geodesic) can be defined. A Euclidean space metric $ds^2 = \sum_i dx_i^2$ could only be used in the case of a flat manifold. The difference between these two metrics is shown in Figure 3.a for a sample object. The red line depicts a distance measured using the Euclidean metric and the green line depicts a distance measured using the Riemannian metric.

4.1 Exact maps

Flat embedding is used to measure a distance between a set of points and areas on the non-flat manifolds. It is a process of mapping a space of a set of points equipped with a geodesic into a finite-dimensional Euclidean space. It is also known as "map-maker-problem". An ideal embedding of the sphere without distortion does not exist, hence there are different types of surface parametrization which preserves different properties, such as length, angle and area preservation [21].

[1] The implementation of this algorithm was provided by Anton Semechko.

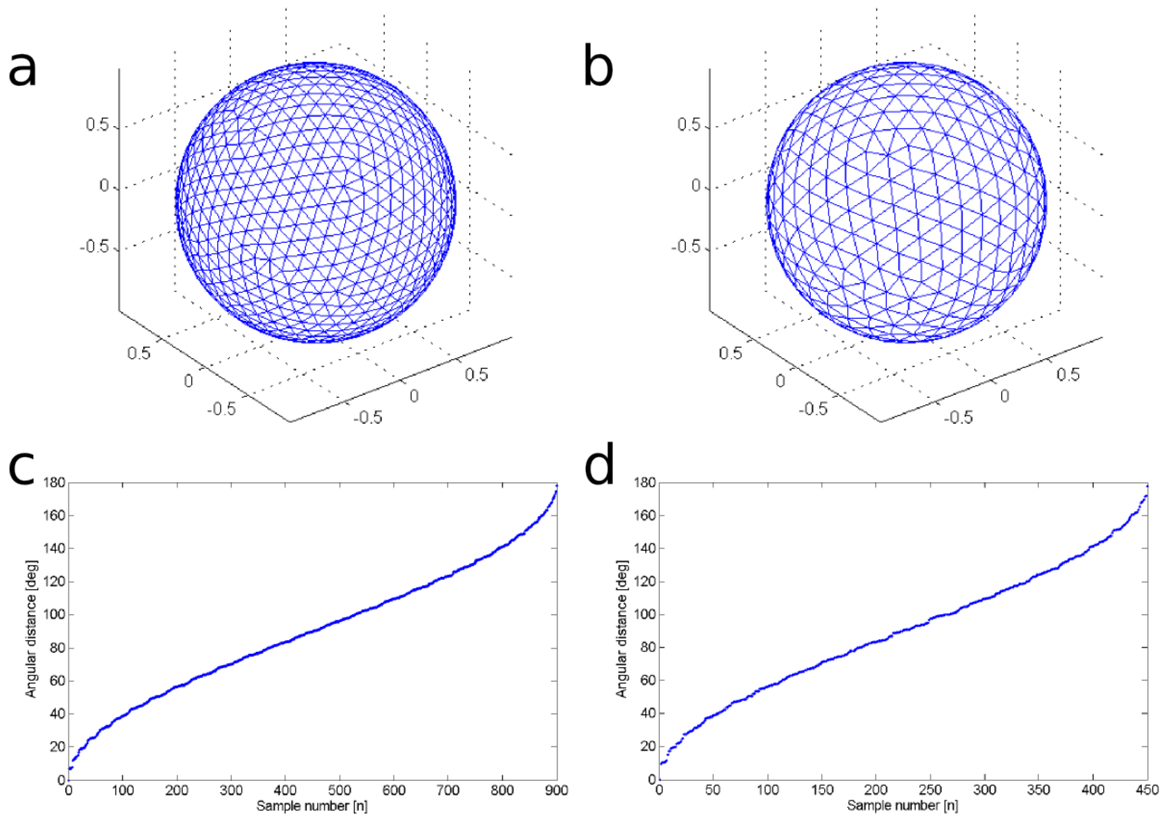


Figure 2: Normals sampling with (a) 900 and (b) 450 particles. Corresponding distances between selected particle and all the other particles for (c) 900 and (d) 450 particles.

The surface parametrization for a sample object is shown in Figure 3.c. For the cone shown in Figure 3.c, its flat embedding was produced by manually indicating the singular points and cut vertices.

4.2 Surface Parametrisation

One of the most popular techniques of intrinsic parametrization of surfaces is triangulation which is employed for minimization of a distortion of the intrinsic measures of the original mesh when a piecewise linear mapping between a 3D mesh and and isomorphic flat mesh is computed [7].

State-of-the-art results of the methods which employs parametrization with quad meshes were presented in [8] and the concept of a cross-field was introduced. A single cross is viewed as a parallelogram which poses four degrees of freedom. Cross field can be generated automatically and the methods are driven by the principal curvature information. The encoding of the cross field is often given in a polar representation where the cross is split into angular and length components.

Generation of regular quad meshes is pertained to generation of a cross field with few singularity points. In order to employ field guided methods for quad mesh generation, first an orientation field (local representation of quad elements) is computed for a given input triangle mesh. Then, sizing field is obtained by computing sampling density. Finally, consistent quad mesh is generated. In order to generate the quadrangulation, global optimization techniques, such as Mixed-Integer Optimization, are used. In the optimization process, a consistent alignment of the polygons is computed. However, it is not possible to obtain the alignment for objects with *complex* geometric properties. Therefore, the minimum number of singularity points which represent the changes in the orientation of

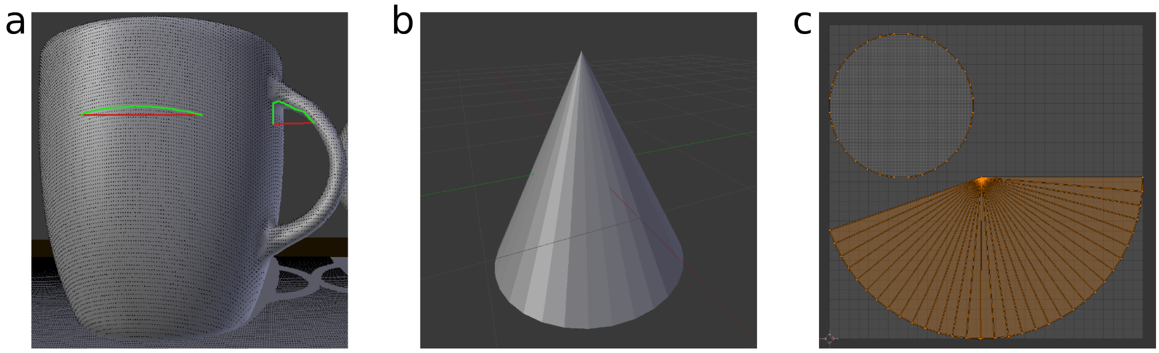


Figure 3: Measuring the distance on a surface. (a) Red line depicts a distance measured using a Euclidean metric, and green line depicts a distance measured using a Riemannian metric. (b) Simple cone and (c) its parametrization.

the surfaces are considered in the optimization process. The latest improvement over this approach is the work proposed in [9]. The authors introduce frame fields, which generalize cross fields by incorporating anisotropy, scaling and skewness, while maintaining the topological structure of a cross field. This improvement allows the user to control not only the alignment of the polygons but also density and anisotropy of their distribution.

4.3 Approximations

A flat embedding which is obtained by parametrization of surfaces is employed to use a Euclidean metric when the object surface is flattened out.

If an exact parametrization of the surface is not needed or it is not possible to obtain it, approximation methods are used. In [22], a method is proposed for computing a geodesic using a given range image. Given two points $\bar{p}_0, \bar{p}_n \in \mathbf{D}$, the distance is defined as

$$d(\bar{p}_0, \bar{p}_n) \approx \sum_{\bar{p}_i \in \mathcal{P}(\bar{p}_0, \bar{p}_n)} \|\mathbf{R}(\bar{p}_i) - \mathbf{R}(\bar{p}_{i+1})\|, \tag{17}$$

where $\mathcal{P}(\bar{p}_0, \bar{p}_n)$ is a set of vertex points residing on the path between $\bar{p}_0$ and $\bar{p}_n$ in the range image. It is assumed that each point obtained from the range image encodes a 3D coordinate of a surface point using a mapping $\mathbf{R} : \mathbf{D} \to \mathbb{R}^3$, where $\mathbf{D} \subset \mathbb{R}^2$. If the path between $\bar{p}_0$ and $\bar{p}_n$ crosses an unsampled point, then the geodesic is defined as infinity. The geodesic is approximated as the length of the line segments of the surface on the path for two points. The approximation holds for small distances. In other case, it is required to perform flat embedding.

5 Geometric Properties of Scale Space

The notion of scale variance is associated with a projection from the 3D world onto a 2D image plane for intensity images. The affect of the size of a support region on modeling the surface type using Mean (H) and Gaussian (K) curvature sign is shown in Figure 4. For a given range image shown in Figure 4.a, the HK-map is computed for a neighborhood with 9 pixels (Figure 4.b), 11 pixels (Figure 4.c), 13 pixels (Figure 4.e), and 15 pixels (Figure 4.f). The surface types are shown in table (Figure 4.e) as a reference. The relationship between the size of the support region and the surface type is analyzed by Unnikrishnan et al. [23] using a Multi-Scale Interest Region (MSIR) approach. However, the proposed MSIR is not accurate for basic shapes and is not stable for object recognition

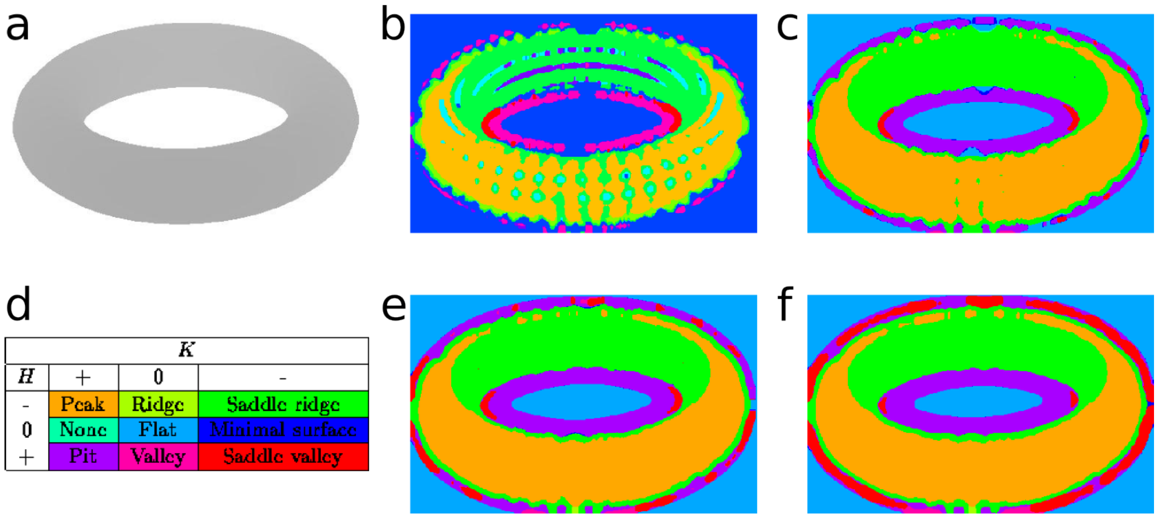


Figure 4: Influence of the support region size on the surface type estimation. (a) A range image of the sample object. The HK-maps are computed for a support region of size (b) 9 pixels, (c) 11 pixels, (e) 13 pixels, (f) 15 pixels. Surface types classified according to the sign of the Mean and Gaussian curvature (d).

[24]. They provided an experimental analysis of Shape Index [25] which is computed using principal curvatures.

5.1 Scale space in 2D

We first briefly describe the concepts of the automatic scale selection proposed by Lindeberg [26]. The notation would follow [27]. Scale selection refers to methods which are used for estimation of characteristic scales in the images. Multi-scale representation of the image data is achieved through embedding the original signal into a one-parameter family of signals using scale as the parameter. For a given N-dimensional signal $f : \mathbb{R}^N \to \mathbb{R}$ the scale space representation of f is defined by the convolution operation

$$L(x;t) = \int_{\xi \in \mathbb{R}^N} f(x-\xi) g(\xi, t) d\xi \tag{18}$$

where $g : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}$ denotes a Gaussian kernel

$$g(x.t) = \frac{1}{(2\pi t)^{N/2}} e^{-|x|^2/2t}. \tag{19}$$

The variance $t = \sigma^2$ of this kernel is referred as a scale parameter. Gaussian derivatives could be computed by differentiating the scale space representation or by convolving the original image with the Gaussian derivative kernels as

$$L_{x^\alpha}(\cdot;t) = \partial_{x^\alpha} L(\cdot;t) = (\partial_{x^\alpha} g(\cdot;t)) * f(\cdot), \tag{20}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $\partial_{x^\alpha} = \partial_{x_1^{\alpha_1}} \partial_{x_2^{\alpha_2}} \ldots \partial_{x_N^{\alpha_N}}$.

5.2 Scale space in 3D

Scale selection methods proposed for 2D images can be used for processing 3D data by replacing intensity values with 3D vertex coordinates of the model. Such attempts are reported in [22], where two main problems with this approach were indicated. First,

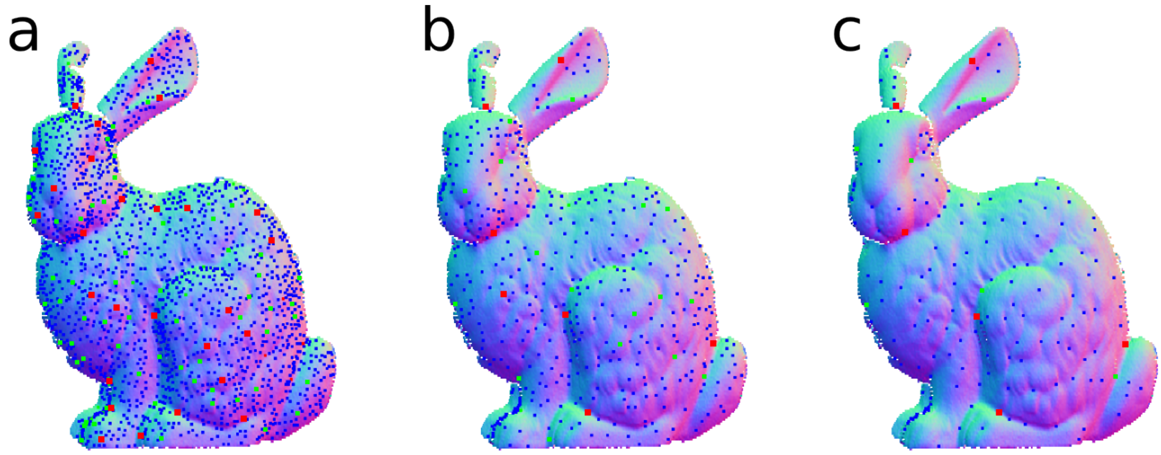


Figure 5: Computation of corners in different geometric scale. Blue dots depict the corners which are detected at the finest scale, and the red dots depict the corners which are detected at the coarsest scale. Corners detected using non-maxima suppression with radius (a)1%, (b) 2% and (c) 3%. Radius is defined as the percentage of the axis length of the largest bounding box of the object.

direct smoothing of 3D points modifies the extrinsic geometry of the original model. In consequence, change of the global topology of the data may occur e.g. fragmentation of original model. Second, most of the methods use Euclidean distance between 3D points. This can lead to the incorrect features in the scale space caused by the topological changes within the support region e.g. two surface of different parts could lie close to each other using Euclidean 3D distance metric like point at the body of the mug and at the handle (Figure 3a – red line) but the correct geodesic distance (Figure 3a – green line) is substantially larger. The topology of the shape was taken into account.

Nishino and co-authors in the series of publications concluded with the journal article [22] investigated thoroughly the scale space in 3D. They proposed correct measures of the distance and the the algorithms, which exploits those metrics, for object registration and object recognition. The main focus of their work is on the 2D representation of the 3D surface geometry which allows to work with geodesics, which in consequence leads to the results which are consistent with mathematical framework of the scale selection presented in section 5.1.

The search of the features for different local scales is performed by convolving range data with Gaussian kernel of different radius, but importantly, this radius is calculated on the surface hence using geodesic distances. Authors are also providing the software implementation of their algorithms. This allowed us to have a closer look at the properties of the approach.

For a given range image, corners were visualized on different local scales in Figure 5. Blue dots depict the corners which are detected at the finest scale, and the red dots depict the corners which are detected at the coarsest scale. As it is shown in Figure 5.a, low non-maxima suppression level causes the abundant number of fine scale corners, since the fine scale corners could have been formed from a noise. Whereas, for a high non-maxima suppression level (Figure 5.c), the number of blue dots is significantly lower and the number of red dots remains the same. In other words, the patterns observed at the corners represent larger parts of shape of an object which are locally unique and not influenced by noise at the coarser scales.

6 Conclusion

We have provided a conceptual overview of the state of the art methods employed for the analysis of geometric properties of shapes. To summarize, the impact of the presented concepts on the compositional hierarchy formation process are threefold:

- **Designing low-level representations:** The knowledge of the surface parametrization is indispensable to obtain an approximate discretization of shape of an object at the first layer. Singularity points should be considered while designing distance functions of parts of shapes in order model hierarchical compositional representations.
- **Modeling shapes at different scales:** Each layer of a compositional hierarchy operates on a different scale. Hence, the results achieved in scale space theory of 3D shapes have provided a basis for modeling the abstraction of the perceived object at different layers of the compositional hierarchy.
- **Part selection in compositional hierarchies:** The results obtained from the analysis of geometric properties of shapes should be employed for modeling representations which can be used for the development of robotic object manipulation and grasping algorithms.

References

[1] P. J. Besl, *Surfaces in Range Image Understanding*, ser. Springer Series in Perception Engineering. Springer-Verlag, 1988.

[2] O. Faugeras, *Three-Dimensional Computer Vision. A geometric view point.* MIT Press, 1993.

[3] A. Leonardis, A. Gupta, and R. Bajcsy, "Segmentation of range images as the search for geometric parametric models," *Int. J. of Computer Vision*, vol. 14, no. 3, pp. 253–277, 1995.

[4] T.-W. R. Lo and J. P. Siebert, "Local feature extraction and matching on range images: 2.5D {SIFT}," *Computer Vision and Image Understanding*, vol. 113, no. 12, pp. 1235–1250, 2009, special issue on 3D Representation for Object and Scene Recognition.

[5] C. Redondo-Cabrera, R. Lopez-Sastre, J. Acevedo-Rodriguez, and S. Maldonado-Bascon, "SURFing the point clouds: Selective 3D spatial pyramids for category-level object recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3458–3465.

[6] C. Bennis, J.-M. Vézien, and G. Iglésias, "Piecewise Surface Flattening for Non-distorted Texture Mapping," *SIGGRAPH Comput. Graph.*, vol. 25, no. 4, pp. 237–246, Jul. 1991. [Online]. Available: http://doi.acm.org/10.1145/127719.122744

[7] M. Desbrun, M. Meyer, and P. Alliez, "Intrinsic Parameterizations of Surface Meshes," *Computer Graphics Forum*, vol. 21, no. 3, pp. 209–218, 2002. [Online]. Available: http://dx.doi.org/10.1111/1467-8659.00580

[8] D. Bommes, M. Campen, H.-C. Ebke, P. Alliez, and L. Kobbelt, "Integer-grid Maps for Reliable Quad Meshing," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 98:1–98:12, Jul. 2013. [Online]. Available: http://doi.acm.org/10.1145/2461912.2462014

[9] D. Panozzo, E. Puppo, M. Tarini, and O. Sorkine-Hornung, "Frame Fields: Anisotropic and Non-orthogonal Cross Fields," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 134:1–134:11, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2601097.2601179

[10] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3D Object Recognition in Cluttered Scenes with Local Surface Features: A Survey," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 36, no. 11, pp. 2270–2287, 2014.

[11] T.-H. Yu, O. Woodford, and R. Cipolla, "A Performance Evaluation of Volumetric 3D Interest Point Detectors," *Int. J. of Computer Vision*, vol. 102, no. 1-3, pp. 180–197, 2013.

[12] C. Akgul, B. Sankur, Y. Yemez, and F. Schmitt, "3D Model Retrieval Using Probability Density-Based Shape Descriptors," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 31, no. 6, pp. 1117–1133, 2009.

[13] S. Fidler and A. Leonardis, "Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[14] B. Ommer, M. Sauter, and J. M. Buhmann, "Learning Top-Down Grouping of Compositional Hierarchies for Recognition," in *Proceedings of the 2006 Conf. on Computer Vision and Pattern Recognition Workshop*, ser. CVPRW '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 194–211.

[15] L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille, "Part and appearance sharing: Recursive Compositional Models for multi-view," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1919–1926.

[16] R. B. Rusu, "Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments," Ph.D. dissertation, Computer Science department, Technische Universitaet Muenchen, Germany, October 2009.

[17] E. Saff and A. Kuijlaars, "Distributing many points on a sphere," *The Mathematical Intelligencer*, vol. 19, no. 1, pp. 5–11, 1997. [Online]. Available: http://dx.doi.org/10.1007/BF03024331

[18] F. C. Park and B. Ravani, "Smooth Invariant Interpolation of Rotations," *ACM Trans. Graph.*, vol. 16, no. 3, pp. 277–295, Jul. 1997.

[19] D. Huynh, "Metrics for 3D Rotations: Comparison and Analysis," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 2, pp. 155–164, 2009.

[20] J. Kuffner, "Effective sampling and distance metrics for 3D rigid body path planning," in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 4, April 2004, pp. 3993–3998.

[21] M. Floater and K. Hormann, "Surface Parameterization: a Tutorial and Survey," in *Advances in Multiresolution for Geometric Modelling*, ser. Mathematics and

Visualization, N. Dodgson, M. Floater, and M. Sabin, Eds. Springer Berlin Heidelberg, 2005, pp. 157–186. [Online]. Available: http://dx.doi.org/10.1007/3-540-26808-1_9

[22] P. Bariya, J. Novatnack, G. Schwartz, and K. Nishino, "3D Geometric Scale Variability in Range Images: Features and Descriptors," *International Journal of Computer Vision*, vol. 99, no. 2, pp. 232–255, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11263-012-0526-7

[23] R. Unnikrishnan and M. Hebert, "Multi-scale interest regions from unorganized point clouds," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, June 2008, pp. 1–8.

[24] U. Bonde, V. Badrinarayanan, and R. Cipolla, "Multi Scale Shape Index for 3D Object Recognition," in *Scale Space and Variational Methods in Computer Vision*, ser. Lecture Notes in Computer Science, A. Kuijper, K. Bredies, T. Pock, and H. Bischof, Eds. Springer Berlin Heidelberg, 2013, vol. 7893, pp. 306–318. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-38267-3_26

[25] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing*, vol. 10, no. 8, pp. 557–564, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/026288569290076F

[26] T. Lindeberg, "Feature Detection with Automatic Scale Selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998. [Online]. Available: http://dx.doi.org/10.1023/A%3A1008045108935

[27] ——, "Scale Selection," in *Computer Vision*, K. Ikeuchi, Ed. Springer US, 2014, pp. 701–713. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-31439-6_242

Statistical Learning of Multi-modal Representations of Objects using Multi-view 2D and View-based 3D Models

Technical Report
University of Innsbruck

Mirela Popa, Mete Ozay, Vladislav Kramarev, Rusen Aktas, Krzysztof Walas, Justus Piater, Ales Leonardis, and Jeremy Wyatt

February 28, 2015

1 Introduction

Visual object recognition and categorization has many potential applications, enabling scene interpretation, semantic visual retrieval, and cognitive robotics. Furthermore, efficient object representations can be useful in a number of robotic application domains, such as object manipulation by a robotic arm, including tasks such as loading and unloading a dishwasher, grasping objects or opening a door.

In recent years, many approaches have been proposed for object class detection using 2D features from single-view, such as the deformable part model [1], but also multi-view based methods which are useful at predicting not only the category but also the pose of an object in a scene [2]. Furthermore, the availability of depth sensors such as Kinect had a great impact on rapid development of algorithms for view-based 3D object models construction and acquisition. Employing view-based 3D models for object recognition is useful, as they capture the intrinsic geometric structure of objects and provide a compact object representation [3].

Provided that 2D and view-based 3D information are complementary, we aim to combine them in an efficient manner, enabling improved recognition performance. There are different fusion methods available, such as early-level fusion, where different features are concatenated at an early stage, or decision-level fusion, case in which the recognition is performed individually for each input channel and the final decision is taken by combing the independent results, using different criteria such as maximum or weighted sum [4].

Any of these methods is not sufficient for our goals, as we aim to take advantage of the correlations between 2D and view-based 3D information not only at the beginning or at the end of the recognition process, but to model them in a hierarchical manner, by considering relations between layers in the 2D and the view-based 3D hierarchies. Therefore, in this work, we propose using both 2D and view-based 3D primitives, constructed using hierarchical representations, in which the effect of low-level visual features to higher-level entities is learned, leading to detection of object models in a scene.

The principles of hierarchy construction have been introduced in [5] for 2D and in [6] for view-based 3D data. The relations between the two types of information will be captured using an efficient probabilistic model. The model provides better object representations, by reinforcing the evidence provided by one input channel, while dealing in a better way with the limitations of each input primitive type. Furthermore, it can be employed in the case of missing observations, for inferring the most likely view-based 3D primitives from observed 2D ones and vice-versa.

In the following section we present a review of related works, while in section 3 we introduce our model for capturing correspondences between 2D and view-based 3D primitives.

2 Related Work

One approach towards fusing 2D and 3D information consists of projecting a 3D shape into 2D and finding a set of meaningful multi-view 2D projections [7], option supported by [8] which argue that human perception of 3D shapes is based on 2D observations. The relation between 2D and 3D is exploited for 3D shape segmentation. First, the 2D projections of a 3D shape are labeled, then they are transferred to the 3D shape via back-projection, while finally the integration of multiple labels is done for obtaining a unified segmentation. Projection of 3D information into 2D as a means to recognize 3D objects and their pose is not new in computer vision and has been investigated also by [9]. This approach is justified by the simplification of the problem, obtained by working in a low-dimensional space and by making use of the bigger volume of 2D image data available in the online environment. The disadvantages of this method reside in the complexity of the process, which requires a manual labelling of parts, an assumption regarding the upright orientation of objects, as well as a slightly worse performance on articulated objects in comparison with rigid ones.

Another approach for capturing 2D to view-based 3D correlations is described by [10]. In this work, the assumption that edge information in an image has a high probability of being the edge of the depth map is used. Next, an image is divided into blocks and depth is assigned based on neighbourhood information and smoothed using a bilateral filtering.

An adaptive fusion approach of 2D and 3D features is proposed by [11] with an application in face recognition. The integration of the two modalities is beneficial, as 2D features are more precise at capturing facial details, while 3D features are more robust under pose or lighting variations. The two types of features are fused at an early stage, while in the recognition process, only the most reliable features are used. Discarding of unreliable features is achieved using a matching criterion, which assigns weights based on a similarity measure. The same topic is investigated also in [12], where the fusion strategy consists of an offline and an online weight learning process. In both cases the most relevant weights of all the scores for each sample in each modality is automatically selected.

3D object retrieval based on a 2D image represent another source of inspiration for linking 2D and 3D information. In [13], the authors propose using a sketch representation based on optimized Gabor filters, followed by a bag-of-words approach. The feature extraction technique can be further refined into a hierarchical representation using concepts proposed by [14], which guides the learning of a dictionary of templates in an unsupervised manner. Each template is a composition of Gabor wavelets which are allowed to shift their location and orientation relative to each other. The invariance property is achieved using the active basis model [15].

The review of the works dealing with 2D and view-based 3D information showed that fusing them is beneficial, while a great care needs to be given to challenging issues, such as conflicting data, outliers and noise, or data registration and correlation assessment. Without a proper pre-analysis of all these issues, the success rate of the fusion algorithm can be affected. In the next section, we introduce our proposed probabilistic model.

3 Statistical Modeling of Multi-modal Representations

Our goal consists of matching 2D and view-based 3D parts in order to permit inference of 2D parts from observed view-based 3D parts and vice-versa and also to achieve better object recognition and/or object pose recognition accuracy.

We assume that the 2D and the view-based 3D hierarchical representations are based on similar principles and encode spatial locality and co-occurrence of features, while the difference consists in the different types of views associated with the two representations.

Furthermore, we consider that each hierarchical representation is composed of a number of layers and a number of parts on each layer. Another assumption that we use is that the two hierarchies are trained on the same type of data consisting of both 2D images and the corresponding depth images or point clouds.

We analyze the statistical relationship between representations obtained from 2D models computed using Compositional Hierarchy of Parts (CHOP) algorithm [5] and view-based 3D models computed using a 3D compositional hierarchy [6].

The hierarchical compositional 2D model is formed of n_{2D} layers, each layer being composed of a number of 2D parts.

We denote the i^{th} part constructed at the l^{th} layer of a hierarchical vocabulary of 2D shapes as Γ_i^l, where $i \in \{1, \ldots, n_l\}, l \in \{1, \ldots, n_{2D}\}$. The j^{th} part constructed at the l'^{th} layer of a hierarchical vocabulary of 3D shapes as $\Omega_j^{l'}$, where $j \in \{1, \ldots, n_{l'}\}, l' \in \{1, \ldots, n_{3D}\}$.

3.1 Statistical Analysis of the Relationship between Multi-view 2D and View-based 3D Models

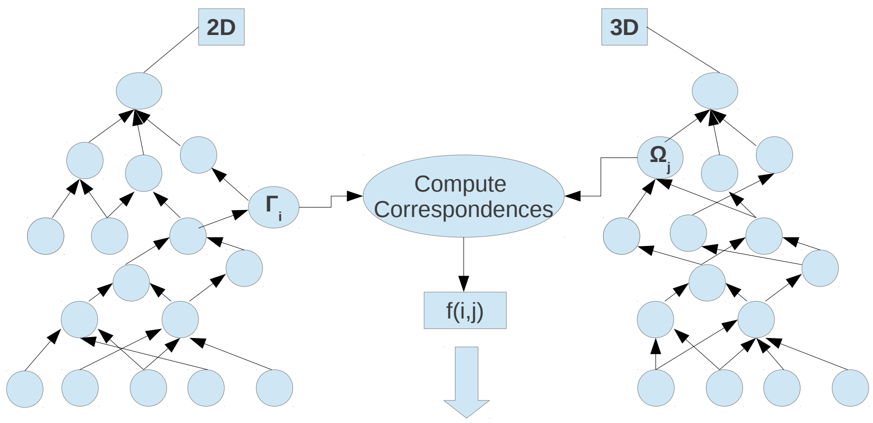
For establishing correspondences between the parts in the two hierarchies, we build a *correspondence table* with row indices $(1, \ldots, N)$, (where $N = \sum_{l=1}^{n_{2D}} n_l$) corresponding to the 2D parts and columns indices $(1, \ldots, M)$, where $M = \sum_{l'=1}^{n_{3D}} n_{l'}$) corresponding to the view-based 3D parts, as depicted in Figure 1 below.

Ideally, we would like to establish level-by-level correspondences between 2D and view-based 3D parts. However, since the two hierarchies are constructed in distinct fashions, there are no directly corresponding layers. Thus, in the most general sense, one might consider correspondences between all pairs of 2D and view-based 3D parts $((\Gamma_i, \Omega_j), i \in (1, \ldots, N), j \in (1, \ldots, M))$, irrespective of which layers they belong to. However, one heuristic that can be used to restrict the number of parts associations would be to take into consideration the sizes of the parts in the two hierarchies.

The correspondences between the two types of parts are build using an observation function f, which quantizes the occurrences of one 2D part in the neighbourhood of a view-based 3D part and vice-versa. For computing the distance between parts in the two hierarchies, we use 2D projections of view-based 3D parts in the Euclidian space.

Next, we illustrate the parts binding mechanism in Figure 1.

To every correspondence $f(i, j)$, where i is the index of the 2D part Γ_i and j the index of the view-based 3D part Ω_j, we assign a real value. As for each 3D/2D part there might be several 2D/3D parts corresponding to it, all associated correspondences will be activated. Therefore, the value $f(i, j)$ is



f	1	...	j	...	M
1		0.4			0.2
...	0.1				
i			**0.6**		0.1
...	0.3				
N		0.4		0.2	

Figure 1: Correspondences between the 2D and view-based 3D hierarchies.

computed as the joint probability of observing parts Γ_i and Ω_j, in the same neighbourhood $N_1(i,j)$. We define the neighbourhood $N_1(i,j)$ as the set of all 2D and 3D parts which are close to the center $C_{\Gamma_i} = (c_{x\Gamma_i}, c_{y\Gamma_i})$ of the 2D part Γ_i and the center $C_{\Omega_j} = (c_{x\Omega_j}, c_{y\Omega_j})$ of the 2D projection of the 3D part Ω_j. The notion of closeness is defined using the Euclidian distance $d(\Gamma_i, \Omega_j) = \sqrt{(c_{x\Gamma_i} - c_{x\Omega_j})^2 + (c_{y\Gamma_i} - c_{y\Omega_j})^2}$, such as $d(\Gamma_i, \Omega_j) < \tilde{d}$. The parameter $\tilde{d}$ is defined, using the same distances considered when grouping sub-parts into parts for constructing the 2D and 3D hierarchical representations.

$$f(i,j) = P(\Gamma_i, \Omega_j), (\Gamma_i, \Omega_j \in N_1(i,j)). \tag{1}$$

In general, there will not be a one-to-one relation between 2D and view-based 3D parts, as the 2D views depend on viewpoint. The correspondence table represents the set of possible correspondences over all viewpoints.

Moreover, the joint probability table of 2D and view-based 3D parts can be used for performing inference of view-based 3D parts/2D parts coherently and transparently across the 2D and view-based 3D hierarchies.

Still, the inference process can be regraded from two perspectives, given the type of parts we want to predict, either 2D or 3D.

Therefore, for he task of inferring a 2D part Γ_i from an observed view-based 3D part Ω_j, we employ the conditional probability:

$$P(\Gamma_i|\Omega_j), i \in (1, \ldots, N), \Gamma_i \in N_2(j), \tag{2}$$

operation, which will retrieve all 2D parts Γ_i which are probable to be observed in the neighbourhood $N_2(j)$ of the view-based 3D part Ω_j. The neighbourhood $N_2(j)$ consists of all the 2D parts which are close to the 2D projection of the 3D part Ω_j in the Euclidian space.

The relation to equation (1) is build using the Bayes rule such as:

$$P(\Gamma_i|\Omega_j) = \frac{P(\Gamma_i, \Omega_j)}{P(\Omega_j)}. \tag{3}$$

Respectively, inference of a view-based 3D part Ω_j can be performed using the conditional probability:

$$P(\Omega_j|\Gamma_i), j \in (1, \ldots, M), \Omega_j \in N_3(i), \tag{4}$$

which permits retrieving all 3D parts Ω_j which could be observed in the neighbourhood $N_3(i)$ of the 2D part Γ_i. The neighbourhood $N_3(i)$ consists of all the 3D parts whose 2D projections are close to the 2D part Γ_i in the Euclidian space.

Furthermore, the relation to equation (1) is given by:

$$P(\Omega_j|\Gamma_i) = \frac{P(\Omega_j, \Gamma_i)}{P(\Gamma_i)}. \tag{5}$$

The correspondence table introduced above can be considered as a starting point for different tasks, such as inferring view-based 3D information from observed 2D parts, and respectively inferring 2D information from observed view-based 3D parts, object recognition, object reconstruction, or pose estimation.

3.2 Inference of View-based 3D Parts using 2D and 3D Hierarchical Compositional Models

We propose a probabilistic model for performing inference in the view-based 3D hierarchy, by reusing the inference mechanism existing in the 2D hierarchy in combination with the joint probability table introduced above.

We use the assumption that parts in the view-based 3D hierarchy are formed by propagating evidence from lower levels to higher ones in a probabilistic manner.

Next, we condition the probability of a view-based 3D part Ω_j^l at layer l not only on the the sub-parts on the previous layer $l-1$, but also on the 2D parts that are probable to co-occur with it. We will replace the conditional probability:

$$P(\Omega_j^l|\Omega_i^{l-1}), i \in (1, \ldots, n_{l-1}), \tag{6}$$

with:

$$P(\Omega_j^l|\Omega_i^{l-1}, \Gamma_k), i \in (1, \ldots, n_{l-1}). \tag{7}$$

2D parts Γ_k are retrieved using the joint probability table introduced in Section 3.1 and consist of the set of 2D parts which maximize the $P(\Gamma_k, \Omega_j)$. The threshold for part selection τ, where $P(\Gamma_k, \Omega_j) > \tau$ is learned according to the purpose of the inference process.

Furthermore, according to Bayes rule we can estimate the conditional probability introduced in (7) as follows:

$$P(\Omega_j^l|\Omega_i^{l-1},\Gamma_k) = \frac{P(\Omega_i^{l-1},\Gamma_k|\Omega_j^l)P(\Omega_j^l)}{P(\Omega_i^{l-1},\Gamma_k)}. \tag{8}$$

For estimating the following conditional probability:

$$P(\Omega_i^{l-1},\Gamma_k|\Omega_j^l) = P(\Gamma_k|\Omega_j^l,\Omega_i^{l-1})P(\Omega_i^{l-1}|\Omega_j^l), \tag{9}$$

we take into account the introduced definition of neighbourhood, which suggests that if the pairs of parts (Γ_k,Ω_j^l) and $(\Omega_j^l,\Omega_i^{l-1})$ are in the same neighbourhood and can be correlated, then the parts $(\Gamma_k,\Omega_i^{l-1})$ are conditionally independent, given part Ω_j^l, as the correspondences between parts are established at the highest possible layers in the two hierarchies and therefore:

$$P(\Omega_i^{l-1},\Gamma_k|\Omega_j^l) = P(\Gamma_k|\Omega_j^l)P(\Omega_i^{l-1}|\Omega_j^l). \tag{10}$$

By introducing equation (10) into (8) we obtain:

$$P(\Omega_j^l|\Omega_i^{l-1},\Gamma_k) = \frac{P(\Gamma_k|\Omega_j^l)P(\Omega_i^{l-1}|\Omega_j^l)P(\Omega_j^l)}{P(\Omega_i^{l-1},\Gamma_k)}, \tag{11}$$

where the conditional probability $P(\Gamma_k|\Omega_j)$ can be obtained from equation (3) and $P(\Omega_i^{l-1}|\Omega_j^l)$ can be computed using the initial conditional probability introduced in (6) as follows:

$$P(\Omega_i^{l-1}|\Omega_j^l) = \frac{P(\Omega_j^l|\Omega_i^{l-1})P(\Omega_i^{l-1})}{P(\Omega_j^l)}. \tag{12}$$

Finally, by introducing equation (12) into (11) we obtain:

$$P(\Omega_j^l|\Omega_i^{l-1},\Gamma_k) = \frac{P(\Gamma_k|\Omega_j^l)P(\Omega_j^l|\Omega_i^{l-1})P(\Omega_i^{l-1})}{P(\Omega_i^{l-1},\Gamma_k)}. \tag{13}$$

The view-based 3D part prior $P(\Omega_i^{l-1})$ can be considered uniform or computed using the inverse frequency operator on an input database, which takes into account the part frequency across the database, while the joint probability $P(\Omega_i^{l-1},\Gamma_k)$ cane be obtained from equation (1) presented in Section 3.1.

The proposed model will be initialized with probabilities of 2D parts on the first layer of the 2D hierarchy, which correspond to feature responses, (features represented in this case by a set of Gabor filters). Next, the observations in the 2D hierarchy will be used to propagate information in the view-based 3D hierarchy and finally to obtain hypotheses about view-based 3D parts.

Next, we analyze a set of statistical properties of the correspondences between parts in the 2D and view-based 3D hierarchies. This analysis will be used to obtain a better understanding of the benefits achieved by fusing them.

3.3 Statistical Evaluation

The purpose of the statistical evaluation of the correspondences between the two information channels 2D and 3D is to assess the relations which can be formed between them and furthermore, to draw conclusion regarding the efficiency of the inference process.

For quantizing correspondences between parts in the 2D and 3D hierarchies, we use several statistical measures.

First, we analyze the correspondences of each 2D/3D part, given the set of 3D/2D parts, by constructing a pdf over all computed conditional probabilities of that part. A peaked pdf suggests that the considered 2D/3D part can be matched successfully to a part or to a limited set of 3D/2D parts, while a multi-peak distribution indicates that more information is needed in order to predict a good matching. The pairs of parts which can be matched with a lower uncertainty between the two hierarchies, should be given a higher weight in the inference process, as they are better at predicting the unknown part in the other hierarchy.

Next, we compute the conditional entropy of 2D parts given 3D parts and vice-versa, using the conditional probabilities introduced in equations (2) and (4). This measure computes the amount of uncertainty remaining in predicting 2D parts given 3D parts and vice-versa, thus helping us to determine which information, either 2D or 3D, is better at predicting the other type of information in the inference process.

$$H(\Delta|\Theta) = -\sum_{j=1}^{M} P(\Omega_j) \sum_{i=1}^{N} P(\Gamma_i|\Omega_j) \log P(\Gamma_i|\Omega_j), \tag{14}$$

$$H(\Theta|\Delta) = -\sum_{i=1}^{N} P(\Gamma_i) \sum_{j=1}^{M} P(\Omega_j|\Gamma_i) \log P(\Omega_j|\Gamma_i), \tag{15}$$

where Δ consists of all 2D parts $\{\Gamma_i, i \in (1, \ldots, N)\}$ and Θ consists of all the 3D parts $\{\Omega_j, j \in (1, \ldots, M)\}$.

Moreover, we also analyze the mutual information of the parts in the 2D and 3D compositional representations, as a measure of the parts mutual dependence, by employing the joint probabilities introduced in equation (1).

$$I(\Gamma|\Omega) = \sum_{j=1}^{M} \sum_{i=1}^{N} P(\Gamma_i, \Omega_j) \log \frac{P(\Gamma_i, \Omega_j)}{P(\Gamma_i)P(\Omega_j)}. \tag{16}$$

In the following section we provide the results of the performed statistical evaluation.

4 Experimental Results

We evaluated the learned correspondences between parts in the 2D and view-based 3D hierarchies on a database of standard shapes consisting of cubes, cones, spheres, etc., from different view-points, which is illustrated in Figure 2.

Our aim consists in quantifying the predictive power of the learned correspondences between 2D and 3D information channels, at each layer of their compositional representations, aim which is assessed using multiple statistical measures.

We performed the analysis in a bottom-up manner, starting from the second layer in both hierarchies and proceeding up to Layer 4. We decided to start at Layer 2, as the first layer in both hierarchies is not informative for building correlations, consisting of six Gabor filters in the 2D hierarchy and respectively, nine depth oriented edges in the 3D hierarchy.

We depict in Figure 3, the obtained correspondence matrix for parts at Layer 2 in both hierarchies. As it can be observed, at this level there are many-to-many correspondences, suggesting that matching parts on this layer is ambiguous.

We continue the analysis at Layer 3. A selection of pairs of parts, which are correlated with a high probability are shown in Figure 4, proving that meaningful correspondences can be learned.

Moreover, we inspected visually the learned correspondences in order to interpret the results of the statistical analysis and we depict in Figure 5 the instantiations on training images of a set of pairs of co-occurring parts (2D,3D).

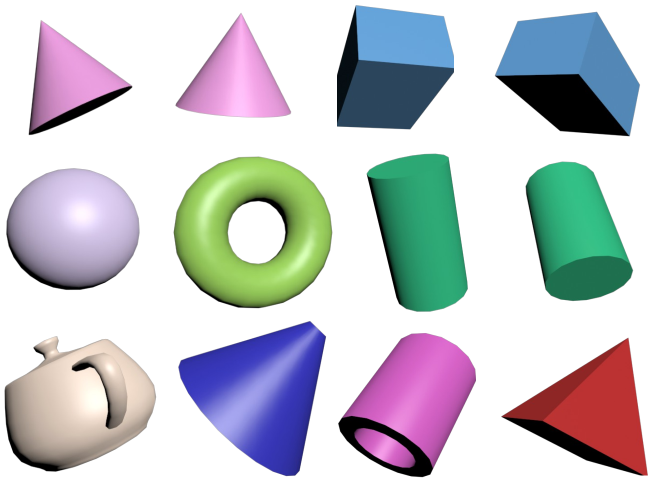


Figure 2: Examples of standard shapes from the ShapeDatabase.

Finally, for Layer 4, we only illustrate the matrix of statistically meaningful correspondences in Figure 6 and highlight the rows corresponding to specific 2D parts and the columns corresponding to specific 3D parts which are co-occurring frequently. In this case, the relation of one 3D part to many 2D parts is depicted, corresponding to a corner in both hierarchies, while the 2D parts capture it from different viewpoints.

Another important information which can be extracted from the learned correspondences between 2D and 3D parts refers to the type of parts which do not co-occur in the same neighbourhood. We depict in Figure 7 examples of pairs of parts at Layer 4 which are not occurring together, showing that the learned statistics between parts are meaningful in both directions.

In our strive for assessing part correspondences, we should take into account the large number of parts at each Layer (see Figure 8 for a subset of 2D and 3D parts at Layer 4), many of each are similar to some extent, fact which influences the spread of the distributions of correspondences over many similar parts. In other words, the layer-by-layer predictability measures are heavily influenced by the raw number of parts, in addition to predictability.

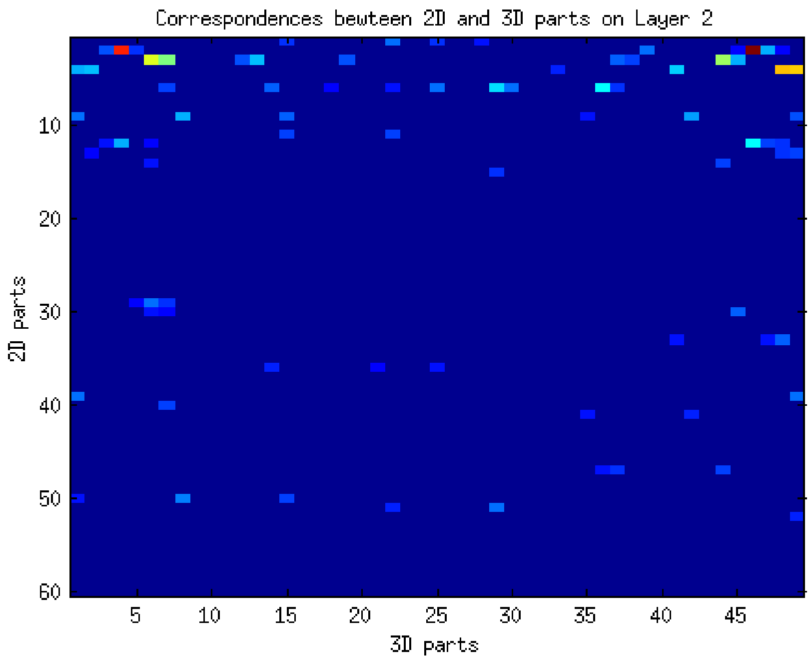


Figure 3: Correspondences between parts at Layer 2.

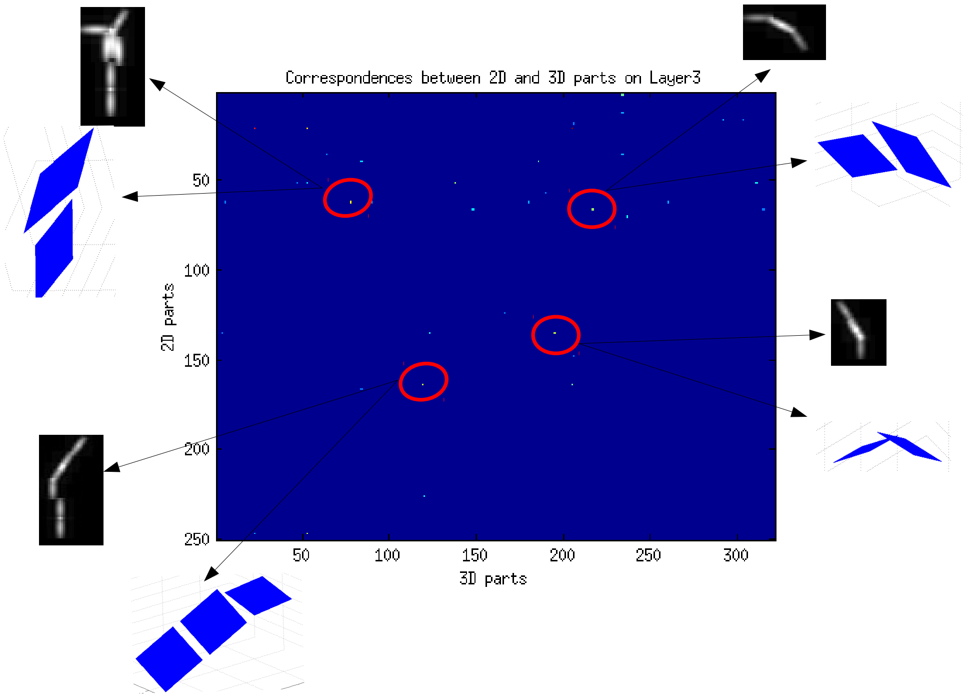


Figure 4: Correspondences between parts at Layer 3.

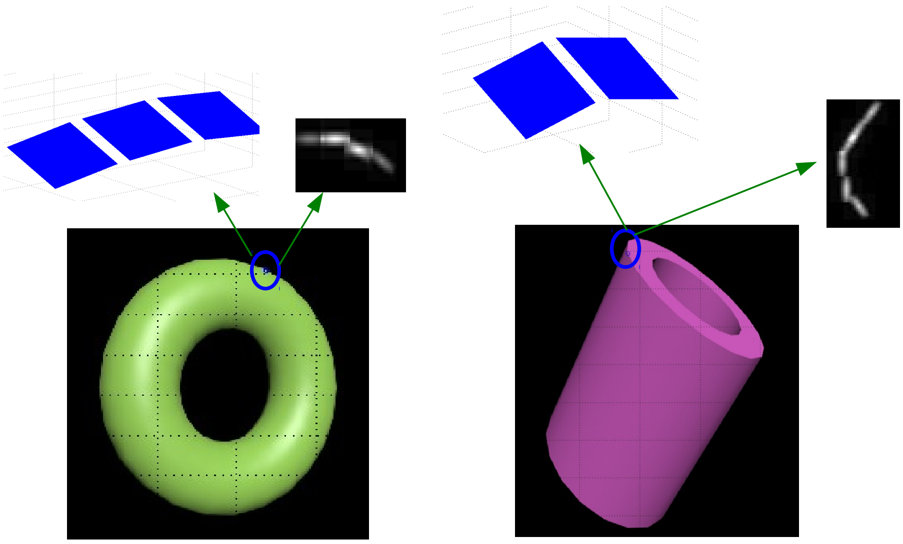


Figure 5: Examples of pairs of parts(2D,3D) instantiated in training images at Layer 3.

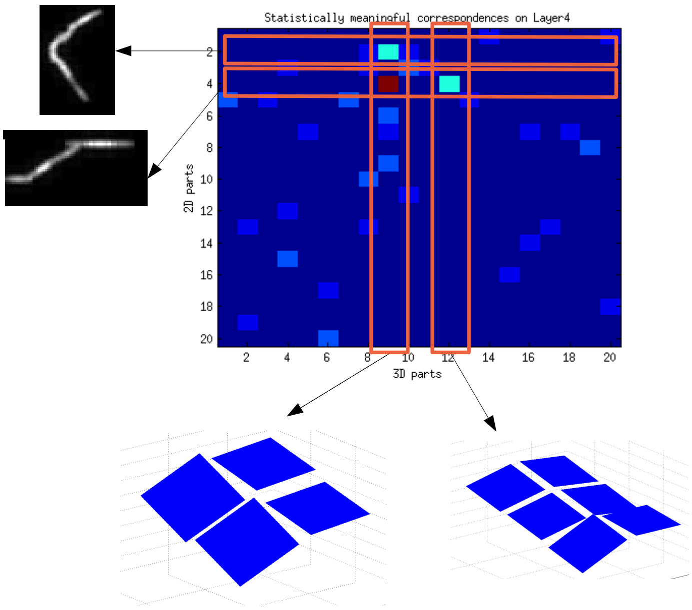


Figure 6: Correspondences between parts at Layer 4.

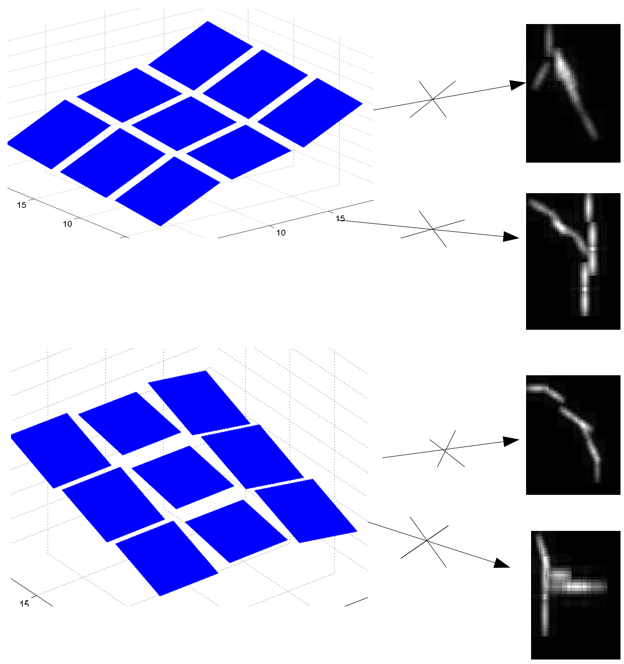
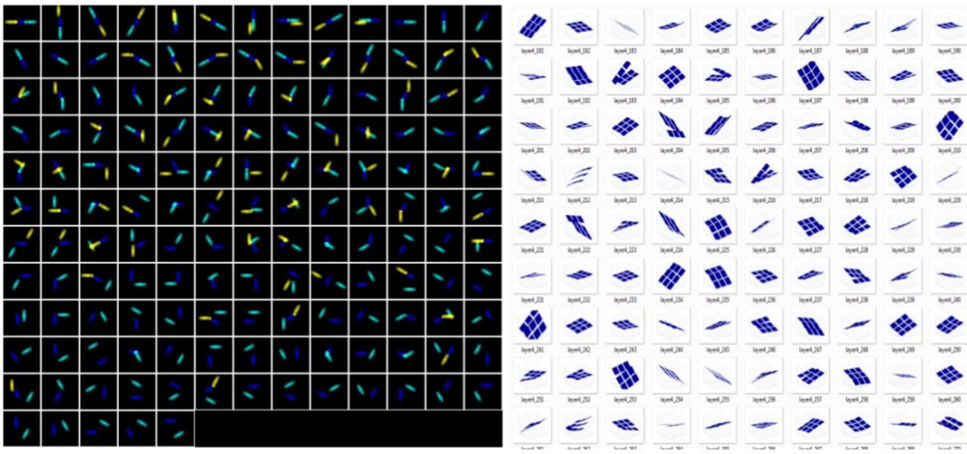


Figure 7: Examples of pairs of parts(2D,3D) at Layer 4, which are not statistically correlated.



a) 2D parts at Layer 4 b) 3D parts at Layer 4

Figure 8: Examples of 2D and 3D parts at Layer 4.

Γ,Ω	$H(\Gamma\|\Omega)$	$H(\Omega\|\Gamma)$
$\Gamma^{l_2}, \Omega^{l_2}$	0.17	0.26
$\Gamma^{l_3}, \Omega^{l_3}$	0.07	0.10
$\Gamma^{l_4}, \Omega^{l_4}$	0.48	0.26

Table 1: Conditional entropies of parts belonging to pairs of layers in the 2D and 3D hierarchies.

Γ	$H(\Gamma)$	Ω	$H(\Omega)$
Γ^{l_2}	3.75	Ω^{l_2}	3.6
Γ^{l_3}	4.81	Ω^{l_3}	4.77
Γ^{l_4}	2.47	Ω^{l_4}	2.03

Table 2: Entropies of parts at Layers 2-4 in the 2D and 3D hierarchies.

We analyzed the conditional entropies of 2D parts given 3D parts and vice-versa, introduced in Section 3.3 for parts belonging to pairs of layers in both hierarchies. This measure is useful at quantifying the amount of uncertainty remaining in predicting 2D parts given 3D parts and vice-versa, thus helping us to determine which information, either 2D or 3D, has a better predictive power in the inference process. The obtained results are included in Table 1.

The performed evaluation shows that a lower amount of uncertainty over the outcome for both types of correspondences is observed for pairs of parts belonging to Layer 3, while the highest amount of uncertainty is observed at Layer 4. Furthermore, the uncertainty of predicting 3D parts given 2D parts is lower than the uncertainty of predicting 2D parts given 3D parts at Layer 4, while the opposite conclusion can be drawn for Layers 2 and 3.

Furthermore, for measuring the dependency between the two hierarchies, we also compute entropies of 2D parts and entropies of view-based 3D parts on each layer of the hierarchy. The obtained values are included in Table 2.

By comparing the results presented in the above tables, we can the measure the gain obtained by using one hierarchy for inference on the other hierarchy and vice-versa. For all layers, it can be noticed that the information obtained from both hierarchies are dependant, as $H(\Gamma) > H(\Gamma|\Omega)$ and $H(\Omega) > H(\Omega|\Gamma)$.

Γ,Ω	$I(\Gamma, \Omega)$
$\Gamma^{l_2}, \Omega^{l_2}$	0.81
$\Gamma^{l_3}, \Omega^{l_3}$	2.68
$\Gamma^{l_4}, \Omega^{l_4}$	1.58

Table 3: Mutual information of parts belonging to pairs of layers in the 2D and 3D hierarchies.

Apart from measuring the properties of the conditional distributions of parts, we think is also relevant to assess the mutual information of 2D and 3D parts, which was introduced in Section 3.3. We present the computed values for Layers 2-4 in Table 3. As it can be noticed, the mutual information is highest at Layer 3, while Layer 2 has the smallest value. This result is explained by the number of found correspondences at Layer 3, where 30% of parts are useful in the matching process, while at Layer 4 only a fraction of 13% of parts are involved in forming correspondences.

5 Conclusion

In this report, we evaluated the statistical relationship between representations obtained from 2D and view-based 3D models, analysis which facilitates also the understanding of the geometric properties of the two compositional representations.

The evaluation of the learned correspondences between parts in the two hierarchies showed that meaningful correlations can be formed for a set of parts.

Next, we introduced a probabilistic fusion model for inferring missing observations, by exploiting the correlations between 2D and view-based 3D primitives.

The fused model enables an improved and refined representation of the world, by encompassing both 2D and 3D aspects of a scene. Moreover, it facilitates inference of one view from the other, in the case when only one hierarchical representation is available, or under challenging conditions, such as occlusions or noisy sensor data.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," vol. 32, no. 9, 2010, pp. 1627–1645.

[2] D. Teney and J. Piater, "Multiview feature distributions for object detection and continuous pose estimation," in *Computer Vision and Image Understanding Journal*, vol. 125, 2014, pp. 265–282.

[3] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3d geometric model," in *IEEE Conference on Computer Vision & Pattern Recognition*, jun 2010, pp. 1688–1695.

[4] F. Mahmoudi, F. Samadzadegan, and P. Reinartz, "A decision level fusion method for object recognition using multi-angular imagery," in *International Conference on Sensors and Models in Photogrammetry and Remote Sensing*, 2013, pp. 409–414.

[5] U. R. Aktas, M. Ozay, A. Leonardis, and J. Wyatt, "A graph theoretic approach for object shape representation in compositional hierarchies using a hybrid generative-descriptive model," in *Computer Vision âĂŞ ECCV 2014*, vol. 8693. Springer International Publishing, 2014, pp. 566–581.

[6] V. Kramarev, S. Zurek, J. L. Wyatt, and A. Leonardis, "Object Categorization from Range Images using a Hierarchical Compositional Representation," in *IEEE International Conference on Pattern Recognition*, 8 2014.

[7] Y. Wang, M. Gong, T. Wang, D. Cohen-Or, H. Zhang, and B. Chen, "Projective analysis for 3d shape segmentation," *ACM Trans. Graph.*, vol. 32, no. 6, 2013.

[8] R. W. Fleming and M. Singh, "Visual perception of 3d shape," in *ACM SIGGRAPH*, 2009, pp. 1–9.

[9] C. M. Cyr and B. B. Kimia, "3D object recognition using shape similarity-based aspect graph," in *ICCV*, 2001, pp. 254–261.

[10] C.-C. Cheng, C.-T. Li, and L.-G. Chen, “A novel 2D-to-3D conversion system using edge information,” in *IEEE Transactions on consumer electronics*, vol. 56, no. 3, 2010, pp. 1739–1745.

[11] S. Arca, R. Lanzarotti, and G. Lipori, “Face recognition based on 2D and 3D features,” in *International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, 2007, pp. 455–462.

[12] W. B. Soltana, D. Huang, M. Ardabilian, and L. Chen, “Comparison of 2D/3D features and their adaptive score level fusion for 3D face recognition,” in *3D data processing, visualization and transmission (3DPVT)*, 2010, pp. 455–462.

[13] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, “Sketch-based shape retrieval,” vol. 31, no. 4, 2012, pp. 1–10.

[14] J. Dai, Y. Hong, W. Hu, S.-C. Zhu, and Y. Nian Wu, “Unsupervised learning of dictionaries of hierarchical compositional models,” June 2014.

[15] Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu, “Learning active basis model for object detection and recognition,” in *International Journal of Computer Vision*, vol. 90, no. 2, 2010, pp. 198–235.