Markerless Motion Analysis in Diffusion Tensor Fields and Its Applications

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktor-Ingenieur (Dr.-Ing)

von

M.E. SANG MIN YOON
Geb, in Daejeon, KOREA

Referenten der Arbeit: Prof. Dr.-Ing. Dr. h. c. Dr. E. h. José Encarnação
Prof. Dr. techn. Dieter Fellner

Tag der Einreichung: 27. April 2010
Tag der mündlichen Prüfung: 28. Juni 2010

Darmstädter Dissertationen D17
Darmstadt 2010

Abstract

The analysis of deformable objects which have a high-degree of freedom has long been encouraged by numerous researchers because it can be applied to such diverse areas as medical engineering, video surveillance and monitoring, Human Computer Interaction, browsing of video databases, interactive gaming and other growing applications. Within the computerized environments, the systems are largely separated into marker based motion capture and markerless motion capture. In particular, markerless motion capture and analysis have also been heavily studied by numerous researchers using local features, color, shape, texture, and depth map from stereo vision, but it is still a challenging issue in the area of computer vision and computer graphics due to partial occlusion, clutter, dependency of camera viewpoints, high-dimensional state space and pose ambiguity within the target object.
In this thesis, we address the issue of the efficient markerless motion capture and representation methodology using skeletal features for the purpose of analysis and recognition of their motion patterns in video sequences. To localize the motion of the target object in a 2D image and 3D volume, we extract the skeletal features by analyzing its Normalized Gradient Vector Flow in the space of diffusion tensor fields since skeletal features are more robust and efficient than other features in recognizing and analyzing the deformable object. The skeletal features within the target object are automatically merged and split by measuring the dissimilarity of tensorial characteristics between neighbor pixels and voxels. The split skeletal features are used as features in human action recognition to understand human motion and target object detection and retrieval for Content based Image Retrieval.

This thesis provides the following contributions to the fields of computer vision and computer graphics:
(i) it introduces the notion of the features in the space of diffusion tensor fields and evaluates the successful analysis method of such features for motion interpre-

tation,
(ii) it presents a theory and an evaluation of the methods for automatic skeleton splitting and merging with respect to similarity measure between neighbor pixels in two dimension or voxels in three dimension and,
(iii) it presents and demonstrates our proposed principle methodologies for diverse applications such as human action recognition or sketch-based image retrieval.

With our system we can robustly handle several computer vision tasks to recognize and understand the motion of the target object without any prior information. In particular, the human action recognition using 3D reconstruction from multiple images and the skeleton splitting procedure is firstly proposed in this thesis and shown to be a useful and stable methodology. Furthermore, users can easily express their intention by sketching the characteristics of a target object and derive available related objects from a data base by using our proposed method.

Indexwords: Markerless Motion Capture, Diffusion Tensor Fields, Skeleton Extraction, Similarity measure, Human action recognition, Sketch-based image retrieval, Theses (academic).

Acknowledgments

This dissertation has been done during the time I spent in GRIS, Informatik, TU Darmstadt. First of all, I would like to thank my supervisor of my thesis, Prof. Dr.techn. Dieter Fellner and Prof. Dr.-Ing. José L. Encarnação. From all staff of GRIS, TU Darmstadt, Fraunhofer, IGD, and ZGDV, I am always motivated by the great atmosphere, open discussion, and friendship.
In particular, PD Dr. Arjan Kuijper and Prof. Stefan Roth, PhD., helped me to go to the right way in the study in computer graphics and computer vision. Beyond technical contribution, I appreciate my family for the support and understanding of abroad study.

Preface

Motion capture and its analysis was carried out by artists, medical doctors, and photographers by the end of 19th century. Through the development of computer technology, this topic moved to researchers in computer vision and computer graphics for video surveillance, computer animation and the gaming industry. This thesis introduces skeletal feature based motion analysis without any prior model to be used for various deformable objects. In numerous markerless motion capture and analysis methods, various features like shape, color, texture, skeleton and depth map using stereo vision are proposed to understand the complex motion of deformable objects.

This thesis provides fundamentals which are related to markerless motion capture and diffusion tensor fields, and its applications such as human action recognition and sketch-based image retrieval for comfortable Human Computer Interaction. Skeletal features which are very familiar to human visual perception are very efficient in understanding the characteristics of target objects using few data memory. Chapter 1 introduces the history, issues, and motivation involving motion capture and analysis, and is followed by an in-depth Chapter 2 which discusses the most influential previous work related to markerless motion capture and its applications in the field of computer vision and computer graphics. Our approach for skeleton extraction and splitting in a 2D image or 3D volume data is explained in Chapters 3 and 4. We have applied our proposed basic principles for human action recognition and sketch-based image retrieval in Chapter 5.

In each chapter, we show that our proposed methods are very efficient to analyze the deformable object without any prior information of target objects. We discuss the comparative experiments undertaken by us and conclude with comments for each topic.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Theory is...to demonstrate and explain the proportions of dexterity on the principles of proportion. Vitruvius. 1486:1.1.1

1.1 Overview

The initiation of motion analysis goes back a long way in the history of mankind. In ancient ages, people painted very detailed motion patterns of subjects when they were involved in some particular activities such as fishing, hunting or fighting with other tribes. In their numerous paintings, different disciplines are highlighted in various aspects of subjects according to their point of view or purpose. In particular, human motion analysis and understanding has gained great significance among numerous artists and medical doctors in order to know how our body parts are proportioned. This is because they need to model the human body perspectively to efficiently use the human labor [207].

The well known painting by Italian polymath, painter, architect and writer, Leonardo da Vinci (1452–1519), 'Vitruvian Man', is one example to know the correlations of ideal human proportions with geometry. In his drawing, he depicted a male figure in two superimposed positions with his arms and legs apart and simultaneously inscribed in a circle and square. Figure 1.1 shows his study of the proportions of the human body as described by Vitruvius.

The concerns for motion analysis have been continued and analyzed by a collaboration of scientists, medical doctors and artists in recent years. Their motion capture and analysis has been extended from human motion to various subjects

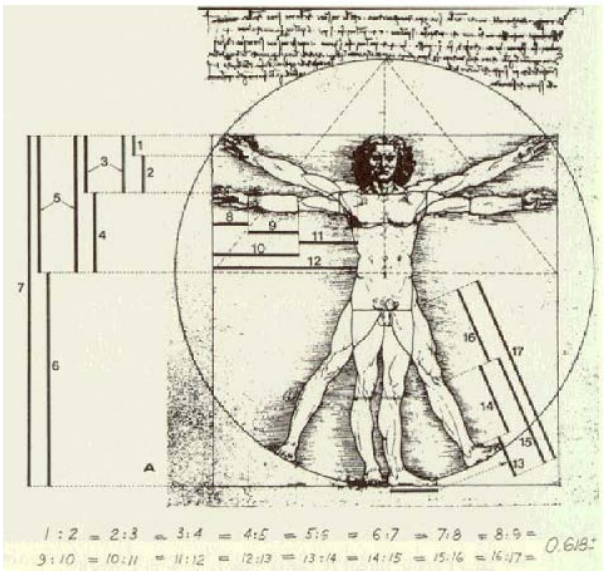


Figure 1.1: The Vitruvian Man drawn by Leonardo da Vinci in 1492.
This picture shows that Leonardo da Vinch tried to draw the variation of human motion according to its change.

like animals, medical cell and so on. Motion analysis using cameras was started by French astronomer Pierre Janssen (1824–1907). He recorded the transit of Venus across the sun using a multi-exposure camera which was invented by himself. By using his camera, the analysis of target objects was more reliable and more realistic in visualizing its characteristics. His work in turn greatly influenced chronophotographics and experiments of animals and humans using cameras. Eadweard J. Muybridge (1830–1904), an English photographer, also succeeded the works of Janssen by capturing the motion of human bodies or animals using multiple cameras and a device for projecting motion pictures that pre-dated the celluloid film strip that was still used. By 1878, Muybridge had successfully photographed a horse in fast motion using a series of twenty-four cameras. The cameras were arranged parallel to the track, with trip-wires attached to each camera shutter triggered by the horse's hooves. As shown in Figure 1.2, Muybridge tried to understand the movement of a horse by using multi-exposure cameras. Muybridge's motion studies, based on multiple images, were extended to walking downstairs, boxing, walking children and so on. They are often cited in the context of the beginning of biomechanics and they were certainly very influential in

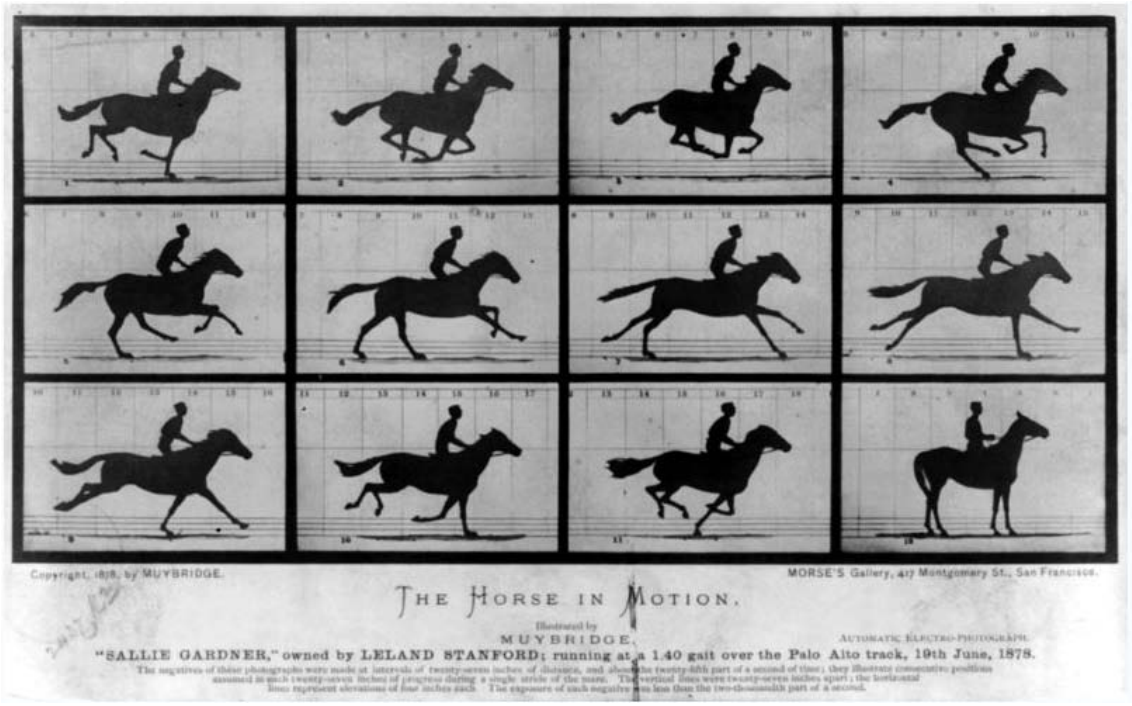
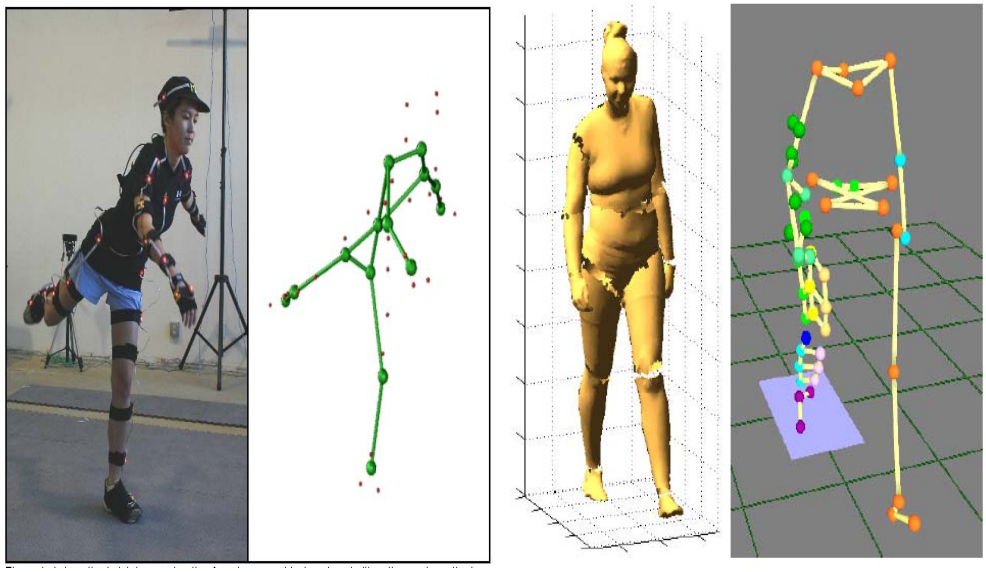


Figure 1.2: The horse motion from experiments of Muybridge to analyze the motion of horse.
Muybridge's experiments using multiple cameras are very efficient in analyzing deformable objects such as animals, humans walking downstairs or boxing.

the beginning of cinematography at the end of the 19th century.

Based on their innovative works for motion capture and analysis and also the rapid development of computers and digital cameras, this topic moved to the area of computer vision and computer graphics. The motion analysis and understanding of deformable objects in a computerized environment are today widely supported in such diverse areas as medicine [122, 159, 164] video surveillance and monitoring [158], computer animation [76], Human Computer Interaction in an augmented reality or virtual reality [55, 133, 217, 271] and sports scene analysis [102].

Motion capture or MoCap is used to describe the process of recording the movement of the target object and translating its motion onto a digital model. Hardware accelerated computers which contain an embedded Graphic Processing Unit (GPU) help parallel processing of the complex calculation of image sequences/3D



(a) Marker based motion capture and analysis(b) Markerless motion capture and analysis

Figure 1.3: Motion capture and analysis methods in a computerized environment.

volume data and high speed cameras have also contributed to real-time volume rendering and processing. Motion capture and analysis in a computerized environment can be separated into two categories namely marker based motion capture [40, 175, 200, 238, 257, 258] and markerless motion capture and analysis [32, 38, 39, 77, 93, 149, 161, 204, 206] as shown in Figure 1.3. These methods provide a technique of how to visualize the realistic 3D model of the users from an arbitrary viewpoint. Figure 1.3-(a) is one example which is based on marker based motion capture and its motion analysis. It first extracts the position of markers using calibrated Infra Red (IR) cameras and connects the markers in order to efficiently visualize and analyze its motion. In Figure 1.3-(a) [258], the skeletal virtual model based on the markers' position provides efficient information of motion change in real-time. Figure 1.3-(b) [204] shows one example of markerless motion capture and analysis from a human body reconstruction and its motion analysis using a skeletal model and fitting this model to a 3D reconstructed human model. In the next section, we will explain the details of marker based motion capture and markerless motion capture and compare the two approaches.

1.1.1 Motion Capture and analysis with Markers and without Markers

Conventional motion capture and analysis have two approaches. One is to attach many sensors to the important joints of a target object, the other is to analyze video sequences by using automatic feature detection, searching for correspondence between the features from multiple views and recovering the motion of the subject. The marker based motion capture and analysis using multiple sensors are already used for comprehensive applications in the analysis of user's performance, medical diagnosis, surveillance and 3D model retrieval. Surveys on existing methods for MoCap can be found in Moeslund et al. [165] and Gavrilla [81]. There are well-known marker based tracking methods existing, e.g. provided by Motion Analysis, Vicon or Simi [1]. A tracking failure sometimes occurs because of intrinsic problems by using surface markers or incorrect tracking of markers. The users also required special lab environments and lighting conditions, but people do not feel comfortable with the markers attached to the body. This often leads to unnatural motion patterns. The marker based motion analysis is also designed to track the motion of markers themselves and thus it must be assumed that the recorded motion of the markers is identical to the motion of the underlying human segments. Since human segments are not truly rigid, this assumption may typically cause an error in the motion analysis of sporting activities. For these reasons, markerless motion capture and analysis is an important field of research. It requires knowledge in biomechanics, computer vision, and computer graphics to overcome the drawbacks of marker based motion capture and analysis [127, 167, 221, 222, 255, 256].
In the area of markerless motion capture and analysis, there is a model based approach and non-prior information based approach. Typically, researchers working in markerless motion capture and analysis in the area of computer vision prefer simplified models, e.g., based on a stick, ellipsoidal, cylindrical or skeleton model [77, 160, 165] . Simplified models for motion analysis provide fast visualization of the target object, but sometimes miss the important information within that target object. In computer graphics, advanced object modeling and texture mapping techniques for an object's motions make it possible to visualize and render the target object without simple models, but the image processing and pose estimation techniques are more complex and require a lot of time. However, recently developed image based 3D reconstruction techniques [35, 60, 117, 132, 145, 157, 166, 216, 218, 245, 246] offer the exact shape representation of deformable object form multiple images in real-time.

1.2 Motivation

The interest in markerless motion capture and analysis is motivated by applications over a wide spectrum of topics such as segmenting the parts of the deformable objects [52], tracking the parts of the objects [31, 63, 66, 205], recovering the underlying 2D/3D body structures for the purpose of computer animation [36, 59, 83, 201, 202], pose reconstruction [3, 47, 87] and medical diagnostics [50]. Although many impressive results have been provided over the last few years, most motion capture and its analysis are very oriented towards their applications and the characteristics of target objects. Their motion capturing methods are based on simple and fast heuristics to increase the efficiency and robustness in a limited environment.

Computational theories in the field of computer vision address the area of motion analysis in a number of different approaches. One type of method attempts to estimate the spatial properties of the rigid body object from motion. For example, motion field [109], the projection of 3D motion of the points in the scene onto an image plane, allows for unique reconstruction and analysis of the 3D structure of rigid body objects. Besides rigid motion, there is a large class of non-rigid body objects including articulated motion, motion of elastic materials, fluids and gases in the world. The activities of the non-rigid body objects are classified into two categories: one is repeatable structure over time and the other is isolated simple motion without spatial and temporal repetition. The properties of the spatial-temporal information of non-rigid body objects are used as one of the important features in understanding the motion of target objects. However, its properties are also very dependent on the characteristics of the target object and do not directly adapt to other objects.

The challenge of using scientific visualization and modeling to understand the complex motion of the deformable object in the topological space of diffusion tensor fields is to decide which attributes of the features should be extracted, how these feature attributes will be visually abstracted into a comprehensible form which is very familiar to human visual perception, and where the extracted features should be applied. This thesis attempts to answer these questions by providing an efficient tool. Our proposed features will be familiar to human visual perception while reducing data memory and processing time. The previous understanding of non-rigid body has been considered in the areas of scalar or vector

fields rather than diffusion tensor fields, but the features in diffusion tensor fields provide much more information and characteristics to measure the similarity between neighboring pixels and voxels within appearance model.
The presence of partial occlusion and clutter in the image or 3D volume data has always disturbed the accurate analysis of a target object in computer vision. It is required to extract features which are very robust against noise and partial occlusion in recognizing the motion of deformable object.
3D volume reconstruction and rendering have been a key issue in computer graphics. From numerous image-based 3D reconstruction methods, Image based visual hull(IBVH) [218] and voxel coloring [245] are representative reconstruction methods in this area, but accuracy in these methods is very dependent on the number of cameras and the size of environment.

1.3 Contribution

In this thesis, we will introduce markerless motion capture and its analysis by extracting its skeletal features in the topological area of diffusion tensor fields. The main contribution of this thesis can be classified as follows:

1. We extract the skeletal features by analyzing the Normalized Gradient Vector Flow in the space of diffusion tensor fields. The eigen-features which come from diffusion tensor fields are used to measure the dissimilarity between neighboring pixels or voxels to automatically merge and split the skeleton (Chapter 3 and Chapter 4).

2. We propose a photo-realistic 3D model reconstruction from multiple images and camera calibration data by tracking the 3D location of the target object (Chapter 4). By tracking its 3D location, we can efficiently reconstruct the 3D model in a large environment. Our proposed 3D reconstruction methodology is also implemented in a GPGPU environment for a real-time rendering. This proposed methodology is implemented in CUDA, whose technology comes from NVIDIA to render the reconstructed 3D object in real-time.

3. Our tensorial feature based motion analysis is applied to various areas in industry such as medical volume visualization and rendering (Chapter 4), sport scene analysis, sketch-based target object detection (Chapter 5) and re-

trieval, human action recognition from 2D image or 3D volume data (Chapter 5).

4. Our proposed method is very effective in solving the problems of 2D image or 3D volume data in the presence of partial occlusion, clutter in the scene, high dimensional state space and pose ambiguity (Chapter 3 and Chapter 4).

Even though we have addressed the basic principles for markerless motion capture and analysis in the diffusion tensor fields, the possible applications of markerless motion capture in the diffusion tensor fields are growing rapidly and visualization of its data is an active area in computer vision and computer graphics.

1.4 Organization of the thesis

This dissertation presents a combination of methods that approach the task of markerless motion capture and its application, such as human action recognition and sketch-based object detection and retrieval in a varying and complementary way.
Chapter 2 presents a brief summary of related work on motion interpretation and its applications: the overview of skeleton extraction of deformable objects and its splitting that create an image or 3D volume data (Section 2.1), plus a comparison of shape based object recognition and skeleton based object recognition (Section 2.2) which shows why skeletal features are efficient in understanding the motion of deformable objects. Photo-realistic 3D reconstruction methodology from calibrated multiple images is also dealt with (Section 2.3). Various similarity measures to segment the unlabeled image or volume data are explained in Section 2.4, and we also briefly survey previous methods for human action recognition (Section 2.5) and content-based image retrieval (Section 2.6). Finally, we will explain the basic concept of diffusion tensor fields to be used for extracting the skeletal features and measuring the similarity within skeletons.
In addition, chapters 3 and 4 describe a theory which is used as a basis for analyzing the deformable objects by segmenting the 2D image or 3D volume data into several subregions which have similar characteristics. For automatic splitting of extracted skeletons within a deformable object, the eigen-features which come from diffusion tensor fields are employed (Section 3.2 and Section 4.2). In particular, in section 4.3, our proposed photo-realistic 3D reconstruction of target objects is introduced for real-time 3D volume data rendering. In both Chapters 3

and 4, we evaluate in a GPGPU environment and demonstrate the robustness and efficiency of our proposed principles. We conduct experiments by comparing our proposed 3D reconstruction method to previous voxel coloring methods. We also compare skeleton extraction methods to model based motion analysis methods using the Zernike Moment based approach.
Chapter 5 presents applications with respect to tensorial features and its similarity measure and describes the methodology for human action recognition in a 2D image or 3D volume data and sketch-based image retrieval. The eigenvalues and eigenvectors of segmented human body parts are used for recognizing basic human actions like walking, jogging and boxing. The Multiple-Kernel based Support Vector Machine is used for classifying human motion. The sketch-based target object detection and retrieval shows that our proposed methodology will be very effective for HCI. The hierarchical image clustering methodology will be good for the multi-label clustering and retrieval.
Finally, Chapter 6 concludes the work with a summary and a discussion of the presented approach and possibilities for extensions.

1.5 Summary

In this Chapter, we have introduced the brief history of motion analysis, motivation for the issue and organization of this thesis.
Interest in deformable objects goes back very far in human history. It has been motivated by curiosity, needs or methods at any given time. In particular, human motion analysis and understanding has gained significance among numerous researchers and artists with regard to knowing how our body parts functions and are proportioned.
Their efforts have been continued in the field of computer vision and computer graphics to be applied to medical engineering, video surveillance and monitoring, Human Computer Interaction and the browsing of video databases. In a computerized environment, marker based motion capture and markerless motion capture and analysis are widely used to record and analyze their motion interactively. With the development of computers and cameras, we can easily capture and reconstruct the non-rigid body due to a high-degree of freedom in real time. The research in markerless motion capture and its analysis from calibrated multiple images has become more focused in order to overcome the drawbacks of marker based motion capturing method.

We are motivated for motion capture to be used in diverse applications without depending on the properties of target objects. Previous motion analysis methods for non-rigid objects have been oriented in their applications and the properties of a target object by simplifying the models and requesting the heuristic parameters. In this dissertation, we will contribute the motion analysis methodology by extracting the skeleton from target objects and analyzing its motions since skeletal features are more robust in recognizing the motion of a target object than curve based approaches. So, skeletal feature based methods can reduce the processing time and memory in regarding motion and retrieve the related models in large databases.

Chapter 2

Related Work on 2D/3D Motion Interpretation

This thesis for markerless motion analysis of deformable object is closely related to several areas in computer vision and computer graphics including 2D/3D skeleton extraction from binarized foreground object and automatic skeleton splitting and segmentation, photo-realistic 3D reconstruction technique from multiple images.
In this chapter, we will survey the previous remarkable researches of motion capture and its analysis methodologies. The organization of this chapter is composed as follows.
In the first section, skeletal feature based motion capture and its analysis are explained in detail. The photo-realistic 3D reconstruction of target object which is required for 3D model's motion analysis will be surveyed in Section 2.2. The skeletal feature splitting which is one of areas of image/volume segmentation is motivated by numerous related works in computer vision and machine learning area and we briefly introduce the previous work in this area in Section 2.4. In Section 2.5, we will apply our proposed methodology to the human action recognition and the sketch-based object detection and retrieval to show that it is very independent on the characteristics of target objects.

2.1 Skeleton Extraction of Deformable Objects

The skeleton which is a set of centers of circles [25] within a deformable object is one of the important areas in image processing and computer vision. The compact

one-dimensional skeletal information which is very familiar to human visual perception has been widely used for shape analysis and object recognition, character recognition, image analysis and biomedical images.
Skeletons have several different mathematical definitions in the technical literature, and there are many different algorithms for computing them. The skeletonization approaches can be classified into four types: thinning algorithm [27], discrete domain algorithms based on the Voronoi diagram [177], algorithms based on the distance transform [23], and algorithms based on mathematical morphology [148]. From extracted skeleton, various approaches to reduce the noisy branches like pruning methods are introduced by measuring the significance assigned to skeletal points or smoothing the boundary before extracting the skeleton.
However, existing skeleton extraction algorithms are very weak because of their high computational complexity, noise sensitivity, centeredness inside the underlying complex shape, partial occlusion or artifacts in a singular region from the given shape. Most of previous methods are also based on vector fields which are generated from a given image by different physical properties. There are only few work to extract the features in the space of diffusion tensor fields from a topological point of view. In next section, we will compare the previous object recognition researches using skeleton and shape features.

2.1.1 Skeleton vs. Shape based Motion Analysis

The deformable object's appearance representation methodologies using local features like SIFT, color, texture, shape, depth map from stereo image, and skeleton can have a significant impact on the effectiveness of motion analysis strategy. A successful recognition technique has to be robust to visual transformations like articulation and deformation of parts, viewpoint variation, occlusion and so on. Thus deformable object representation has to effectively capture the variations in the shape of the target object due to these transformations. In previous shape representation and analysis methodologies, the objects are represented as curves, point sets or feature sets, and skeletons. Sebastian et al. [212] compare two techniques for matching shapes, one is based on matching their outline curves and the second based on matching their skeletons. They proved that the skeleton based shape representation and analysis method was better than curve based representation methods [19]. As the applications of motion analysis, human action recognition from image sequences/volume data and sketch based image detection and retrieval provide numerous literature. The experimental results from Sebastian et al. [212] encourage me to complicate the drawbacks of previous works and over-

come the state of the art.

2.2 3D Reconstruction from Multiple Images

As the preprocessing of 3D model segmentation, the real-time 3D reconstruction from multiple images is one of the important issues in this dissertation. The topic of 3D scene reconstruction of deformable objects based on multiple images has been investigated during the last twenty years and produced numerous results in the area of computer graphics and computer vision. Especially, real-time 3D reconstruction of target objects within a GPU environment [52] has become one of the hot issues nowadays. The 3D reconstruction research starts early on from a stereo vision based reconstruction technique proposed by [152]. Okutomi et al. [179] extend the conventional two-view stereo reconstruction into a multiple camera environment. Kang et al. [112] develop a method of multi-view stereo reconstruction from images to overcome the large occlusions. These methods are designed to reconstruct depth maps from particular viewpoints. Hence, they are usually not suitable for a full 3D scene reconstruction. Image based visual hull reconstruction(IBVH) [218] is a real-time 3D scene reconstruction technique from multiple view images. The visual hull is one of the most robust ways of extracting geometry from photographic input by using a space carving technique. They first found the silhouette contours of a foreground object in the image and then each image region outside of the silhouette represents a region of space where the object cannot be. The carved volume was a conservative approximation to the actual geometry of the object. The algorithm does not need to solve a corresponding problem. Instead, it simply calculates the convex hull of silhouettes in all view images. While the visual hull method works robustly when cameras surround the object, a concave object cannot be reconstructed using the silhouette alone. This problem was solved by a voxel coloring method presented by Seitz et al. [245].

2.3 2D/3D model Segmentation and Splitting

Automatic splitting of skeletal features from deformable objects can be understand as segmentation of unlabeled 2D/3D deformable object into functional parts. A part decomposition not only provides semantic information about the underlying object, but also can be used to guide several types of image processing algorithms,

including skeleton extraction, modeling, morphing, shape based retrieval, and texture mapping. All these applications benefit from segmentations that match human intuition [74, 153, 182].
Segmentation is a classical problem in processing of images, video, audio, surface, and other types of multimedia data. Accordingly, a large number of methods have been proposed for both computing and evaluating segmentations. Over the last decade, many segmentation and splitting algorithms have been proposed, including ones based on K-means [232], graph cuts [84, 116], hierarchical clustering [154], primitive fitting [11], random walks [125], core extraction [115], tubular multi-scale analysis [168], critical point analysis [137], spectral clustering [142], and so on. However, most of these methods have been evaluated only by visual inspection of results, and rarely applied to other applications.

2.4 2D/3D Human Action Retrieval

Some of the recent work which is done in the area of 2D/3D human action recognition can be largely separated into four categories: structural methods [33,62,80, 203,234,236], appearance methods using motion templates [26,94,172], statistical appearance-based methods [65, 67, 107, 192], and event-based motion interpretation methods [72, 195, 208, 278].
The structural methods use parameterized models describing geometric configurations and relative motions of parts in the motion patterns. The structural motion analysis and recognition provides the explicit locations of parts which lead to advantages for application of HCI and motion animation [243], but this approach requires a large number of free parameters that have to be estimated. An appearance-based method using template features needs a lower degree of freedom than structural approach, but it relies on either spatial alignment, or spatial-temporal registration of image sequences prior to reconstruction. A statistical approach is proposed to overcome the difficulty of finding corresponding features between models and structure in test images of structural and appearance based methods. Event based human action recognition methods are suffered from the lack of information about the motion. Most of the above studies are based on computing local space-time gradient or other intensity based features and thus may be unreliable in the cases of low quality video, motion discontinuities and motion aliasing.

2.5 Sketch based Image Retrieval

Our methodology has been motivated by the concept of Content based Image Retrieval (CBIR), Sketch based Image Retrieval (SBIR) and second order symmetric tensor fields. In this section, we survey the previous work which significantly contributed to the innovation within this area.

Research about retrieval of images has been studied in several fields such as computer vision, computational geometry, CAD/CAM, and molecular biology. Several CBIR allows a user to search images for content in media databases. They are mainly based on the derivation and analysis of meaningful features and the measurement of dissimilarities between visual properties by specific distance functions. Many commercial and non-commercial CBIR such as QBIC [75], VIRAGE [88], AMORE [169], MIT Photobook [190], VisualSEEK and WebSEKK [237], NeTra [147], and WBIIS [262] have matured during last years. Several CBIR is focused on feature construction in order to reduce the sensory gap due to the partial occlusion and accidental distortion by using color, texture, and shape abstraction. Especially, color histogram analysis for image indexing [101], the appliance of Gabor filters [110] and wavelet transforms [262] for local shape extraction received significant attention for robust image retrieval.

As one of part of CBIR, SBIR is started from 2D image retrieval [41, 105, 155] to 3D model retrieval [79, 95] and editing [113, 284]. SBIR was developed to overcome the limitations of previously well-known approaches such as keyword or example query based image retrieval, Funkhouser et al. [79] introduced a web-based search engine that has query images based on 2D or 3D sketches using a spherical harmonics shape descriptor. Hou et al. [95] also presented 3D model retrieval using a view-based 3D shape descriptor. The obvious advantage of this method is its easy to use. However, the boundary contours of each target object from different view directions or the information on incomplete shapes are needed to be prepared during a preprocessing phase. Fourier descriptors and Zernike moments are used to match the sketched query image on retrieved images from a database [96, 280].

Chapter 3

Skeleton Extraction from 2D Image In the Space of Diffusion Tensor Fields

3.1 Introduction

The analysis of non-rigid body objects with a high-degree of freedom has always been a challenging topic in the area of computer graphics and computer vision because traditional motion capturing methods gained major popularity not only by its use in medical diagnosis but also mainly in the film and gaming industry advancing the state of the art in modeling and motion reconstruction for computer graphic generated avatars. Aiming at an in-depth analysis of motions in order to better understand the normal and pathological movements, different methods have been introduced for motion analysis ranging from kinematic and kinetic modeling to complex capturing methods based on multiple video sensors recording the positions of markers attached to the target object. Here complex installations based on a large number of video cameras lead to the precise motion mapping of real actors' movements to their virtual counterparts.
Nevertheless, despite the potential of marker based motion capture and analysis, major hurdles for the broad acceptance have been the high cost for their installations, the controlling the environments and the complexity of pre-processing phases for its use. Hence, systems and techniques for low cost installments, easy to use and marker-free capturing methods for accurately measuring the object movements would significantly extend the applicability of motion capturing. At

the same time, marker-free motion capture and analysis imply a paradigm shift away from pure marker based capturing and the reconstruction of meaningful positions in the space. Whereas, marker based motion analysis provides the positions of markers attached to the target objects, marker-free motion capture and analysis rely on the deployment of an articulated model of the deformable objects. This articulated body models provide 'a priori' positions of body segments enabling a proper association of poses as well as the identification of individual body segments which allow for the extraction of the kinematic information.
Within the markerless motion capture and its analysis [32, 38, 39, 77, 93, 149, 161, 204, 206], various features such as local features [128] like Scale Invariant Feature Tracking (SIFT) [146] or optical flow [191, 197], shape [37, 82, 176, 214, 275], texture [259], skeleton [143, 188, 215, 225, 233, 283] and depth information [196] or combination of features are used to analyze the motion of deformable objects. From numerous features, skeleton is one of the most familiar human perceptual features with little data. In particular, skeletal features are better than other shape features in object recognition and retrieval [20, 213].
The skeleton which is a set of centers of circles is one of the important areas in image processing and computer vision. A precise definition of the skeleton or medial axis (MA) in the continuum was given by Blum [25], who postulated the well-known prairie fire analogy. It is a compact one dimensional representation of the complex and deformable objects and also describes an object's geometry and topology using little data. Meanwhile, it is used in many applications, including shape matching [225, 244], computer animation [141, 261] and object registration and visualization [14].
However, the existing skeleton extraction algorithms are still weak because of their high computational complexity, noise sensitivity, centeredness inside the underlying complex shape, partial occlusion or artifacts in a singular region of the given shape. Most previous methods are also based on vector fields which are generated from a given image by different physical properties. Few works have been investigated for extracting the features in the space of diffusion tensor fields from a topological point of view. In this chapter, we develop a skeleton extraction methodology by using a novel topological analysis of deformable target objects investigating the space of associated gradient vector flow fields.
In this chapter, we will present a skeletal feature extraction in the space of diffusion tensor fields. As we analyze the diffusion tensor fields of a Normalized Gradient Vector Flow within a given image, the proposed methodology has the following advantages comparing to other previous vector field based skeleton extraction techniques:

1. There is no need to determine the initial skeleton position in the image.
2. The computational complexity is very low because the skeletal features can represent the target object with one-dimensional data.
3. Our proposed methodology shows an improved skeleton extraction within a singular region of the shape over other existing methods. The normalized gradient vector flow reduces the effects of singularity problem to extract the skeletal features within a target object.
4. The algorithm is robust against the noise and partial occlusion, thus it is very robust to recognize and retrieve the images in database.

Figure 3.1 shows our skeleton extraction and splitting methodology which we will explain in this chapter. We first extract the normalized gradient vector flow from an appearance models which have the dense vector fields derived from images by minimizing an energy function in a variational framework in section 3.2. In Section 3.3, we extract the skeletal features by decomposing the normalized gradient vector flow in the space of diffusion tensor fields. The Section 3.4 provides the skeletal feature splitting methodology by measuring the dissimilarity between neighboring skeletal elements. The experimental results in Section 3.5 proves the robustness and efficiency of our proposed method by comparing to previous methods and ground truth of joint points.

3.2 Skeleton Extraction in Diffusion Tensor Fields

In this section, we explain the skeleton extraction methodology in the space of two dimensional second-order diffusion tensor fields. In the binarized image from the complex image, we first convert the binarized image to vertical and horizontal gradient vectors to extract the components of diffusion tensor fields. In section 3.2.1, we will introduce the basics of diffusion tensor fields, and then we will explain the detail of Normalized Gradient Vector Flow and its analysis in diffusion tensor fields to extract the skeleton within the target object in section 3.2.2.

3.2.1 Diffusion Tensor Fields

In medical applications, Diffusion Magnetic Resonance Imaging (MRI) is introduced as a powerful way to map white matter fibers in vivo images of bio-

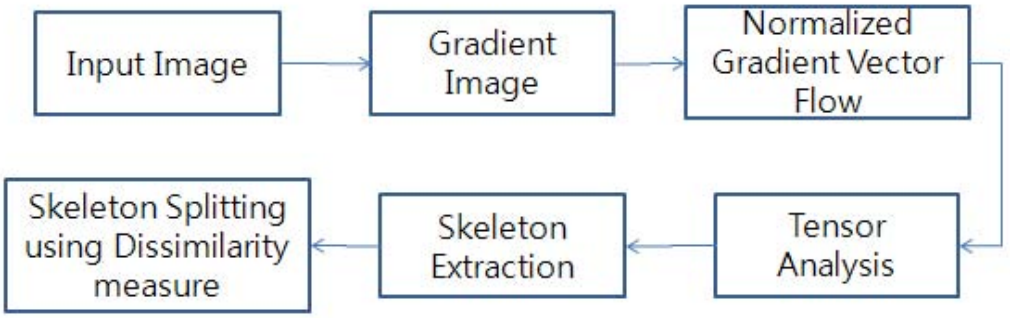


Figure 3.1: Total workflow of our proposed skeleton extraction and splitting methodology in diffusion tensor fields.

Our skeleton extraction and splitting procedure is composed of three categories: (1) Normalized gradient vector flow extraction from images of horizontal and vertical gradient image, (2) Skeleton extraction by calculating the degenerate points in the space of diffusion tensor fields, and (3) skeleton merging and splitting by measuring the similarity between neighbor skeletal pixels.

logical tissues weighted with local microstructural characteristics of water diffusion [99, 121, 211, 248]. Diffusion MRI methods are separated into two large categories: one is the Diffusion Weighted Imaging (DWI) [99, 248] and the other is Diffusion Tensor Imaging (DTI) [99, 211]. The Diffusion Tensor Imaging technique takes advantage of the microscopic diffusion of water molecules, which is less restricted along the direction aligned with the internal structure than along its traverse direction. The measured ratio of water diffusion will differ depending on the direction from which an observer is looking. In DT imaging, each pixel/voxel has one or more pairs of parameters: a ratio of diffusion and a preferred direction of diffusion for which parameter is valid. The properties of each pixel/voxel can be calculated by vector, each obtained with a different orientation of the diffusion sensitizing gradients. Historically, Micahel Moseley [140] reported that water diffusion in white matter is varied dependent on the orientation of tracts relative to the orientation of the diffusion gradient applied by image scanner and described in tensor. Basser et al. [16] showed the classical ellipsoid tensor formulism could be deployed to analyze diffusion MR data. Figure 3.2 shows the ellipsoidal representation of an axial brain section with a rectangular Region of Interest (ROI) and diffusion ellipsoidal representation in the Region of Interest.

Mathematically, the diffusion tensor fields relies on the interpretations of the ge-

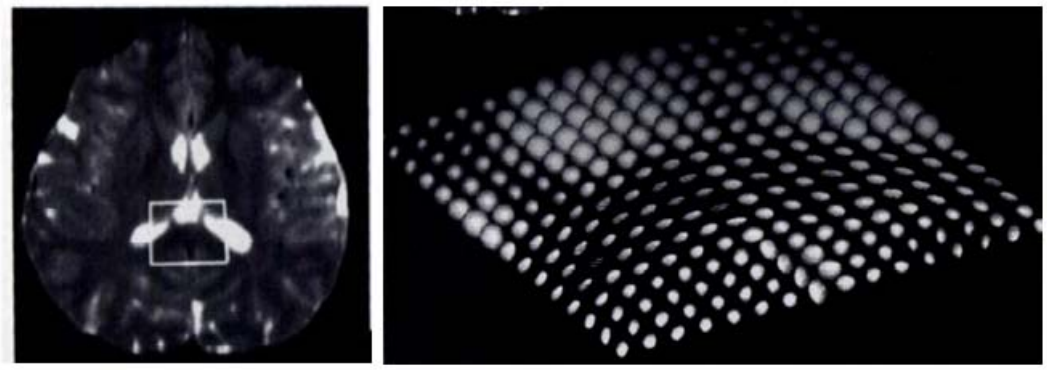


Figure 3.2: Diffusion ellipsoidal representation of ROI of human brain image which is developed by Basser et al. [16].

For each diffusion ellipsoidal model, the degree of diffusion anisotropy is embodied in its shape, the bulk or average diffusivity is related to its size, and the local fiber tract direction is given by the direction of its longest semi-major axis.

ometric quantities known as tensors. Tensors have a real, physical existence in a material or tissue so that they do not move when the coordinate system used to describe them is rotated.

3.2.2 Normalized Gradient Vector Flow of an image

Active snake model which was proposed by Kass et al. [114], has drawn a lot of attentions from researchers in computer vision and image processing. Due to its efficiency of converging to the desired features within a target object by simply defining an energy function, it has used to many applications, including edge detection [97], shape modeling [250], segmentation [130], and motion tracking [251].

Originally, the Gradient Vector Flow fields were proposed to solve the problems of initialization and poor convergence to the boundary within the concave objects yielding a traditional snake form [114]. The Gradient Vector Flow is a vector diffusion approach on Partial Differential Equations (PDEs). It converges towards the object boundary when it is very near to the boundary, but varies smoothly over homogeneous image regions extending to the image border. The main advantage of Gradient Vector Flow fields is to capture a snake from a long range and could force it into concave regions. Mathematically defined, the Gradient Vector Flow

is the vector field **v** that minimizes the following energy functional,

$$\varepsilon = \int\int \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + ||\nabla f||^2 ||\mathbf{v} - \nabla f||^2 dxdy, \tag{3.1}$$

where $\mathbf{v} = [u(x,y), v(x,y)]$, and the initial value of **v(x,y)** is determined by $\nabla f(x,y)$. $\nabla f(x,y)$ is the gradient image derived from a given image. μ is a regularization parameter to be set on the basis of noise present in image. Minimizing this energy will force $\mathbf{v}(x,y)$ nearly equal to the gradient of the edge map where $||\nabla f(x,y)||$ is large. This formula consists of two terms. The first term, the sum of the squares of the partial derivatives of the vector fields, makes the resulting vector flow vary smoothly. The second term stands for the difference between the vector flow and its initial status. Thus minimizing this energy will force nearly equal to the gradient of the edge map where $||\nabla f||$ is large. Nevertheless, the general Gradient Vector Flow method cannot efficiently extract the medial axis (MA) as a weak vector has very little impact on its neighbors that have much stronger magnitudes. Generally, there are several difficulties with this traditional Gradient Vector Flow functions. First, the initial assumption of the contour must be carefully chosen to be close to the true boundaries. This is because the snake moves partially in the direction of external force which is based on the image gradient. The second problem is that it is difficult for the snake to move into the boundary concavities if the external force is not large enough to push the snake into the boundary concavities. The third problem is how to select the initial snakes. A good guess of the initial snakes makes a great impact on the final segmentation.
To overcome these problems, normalized gradient vector flow (NGVF)is proposed which normalized the vectors before applying to the diffusion equation which is shown in equation3.1. A Normalized Gradient Vector Flow [277] can tremendously affect a strong vector, both in magnitude and in orientation by normalizing the vectors over the image domain during each diffusion iteration.
Figure 3.3 shows the Normalized Gradient Vector field from a given image. The traditional Gradient Vector Flow has difficulty preventing the vectors on the boundary from being significantly influenced by the nearby boundaries and thus causes a problem such that the ”snake” may move out of the boundary gap. The Normalized Gradient Vector Flow fields avoid this problem as shown in Figure 3.3-(b). In Figure 3.3-(b), we can see the detail of the Normalized Gradient Vector Flow in the vector around the boundary gap point.

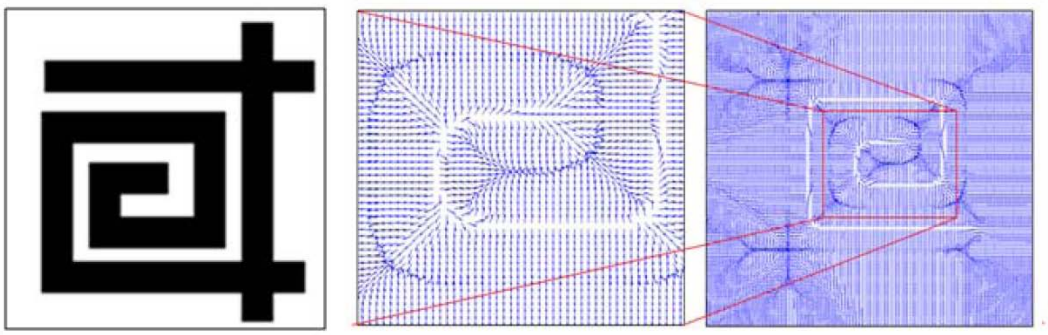


Figure 3.3: Extracting Normalized Gradient Vector Flow from input and gradient image.
(a)Input image (b) Normalized Gradient Vector Flow and its detail

3.2.3 Skeleton extraction in second order diffusion tensor field

In this section, we will explain the automatic skeleton extraction and refinement methodology by using a topological analysis of the Normalized Gradient Vector Flow fields.
Generally, the vector and tensor fields are multivariate and they involve more than one piece of information at every point of space. Representing data in tensor fields have more information than vector fields. The diffusion tensor field, which is defined as a topological representation from a two dimensional, symmetric, second-order tensor field is shown as:

$$T(\bar{x}) = \begin{pmatrix} T_{11}(x,y) & T_{12}(x,y) \\ T_{21}(x,y) & T_{22}(x,y) \end{pmatrix} \tag{3.2}$$

$T(\bar{x})$ is fully equivalent to two orthogonal eigenvectors

$$\bar{T}(\bar{x}) = \lambda_i(\bar{x})\bar{e}_i(\bar{x}), \tag{3.3}$$

where i=1,2. $\lambda_i(\bar{x})$ are the eigenvalues of $T(\bar{x})$ and $\bar{e}_i(\bar{x})$ define the unit eigenvectors [55].

According to [183], we can build a topological analysis of the diffusion tensor fields from the concept of degenerated points, which play an important role of critical points in vector fields. Streamlines in vector fields never cross each other except at critical points. However, the hyperstreamlines in the diffusion tensor fields meet each other only at the degenerated points. Thus, the degenerated

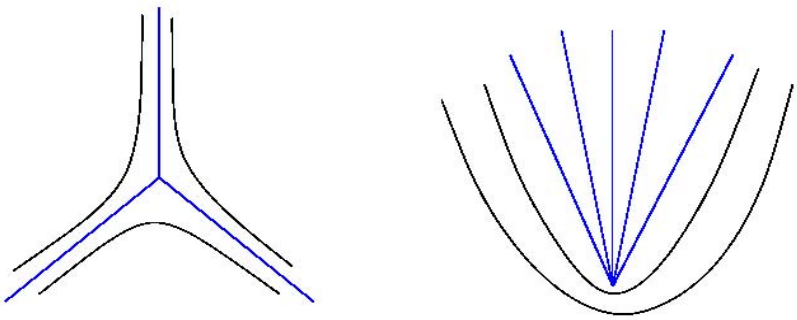


Figure 3.4: Degenerate point separation.
(a)Trisector if δ¿0 (b)Wedge if δ¡0
The black lines shows the silhouette of target object and blue lines present the degenerate points like trisector or wedge according to δ value.

points are the basic singularities underlying the topology of tensor fields. Mathematically, those points are defined as the two eigenvalues of $T(\bar{x})$ which are equal to each other. The degenerated points in the diffusion tensor fields are the basic constituents of critical points in vector fields. There are various types of critical points - such as nodes, foci, centers, and saddle points - that correspond to different local patterns of the neighboring streamlines. Delmarcelle [61] has proven that the local classification of line fields or degenerate points can be determined by constraints.
From a degenerated point, x_0, the partial derivatives are evaluated according to

$$\begin{array}{cc} a = \frac{1}{2}\frac{\partial(T_{11}-T_{22})}{\partial x} & b = \frac{1}{2}\frac{\partial(T_{11}-T_{22})}{\partial y} \\ c = \frac{\partial(T_{12})}{\partial x} & d = \frac{\partial(T_{12})}{\partial y} \end{array} \tag{3.4}$$

An important quantity for the characterization of degenerated points is

$$\delta = ad - bc. \tag{3.5}$$

So a simple point topologically should be classified into two types: trisector if $\delta < 0$, and wedge if $\delta > 0$. Within the target object, these points are assumed as

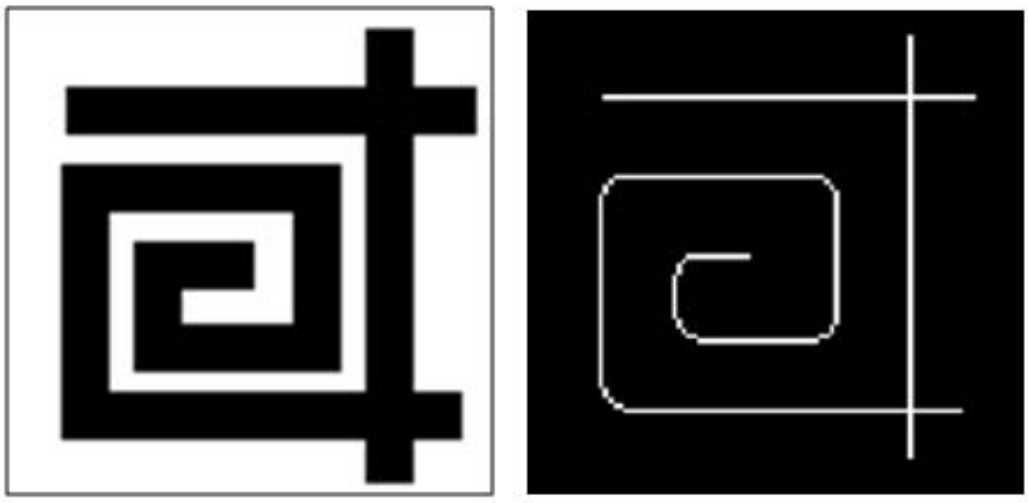


Figure 3.5: Skeleton extraction with our proposed approach
(a)Input image (b)Skeleton of target object

trisector [211].
Degenerate points which are categorized with trisector and wedge are different according to δ value. The local patterns of streamlines such as trisector and wedge are expressed with blue lines in Figure 3.4. Within the target objects which are separated in a binary image, we assume that the δ is always less than 0, and the degenerated point is trisector. Principally, the skeleton of deformable object is connection of degenerate points in the tensor topology. These trisector's degenerate points in tensor fields play the topological role of saddle points in vector fields. The deflect adjacent trajectories in any one of their three hyperbolic sector toward topologically distinct regions of the domain.

Thinning the skeletal features within the target object and connecting the features by continuous degenerated points can be very efficiently done by using the fact that a point within the object which has not at least one background point as an immediate neighbor cannot be removed, since this would create a hole. Therefore, the only potentially removable points are at the border of the object. Once a border point is removed, only its neighbors may become removable. Figure 3.5 is the extracted skeleton within the target object using our proposed approach.

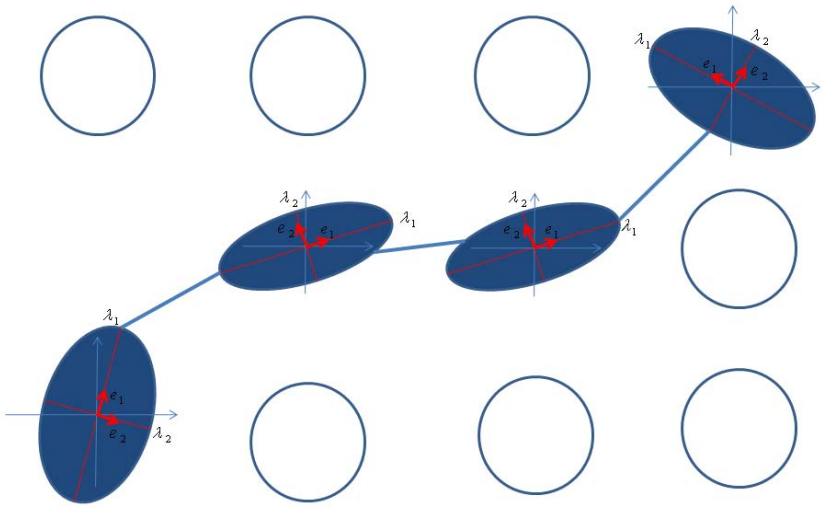


Figure 3.6: Ellipsoidal representation of extracted skeletal elements.
Its scale and rotation of ellipse is determined by extracted its eigenvalues and eigenvectors.

3.3 Automatic Skeleton Splitting using Diffusion Tensor Similarity Measure

After obtaining the skeleton of deformable objects, the skeleton is split into several branches by analyzing its tensorial characteristics. From extracted skeleton, we can separate the elements by using the following definition.

1. **branch point** is the pixel inside the skeleton that connects each branch.

2. **end point** is the pixel inside the skeleton with only one neighbor.

3. **joint point** is the pixel inside a branch that separate the neighbor.

End points can be interpreted as the polar points in the space of diffusion tensor fields and branch points can also be understood as the combination of various eigenvalues between neighboring pixels.

The skeletal elements in the target object can be decomposed as an ellipse model whose scale and rotation are determined by the extracted eigenvalues and eigenvectors.

Figure 3.6 shows how each skeletal element is represented by using its eigenvalues and eigenvectors. The characteristics of ellipse will be the features to separate the skeleton into several joints within a branch.

In a branch, we split the skeleton using the similarity measure between neighboring skeletal elements. For each pixel I_i which is recognized as the skeleton, we measure the dissimilarity between neighboring skeletal elements and measure the dissimilarity using tensorial characteristics. Given two tensors T_i and T_j between neighboring pixels, there are some dissimilarity measures that might be used to compare with neighboring pixels. The tensor can be represented by an ellipsoid, where the lengths of medical axis are proportional to the square roots of the tensor eigenvalues λ_1 and $\lambda_2(\lambda_1 ¿ \lambda_2)$ and their direction correspond to the respective normalized eigenvectors. With this properties, we can measure the dissimilarity between neighboring elements. The simplest one is the tensor dot product [7]:

$$d_1(T_i, T_j) = \sum_{i}^{2}\sum_{j}^{2} \lambda_i^1 \lambda_j^2 (e_i^1 \cdot e_j^2)^2 \tag{3.6}$$

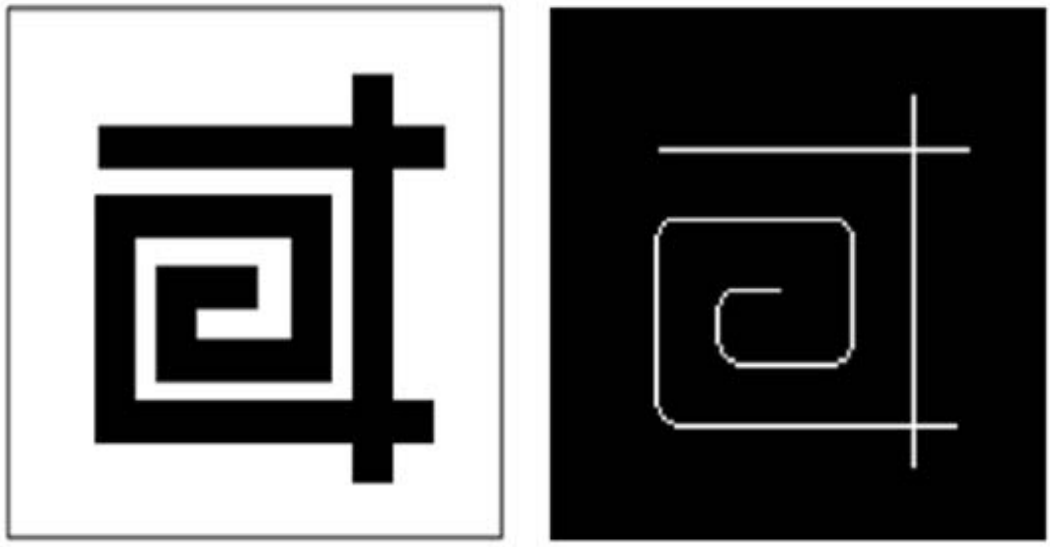
It uses not only the principal eigenvector direction, but the full tensor information. Another dissimilarity measure that uses the full tensor information is the Frobenius norm [7]:

$$d_2(T_i, T_j) = \sqrt{Trace((T_i - T_j)^2)} \tag{3.7}$$

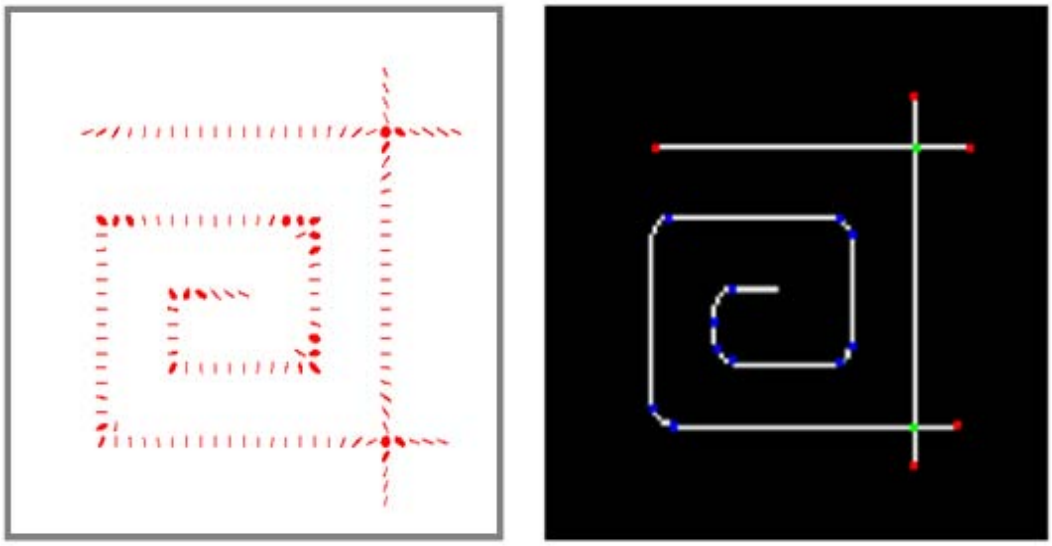
The dissimilarity measure between two elements is the multiplication of d_1 and d_2. Joint points are determined by comparing the similarity measure between neighbor points. Joint points are decided when the direction of Normalized Gradient Vector Flow changes and scale of main and sub eigenvalue is over the threshold. In the Figure 3.7, we visualize the extracted skeleton using ellipsoid representation method. The end points are painted by red, branch points by green, and joint points by blue which are determined by tensorial dissimilarity measure.

3.4 Experiments

We conducted some experiments in order to extract the skeleton and split the kinematics of deformable objects using our proposed methodology. Before we



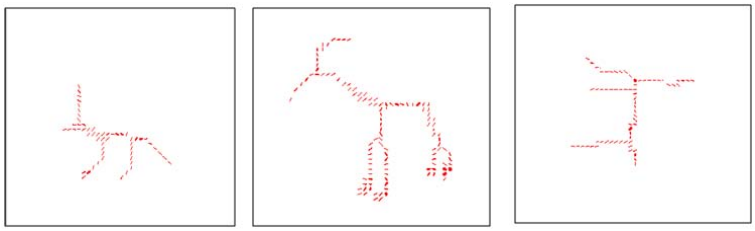
(a)Input image (b)Skeleton extraction



(c)Ellipsoid representation (d) Skeleton splitting

Figure 3.7: Skeleton extraction and splitting in tensor fields
In Figure 3.7 (d), end points are painted are by red, branch points are by green, and joint points by blue.

(a)Skeleton Extraction and Splitting from images



(b)Ellipsoid expression using tensorial elements

Figure 3.8: Extracted skeleton and splitting using our proposed method from images of public database using the characteristics of diffusion tensor fields.

generated the Normalized Gradient Vector Flow, several input images were converted to binary format due to performance and comparison issues with previous approaches. Afterwards, we calculated the eigenvectors and eigenvalues which were extracted from the diffusion tensor fields for identifying the degenerated points. Our experiments were composed of skeleton extraction within the well-known dataset, comparison with previous research to prove the advantages of our proposed method, and accuracy of splitting by measuring the Euclidean distance between our own method and ground truth.

3.4.1 Skeleton extraction and splitting in various image set

We first demonstrated the performance of our proposed skeleton extraction and splitting method in the public image dataset[1] which include various objects. Figure 3.8-(a) shows the extracted skeleton of the binarized input images. Images of Figure 3.8-(b) are the tensorial characteristics of the skeletal features from the input image. The eigenvalues and eigenvectors provide the scale and rotation of the ellipse in the image. Based on these characteristics, we can successfully separate the skeletal features into several joints.

[1] http://www.lems.brown.edu/vision/software

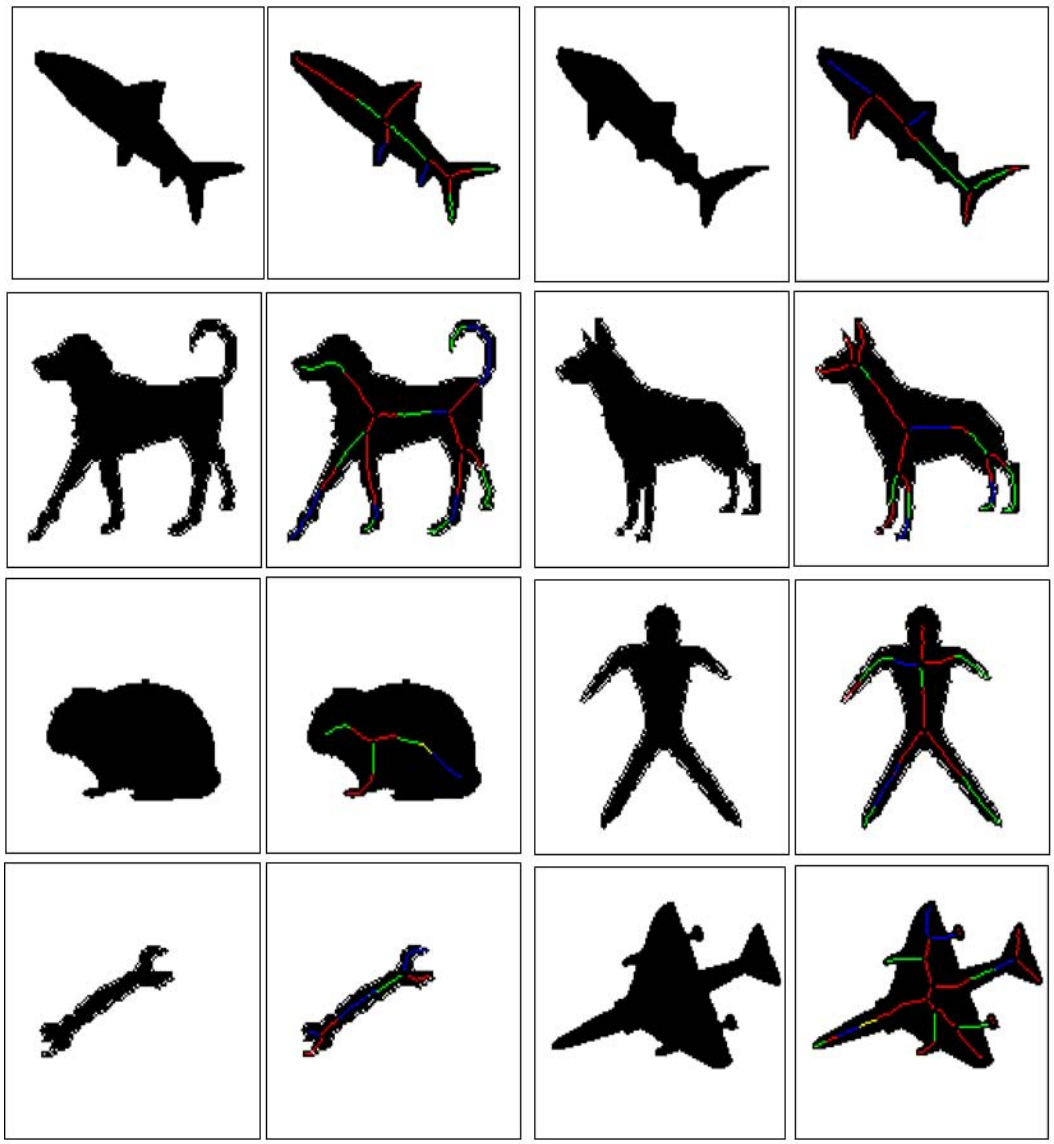


Figure 3.9: Skeleton extraction and its splitting from various deformable objects.

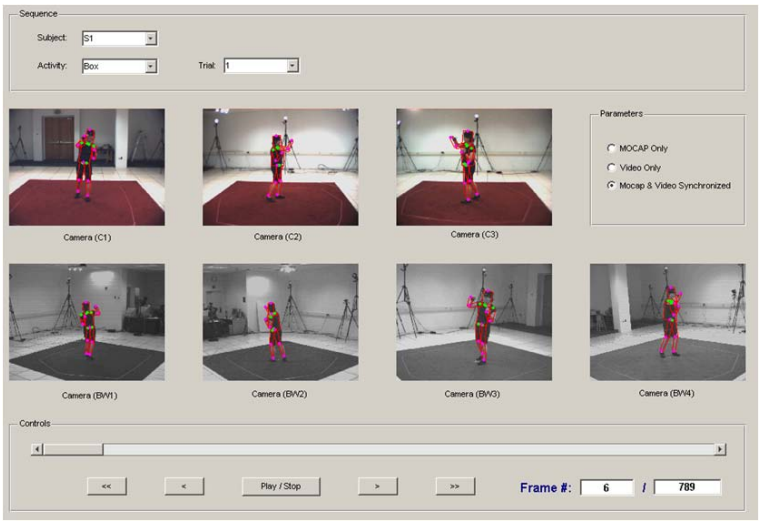


Figure 3.10: Configuration of HumanEva dataset. They provide the MoCap data based on the markers of each joints of human body part from 7 different viewpoint.

The experimental results of various deformable objects were shown in Figure 3.9. The segmented areas in each branch were painted by using different color. Our proposed methodology using Normalized Gradient Vector Flow and eigen-features from diffusion tensor fields did not require any prior information and restrictions to segment the target object, so our approach could be applied to various objects like animals, tools, and human body which have high-degree of freedom.

We also extracted the skeleton and split the skeleton from the images of HumanEva Dataset[2]. HumanEva-I data which are shown in Figure 3.10 contain 7 calibrated video sequences (3 color and 4 gray) which are synchronized with 3D body poses obtained from a motion capture method. The HumanEva-I dataset also contain 4 subjects performing 6 common actions such as walking, jogging, boxing, etc. Figure 3.11 shows the split skeleton of the image from HumanEva dataset. The split areas within each branch is painted by different color. To binarize the image, we compute the background subtraction based on the statistic background information which they provided.

[2]http://vision.cs.brown.edu/humaneva/

3.4.2 Comparison with previous methodology

We compared our proposed skeleton extraction with previous techniques such as skeleton pruning using contour partition [42], and morphological approach [150] in Figure 3.12. Our proposed skeleton extraction method could efficiently represent the characteristics of target object, but very robust in noise effect. We calculated the accuracy of the splitting of the skeleton.

3.4.3 Accuracy test between ground truth and our approach

We lastly conducted the experiments to measure the Euclidean distance between the ground truth and our proposed skeleton extraction and splitting points. To measure the distance, we first converted the color image into binarized image by using background subtraction. We ignored the branch whose size is less than 20 pixels.
Figure 3.13-(b) is the ground truth of human body parts. From numerous split points, we compare the 11 points of human body parts.

Table 3.1 shows the Euclidean distance between the ground truth and extracted split points. The feature points 4 and 5 which are shown in Figure 3.13 have large Euclidean distance than other feature points because despite other features are within the target object and our skeletal features are very close to medial axis of target object, feature points in 4 and 5 are the end of the target object. The standard deviations of Euclidean distance between ground truth and extracted joint points tell us that our proposed method is very robust against various non-repeatable human actions.

3.5 Summary

From numerous representation methodologies to efficiently represent the deformable objects, skeleton is very familiar to human visual perception using one-dimensional data.
In this chapter, we have shown a novel method how we extracted and split the skeleton in the target objects by using the robust, accurate and computationally efficient technique in the topological space of diffusion tensor fields. We extracted the skeleton by using Normalized Gradient Vector Flow. The essential idea to extract the skeletal features was to connect the degenerated points using the eigenvectors and eigenvalues which come from the properties of diffusion tensor fields.

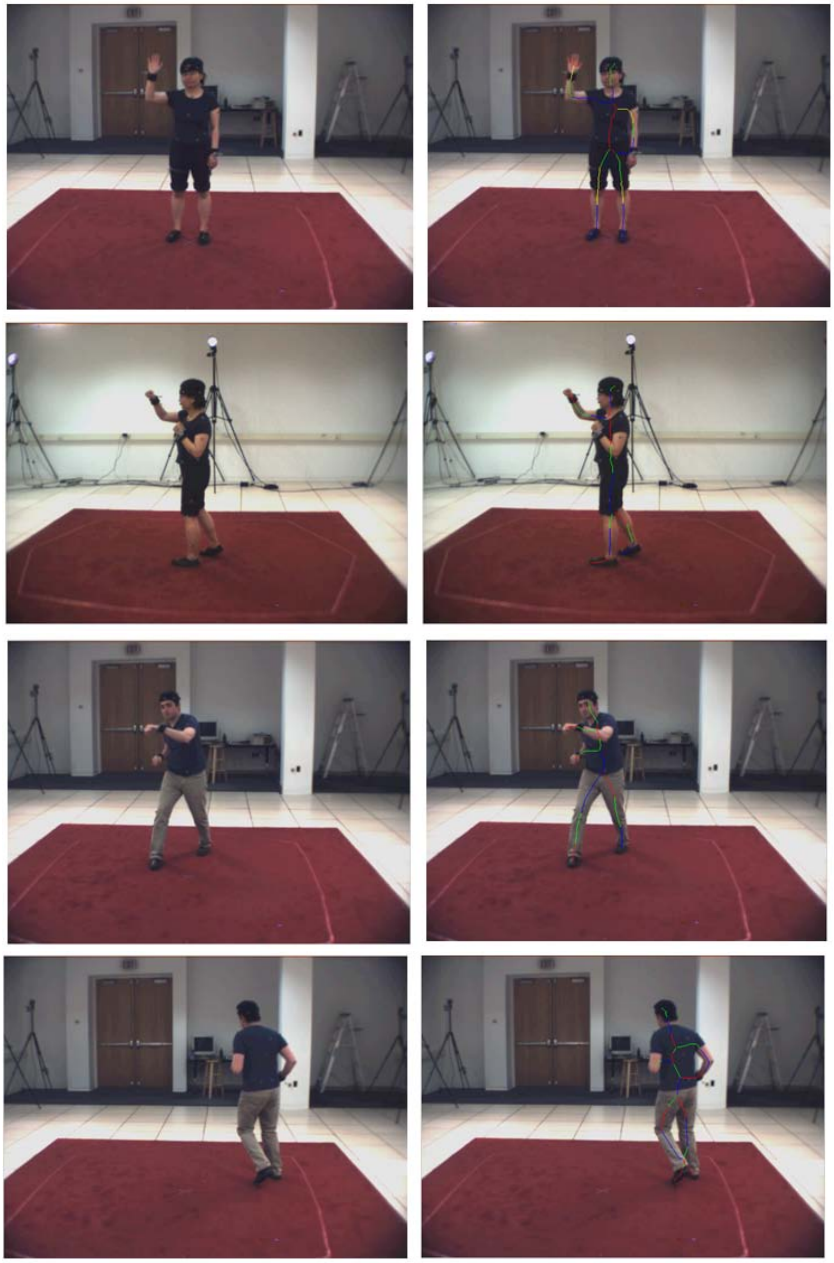
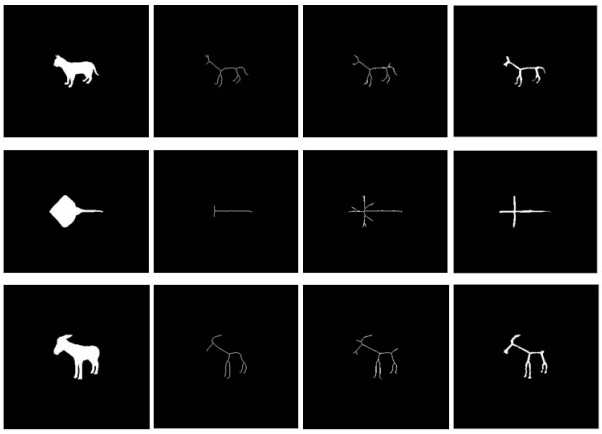
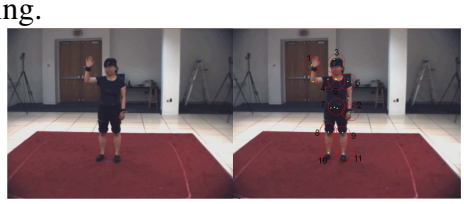


Figure 3.11: Skeleton extraction and splitting of human body parts. We extract the silhouette using background subtraction from the original images of HumanEva dataset.

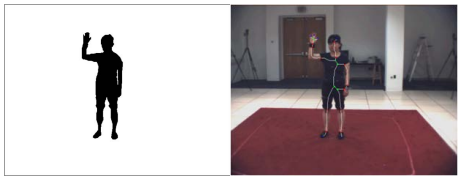


(a)Input image (b)Morphological thinning (c)Skeleton pruning (d) Our approach

Figure 3.12: Comparison of our proposed skeleton extraction methodology with previous skeleton extraction such as morphological thinning and skeleton pruning with contour partitioning.



(a)Input image (b)The ground truth of the split joints are painted by yellow.



(c) Segmentation of target object based on background subtraction. (d)Split skeleton with our own approach. The split joints are pointed by different color.

Figure 3.13: Error measurement between ground truth and measured split points.

Table 3.1: Euclidean distance between ground truth and measured points.

	Euclidean distance (pixel)	Standard Deviation (pixel)
1	5.8182	1.9400
2	10.8860	4.1246
3	16.5479	2.6267
4	21.5287	4.1094
5	26.2333	4.0535
6	5.1329	3.4150
7	3.9196	1.7967
8	2.0991	0.6100
9	0.5245	0.6204
10	4.9123	0.3619
11	3.6964	0.5030

We decomposed the extracted skeletal pixel into ellipsoidal model whose size and direction were determined by its eigen-features. From the extracted skeletal features within the appearance model, the skeleton was separated into several parts by extracting branch point, end points and joint points. In particular, the joint points were extracted by measuring the dissimilarity between neighboring skeletal pixels.

We have illustrated our approach on a variety of 2D deformable objects like animals and human body parts by comparing it to previous techniques. Lastly, we conducted the experiments to show the accuracy of our extracted end, branch, and joint points by comparing the ground truths of target object to our extracted features.
We will focus in particular on a 3D skeleton extraction from volume data in the view of motion analysis of deformable objects in next chapter 4. An additional similarity measure would help us to analyze the objects' motion and retrieve related motion patterns from within a database. The skeletal feature extraction and splitting in the 3D volume data is the dimensional extension of Normalized Gradient Vector Flow fields, three-dimensional second order diffusion tensor fields, and photo-realistic 3D reconstruction procedures from this chapter.

Chapter 4

Skeleton Extraction of 3D Reconstructed Deformable Objects from Multiple Images

4.1 Introduction

The acquisition and visualization of three-dimensional real world objects from a set of images are an important topic in computer graphics as well as computer vision. The main aim of 3D visualization which is used in the close range photogrammetry is the photo-realistic 3D reconstruction of real objects. Most techniques that have been developed during last two decades have focused on how to render the 3D shape of deformable objects in an arbitrary viewpoint using meshes or point-sets. Considering the techniques that recover the shape of target objects, classical approaches which are commonly used for motion analysis rely on the 3D scanners [8, 9, 220, 263, 265]: these sensors are quite expensive but simple to use and various softwares are available for modeling the 3D measurements. They work according to different techniques providing for millions of points, often with related color information. Other techniques try to recover the shape of object with image-based approaches [52, 112, 152, 179, 218, 245]. They can use camera calibration from multiple cameras, silhouette extraction and 3D reconstruction technique.
Even though there is a progress in motion capture and analysis using skeletal features within the 2D image [143, 188, 215, 225, 233, 283], there are still constraints in partial occlusion, clutter and dependency of camera viewpoint. Many research

activities in this area have focused on the problems of tracking the moving objects through an image sequence acquired with multiple cameras and often using predefined 3D models [4, 29, 30, 70, 98, 111, 163, 187]. However, little attention has been directed to the determination of 3D information of the target objects directly from image sequences. 3D structural analysis and visualization of deformable objects can be used to various areas because it can be used in many applications such as pose estimation [253], 3D model matching and retrieval [34] and computer animation [199]. However, 3D structural analysis from the 3D model is still lacking in terms of operating a large set of data. One of representative approaches for 3D shape visualization is a skeletal visualization [28, 48, 51, 92, 194, 247, 249, 282], which is familiar to human visual perception by reducing its volume data to a graph.
In this chapter, we develop a markerless skeleton extraction and its splitting methodology from 3D reconstructed object's structural analysis in the space of associated Gradient Vector Flow fields. As we analyze the diffusion tensor fields of a normalized gradient vector flow, the proposed methodology has the following advantages comparing to previous vector fields based on skeleton extraction techniques and motion analysis methods:

1. There is no need to determine the priori information from 3D model. We do not require the prior model of the target objects to be applied to various non-rigid body objects which have high-degree of freedom such as humans and animals.

2. Our proposed methodology shows an improved skeleton extraction within a singular region of the target object over other existing methods. The eigenfeatures which come from three-dimensional second order diffusion tensor fields provide the efficient features for thinning the skeleton features within the deformable volume data.

3. Our proposed algorithm is robust against noise when the shape of 3D reconstructed model is very complex. So our proposed features are very efficient to object recognition and retrieval.

4. We can analyze the motion of deformable objects by segmenting the skeleton. Skeletal feature splitting procedure is computed by measuring the dissimilarity between neighboring voxels. The tensorial features are used for skeleton splitting and merging.

Figure 4.1 shows our methodology which we will explain in detail further on. Our system is largely separated with 3D reconstruction from a set of input images(blue box), and skeleton extraction and splitting in a space of diffusion tensor fields(red box). In the area of 3D reconstruction from multiple images, we propose a new 3D reconstruction methodology by tracking the target objects. In the section of 3D skeleton extraction and segmentation, we decompose the 3D volume data into three dimensional ellipsoidal models whose properties come from diffusion tensor fields. We then extract and segment the skeletal features by measuring the similarity.

4.2 3D Reconstruction from multiple images

Real-time photo-realistic 3D reconstruction of target object from multiple cameras can be largely separated into three categories: camera calibration, silhouette extraction using background subtraction, and 3D reconstruction technique by tracking the 3D position of the target objects. In this section, we will follow the procedure of 3D reconstruction procedure in the hardware accelerated GPU environment for real-time rendering.

4.2.1 Multiple Camera calibration

In this section, we will discuss how to project the 3D rays onto image planes and how to lift the resulting 2D points back to 3D scene. Camera calibration for 3D reconstruction of target object is a necessary step in order to extract metric information from 2D images. These works have been done [73, 91, 254], starting in photogrammetry community, and recently in computer vision. Through camera calibration procedure, we find the camera's intrinsic and extrinsic parameters that accurately predict the pixel coordinates of a point in 3D from the point's world coordinate system. As input, we have a large number of points with 3D coordinates in the world and the pixel coordinate.
Our camera calibration procedure starts from the assumption that the camera model is pinhole camera. The point C denotes the center of projection of the camera and P denotes an inverse projection matrix that transforms homogenous image coordinates $x = [uv1]^T$ to rays in 3D where u and v are the pixel position in the image. The pair $[P|C]$ specifies the location of the camera in the 3D space. The following equation specifies how to compute a ray in 3D:

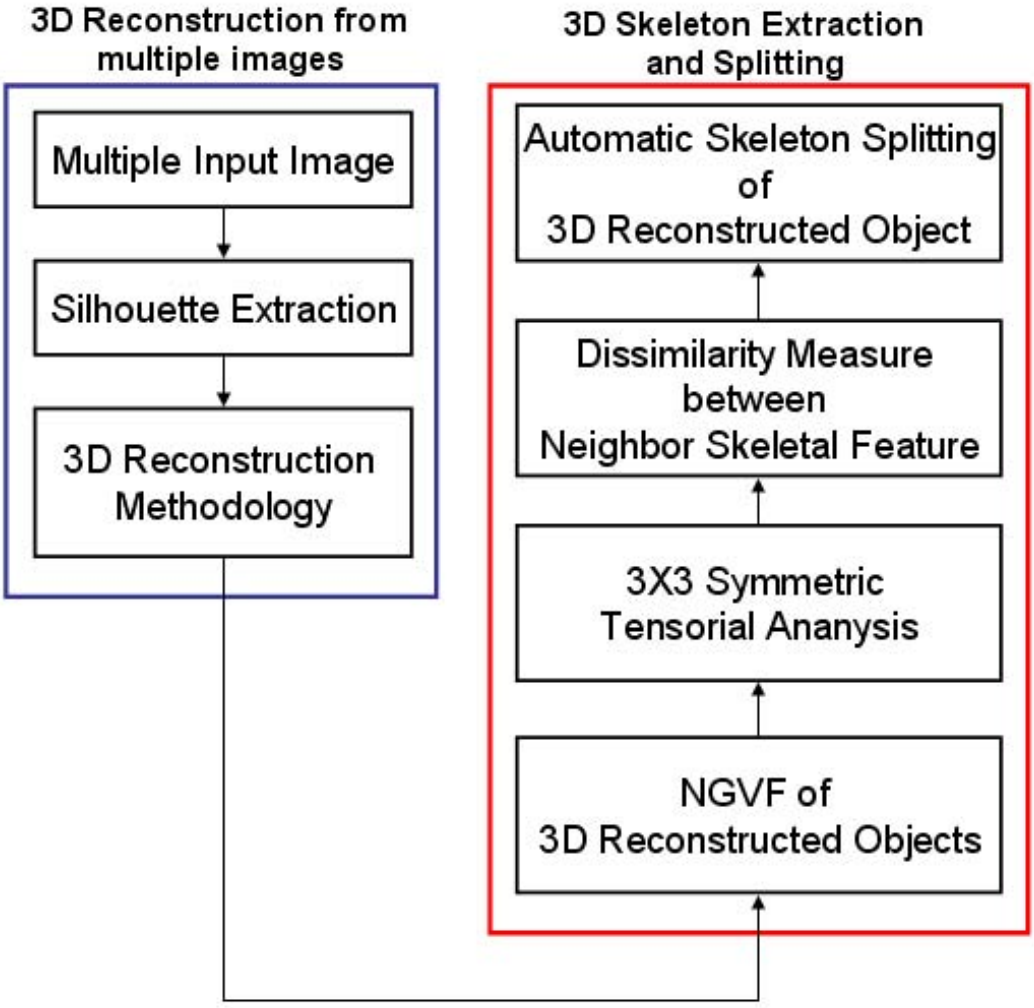
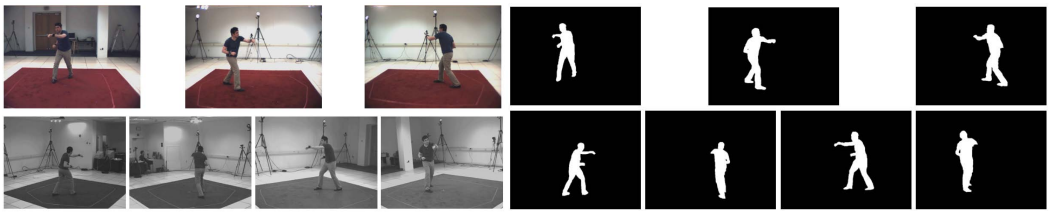


Figure 4.1: Total flowchart to extract the skeleton and splitting of 3D reconstructed object from multiple images. Our system is largely separated with 3D reconstruction form multiple images(blue box) and skeletal elements extraction and splitting in the space of tensor fields(red box).



(a)Input image (b) Silhouette extraction using background subtraction

Figure 4.2: kernel density estimation based background subtraction

$$X(t) = C + tPx \tag{4.1}$$

The following equation is used to obtain the point a, the projection of the 3D point A onto the camera's image plane specified by $[P|C]$:

$$a = P^{-1}(A - C) \tag{4.2}$$

We could compute the ray projection by projecting two of its points onto the reference image and then determining the line through the points.

4.2.2 Target object segmentation using kernel density estimation based background subtraction

There exist many approaches to extract and segment the target objects with the lowest possible false alarm rate from static cameras. Background subtraction [54,71,90,123,144,180,219,239,240,268] is a method typically used to detect the deformable objects in the scene by comparing each new frame to a model of the scene background. We use a non-parametric technique for background modeling and foreground extraction. Our approach is based on kernel density estimation of the probability density function of the intensity of each pixel within each image. Kernel density estimation based background modeling aims on capturing and storing recent information about the image sequence, continuously updating this information in order to capture fast changes in the scene background [90].

The intensity distribution of a pixel can change quickly. So we can estimate the density function of this distribution at any moment of time given only very recent history information if we want to obtain a sensitive detection. Using the

recent pixel information, the probability density function of each pixel will have intensity value $I(x,y)$ at time t and can be non-parametric estimated using the kernel, K as

$$pdf(I_t) = \frac{1}{N}\sum_{i=1}^{N} K(I_t - I_i) \tag{4.3}$$

where N is the recent pixel information for comparing the current image's pixel information. We choose our kernel estimation function to be a Gaussian kernel for color image. Figure 4.2 is the example for detecting the foreground object detection from static cameras by using kernel density estimation based background subtraction. This example images are captures from the HumanEva database.

4.2.3 Tracking based 3D reconstruction methodology

We follow a photo-realistic 3D reconstruction methodology from multiple images which have camera calibration data and silhouette extraction of the target objects. We use the background subtraction from the static cameras by applying kernel density estimation based background subtraction methodology. From numerous previous photo-realistic 3D reconstruction techniques from multiple images, Image Based Visual Hull [218] and voxel coloring [245] promised an efficient 3D reconstruction technique in real-time.

Nevertheless, the Image Based Visual Hull algorithm computes a 3D coarse shape reconstruction of a target object from its 2D projections from a few number of images because it is very dependent on the number of images used, on the position of each viewpoint considered, on the camera's calibration quality and on the complexity of object's shape. Voxel coloring which reconstructs the radiance or color at the surface points by projecting every voxel to each image plane. It takes much time to voxelize the whole 3D scene. Our proposed 3D reconstruction methodology continuously tracks the 3D boundary of target object and carves the voxel by checking the color consistency within the boundary of tracked target object. Our proposed methodology is very efficient and accurate compared to other previous methods in the case of 3D reconstruction from few cameras in a large environment.

We consider a 3D scene observed by n calibrated static cameras and we focus on the state of one voxel in the position V chosen among the positions of the 3D lattice used to discretize the 3D scene. We here model how knowledge about the

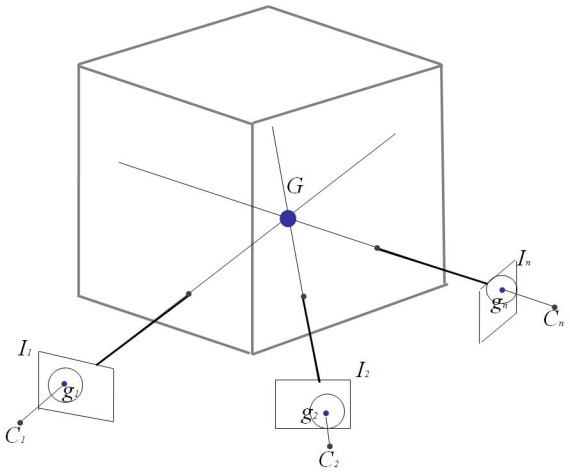
occupancy state of voxel V influences image formation, assuming a static appearance model which is extracted from kernel density estimation based background subtraction.

Figure 4.3 shows the concept of configuration of 3D lattice by tracking the target object and its inverse projection in the 3D scene. We continuously track the center of gravity $g_1, g_2, ..., g_n$ of the appearance model in each image and calculate the G points in the 3D scene which is earned by intersection of n 3D lays. We extract the 3D lattice by combining the silhouette images of target object to be reconstructed with camera calibration information to set the visual rays in the 3D space for all silhouette points, which define a generalized cone within lays the same object. The 3D lattice in a whole scene is determined by its intersection of these cones. Within a 3D lattice, we use the photo-consistency measure to determine if a certain voxel V belongs or not to the object being reconstructed.

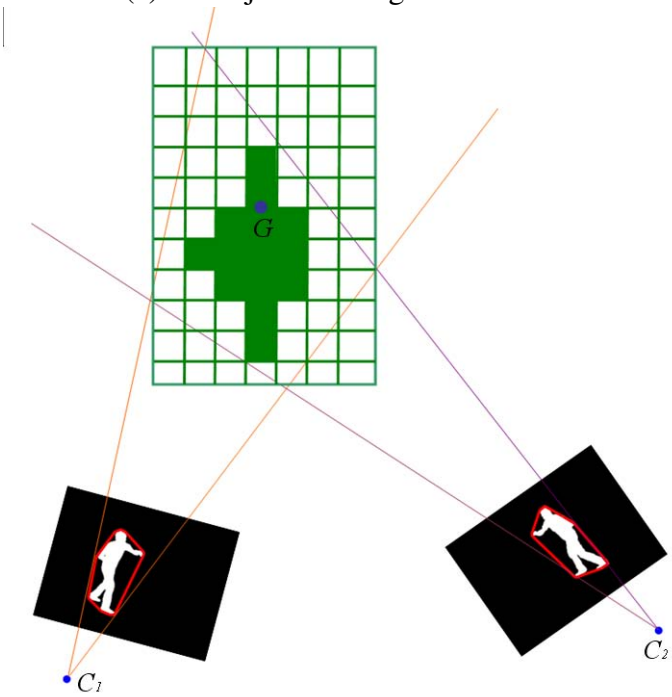
Figure 4.4 visualizes the procedure for 3D target object reconstruction from multiple images by tracking the boundary. Figure 4.4-(a) is the input images from various viewpoints, Figure 4.4-(b) shows the 3D lattice which is constructed from intersection of multiple convex cone by tracking the 3D center of gravity of target object. We visualize the 3D reconstructed object within a 3D lattice. Within the 3D lattice, we carve the voxels by checking the color consistency between multiple images. Figure Figure 4.4-(b) also shows the reconstructed object in various viewpoint of the 3D reconstructed object.

4.3 3D Skeleton extraction in 3D diffusion tensor fields

Several techniques for 3D mesh model based skeleton extraction and analysis have been introduced in the area of computer graphics in order to separate a 3D volume into several joints which have similar characteristics. Skeletal splitting technique, one of fields of 3D model segmentation methodologies, can be divided into several classes according to the different classification schemes, Shamir [224] surveyed previous researches on the 3D segmentation of 3D volume data and classified the problems into two types of segmentation classes: as surface-type [12, 269] and part-type [134,252,281] segmentation. Surface-type based segmentation methods are based on the decomposition of geometric primitives such as planes, cylindrical patches and spherical parts. Part-type segmentation decomposes a 3D object into sub-meshes by segmenting a surface into connected components.

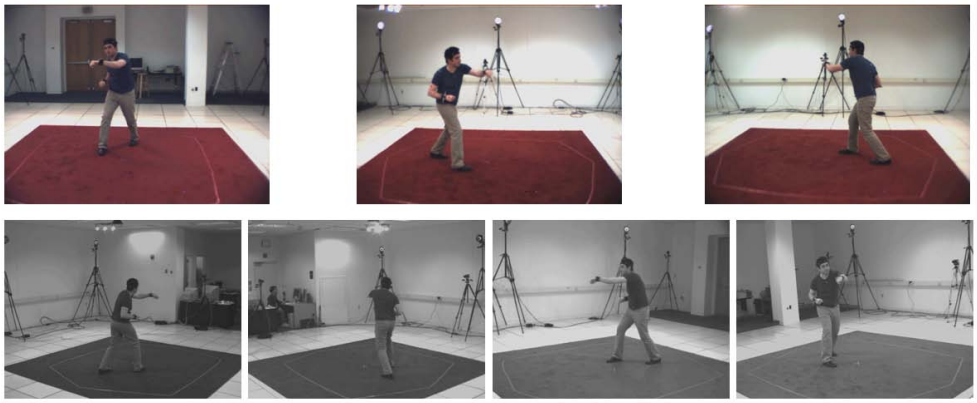


(a) 3D object tracking in 3D scene

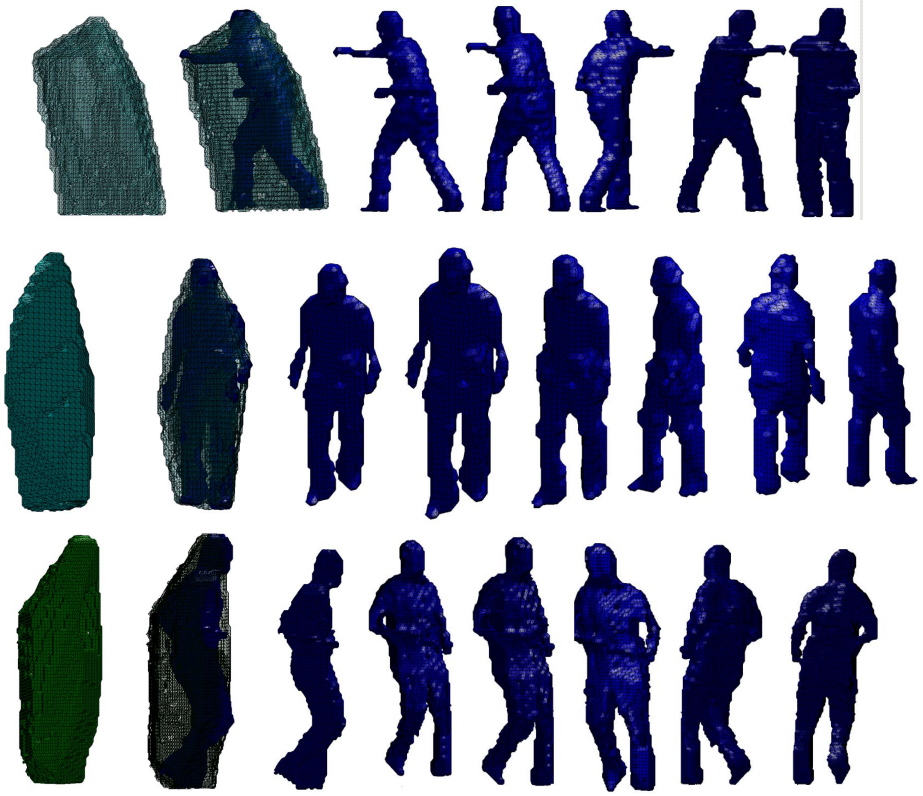


(b)Voxel carving in a tracked 3D lattice by checking the color consistency

Figure 4.3: 3D lattice configuration by tracking 3D boundary of target object and voxel carving using color consistency check within 3D lattice.



(a) Input images which are captured from multiple viewpoints



(b) The 3D lattice and 3D reconstructed object within lattice and its view in different point of boxing, walking, and jogging.

Figure 4.4: Multiple images and its reconstructed object using our proposed methodology in different viewpoint.

In this section, we will explain how we extract the skeletal features from multiple images by using appearance model which is extracted from background subtraction, image based 3D reconstruction method and ellipsoidal decomposition of volume data.

4.3.1 NGVF fields from 3D reconstructed object

Originally, the Gradient Vector Flow (GVF) fields in 3D volume data were proposed to solve the problem of initialization and poor convergence to boundary concave objects yielding a traditional snake form. The Gradient Vector Flow which is a vector diffusion approach using Partial Diffusion Equations(PDEs) is a dimensional extension of 2D Gradient Vector Flow which already explained in last chapter 3. It converges towards the object boundary when very near to the boundary, but varies smoothly over homogeneous image regions extending to the image border. The main advantages of the Gradient Vector Flow fields is that it is able to capture a snake from a long range and could force it into concave regions. Mathematically defined, the Gradient Vector Flow fields are the vector fields v that minimize the following energy functional,

$$E(V) = \int\int\int \mu|\nabla V|^2 + |\nabla f|^2|V - \nabla f|^2 dx \tag{4.4}$$

where $x = (x,y,z)$, μ is a regularization parameter, and $f(x)$ is an edge map derived from the volume, Vox. For a binary volume, $f(x) = -Vox$. Minimizing this energy will force fx nearly equal to the gradient of the edge map where $||\nabla f||$ is large. Nevertheless, the general Gradient Vector Flow method cannot efficiently extract the medial axis as a weak vector has very little impact on its neighbors that have much stronger magnitudes.
The Normalized Gradient Vector Flow fields can tremendously affect a strong vector, both in magnitude and in orientation by normalizing the vectors over the image domain during each diffusion iteration. The traditional Gradient Vector Flow fields have difficulty preventing the vectors on the boundary from being significantly influenced by the nearby boundaries and thus causes a problem such that the ”snake” may move out of the boundary gap. The Normalized Gradient Vector Flow fields complement the weakness of the traditional Gradient Vector Flow fields.

4.3.2 Ellipsoidal Decomposition of 3D Volume Data using Tensorial Features of 3D Model

Previous 3D skeleton extraction and splitting methodologies are mainly based on vector or scalar transformations because similarity measures within these spaces provide a familiar perception of the human eye. Nevertheless, tensorial maps contain and provide more information than scalar ones as to measure the similarity between neighbor regions.

Even though a diverse array of 3D surface representation methodologies are introduced, there is no single representation method which satisfies the needs of all problems in various applications. Among varieties of 3D surface representation methodologies, the mesh based model is popularly used to visualize the target object, but it is still a problem under the research of visualizing the complex objects such as human body model because it requests so much data. The 3D ellipsoidal representation of a 3D deformable object is very efficient and effective to visualize and recognize its characteristics using few parameters. In this section, we briefly explain the concept of three dimensional second-order symmetric tensor fields and their properties allowing us to extract the dedicated features from 3D volume data and to visualize its characteristics with superquadric model [15,43,56,129,189,226,227]. The topological analysis of 3D volume data provides a simple but powerful representation of complex deformable objects or natural phenomena.

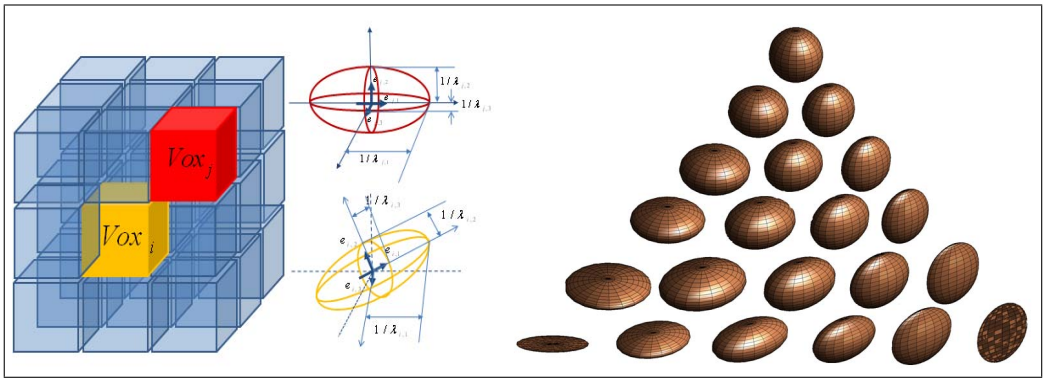
A tensor is the mathematical definition of a geometric or physical quantity whose analytic description consists of an array of scalars. This means that a tensor is an abstract object expressing some definite type of multi-linear concept. Hence, the tensor field commonly defined as a topological representation of a 3D symmetric, second-order symmetric tensor field is presented [61] as :

$$T = \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \tag{4.5}$$

where $T_{xy} = T_{yx}, T_{xz} = T_{zx}, T_{yz} = T_{zy}$ because the tensor is a symmetric positive definite matrix. This matrix can be reduced to its principal axes by solving the characteristic equation as:

$$(T - \lambda \cdot \mathbf{I})\mathbf{e} = \mathbf{0} \tag{4.6}$$

where $\mathbf{I}$ is the identity matrix, λ are the eigenvalues of the tensor and e are the normalized eigenvectors and the corresponding eigenvectors are orthogonal. In



(a) 3D ellipsoidal representation of each voxel from extracted eigenvalues and eigenvectors. (b) The superquadric tensor visualization as the change of eigenvalues' value from three dimensional second-order diffusion tensor fields.

Figure 4.5: Superquadric representation using the eigenvalues and eigenvectors from the properties of diffusion tensor fields. The scale and orientation of the superquadric model is determined by its eigenvalues and eigenvectors. The visualization and analysis using superquadric model of complex 3D model is very familiar with human visual perceptual system.

this case, the tensor in each pixel can be represented by an ellipse, where the main axis lengths are proportional to the eigenvalues λ ($\lambda_1 > \lambda_2 > \lambda_3$).

Figure 4.5 is a conceptual representation of the volume data how the superquadric models can be represented by extracted eigenvalues and eigenvectors from three dimensional second-order diffusion tensor fields. The ratio between eigenvalues determines the shape of the superquadrics, and its principal eigenvector direction defines the rotation of the superqaudrics as shown in Figure 4.5-(b). Hence, we can visualize 3D volume data using an ellipsoidal representation based on the properties of the diffusion tensor field space. Figures 4.6-(a) and 4.6-(b) show the ellipsoidal decomposition of 3D volume data [22] which is reconstructed by multiple images at a certain level of detail, detailing shape and orientation of each superquadric primitive. Figures 4.6-(c) and 4.6-(d) are also superquadric representation of various 3D models like animals or tools that are high-degree of freedom. We can see that the 3D target objects using our superquadric representation still remains familiar to the visual perception and provides a more intuitive access to understand the objects' characteristics. Main advantage is that we only need few parameters for the representation of the model.

4.3.3 Skeleton extraction using ellipsoidal representation

In this section, we will explain an automatic skeleton extraction and refinement using a tensor topological analysis in the space of Normalized Gradient Vector Flow fields. Previous skeleton extraction and its structural analysis methodologies are computed in the vector fields. Its topology is obtained by locating critical points and displaying the set of their connecting streamlines. When it comes for representing and analyzing the directions of Normalized Gradient Vector Flow on a 3D shape of the reconstructed object, tensor fields provide a larger vocabulary of visual elements than vector fields. The degree of anisotropy in each 3D ellipse can be quantified in a single number called a diffusion anisotropy index and it is represented as Fractional Anisotropy (FA). The Fractional Anisotropy representation method geometrically characterizes the shape of 3D ellipse of each voxel.

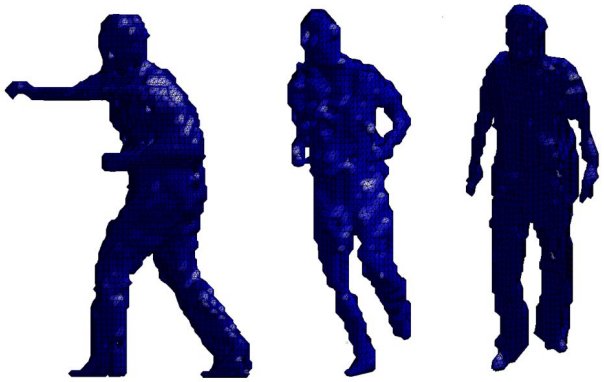
$$FA = \sqrt{\frac{3[(\lambda_1 - \mathbf{T}_{avg})^2 + (\lambda_2 - \mathbf{T}_{avg})^2 + (\lambda_3 - \mathbf{T}_{avg})^2]}{2 \times ((\lambda_1)^2 + \lambda_2)^2 + \lambda_3)^2)}}, \tag{4.7}$$

where $\mathbf{T}_{avg}$ is the average of λ_1, λ_2, and λ_3.

The skeleton of 3D reconstructed object is extracted by analyzing the Fractional Anisotropy values of each voxel and connecting the neighbor voxels by calculating its Fractional Anisotropy values. Figure 4.7 shows the extracted skeleton from the 3D reconstructed object using our proposed methodology. Comparing to skeleton extraction from 2D image which is explained in last chapter, the skeletal features from 3D model have more branches because we have to check 26 neighboring voxels per each skeletal element to extract the skeleton.

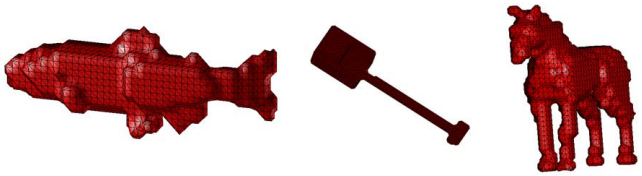
4.4 Automatic Skeleton Splitting

In this section, we will explain how we split the extracted skeleton by measuring the dissimilarity between neighbor skeletal voxels in the space of diffusion tensor fields. To split the extracted skeleton from a target object, we first extract the end points and branch points by checking its connectivity and tensorial properties. In a branch, we separate each branch when skeletal curves other directions.
The 3x3 symmetric tensor matrix can be used to measure the distance between neighbor tensorial elements. Since the key factor in tensorial analysis is the proper

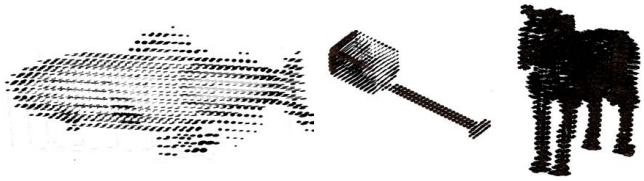


(a) 3D reconstructed human body model of boxing, jogging, and walking

(b)Superquadric decomposition from volume data of boxing, jogging, and walking



(c) Various 3D models like fish, tool, and horse



(d)Superquadric decomposition from volume data of various deformable objects

Figure 4.6: Superquadric decomposition from volume data of 3D volume data.

(a) Image based reconstructed 3D volume data

(a) Skeletal features from unlabeld 3D model

Figure 4.7: 3D model and its skeletal structure using our proposed method.

choice of the similarity measure to be used, several works have been published on this properties [58].

Given two tensorial elements within the 3D skeleton such as $\mathbf{T}_i$ and $\mathbf{T}_j$, the most simple comparison between two tensor quantities is the dot product between the principal eigenvector direction.

$$d_1(\mathbf{T}_i, \mathbf{T}_j) = |e_{1,i} \cdot e_{1,j}| \tag{4.8}$$

where $e_{1,i}$ and $e_{1,j}$ are principal eigenvectors of tensors $\mathbf{T}_i$, and $\mathbf{T}_j$, respectively. Another simple similarity measure is the tensor dot product,

$$d_2(\mathbf{T}_i, \mathbf{T}_j) = \sum_{i}^{3}\sum_{j}^{3} \lambda_i^1 \lambda_j^2 (e_i^1 \cdot e_j^2)^2 \tag{4.9}$$

One such example is the tensorial Euclidean distance obtained by using Frobenius norm. Due to its simplicity, tensorial Euclidean distance has been used extensively in Diffusion Tensor restoration.

$$d_d(\mathbf{T}_i, \mathbf{T}_j) = \sqrt{Trace((\mathbf{T}_i - \mathbf{T}_i)^2)} \tag{4.10}$$

From various similarity measure methods, we measure the similarity measure, $Score_{i,j}$ of two voxels, Vox_i and Vox_j, is defined as $d_1(\mathbf{T}_i, \mathbf{T}_j) \times d_3(\mathbf{T}_i, \mathbf{T}_j)$ because these two similarity measures represent the difference of scale and angle of 3D ellipse between neighbor skeletal voxels. The similarity measure to merge and split the skeletal voxels is as follows:

The 3D model segmentation procedure is computed by iterative region growing method. In the initial state.

STEP0 : initially, the numbers of subregions of human body model is equal to the number of voxel of 3D human model and calculate the $Score_{i,j}$ where similarity measure between voxel i and j.

STEP1 : we progressively merge the neighbor voxels if $Score_{i,j}$ is less than threshold and recalculate the average of *Score* for merged subregions, $Score_{sub} = \frac{1}{n}\sum_{k=1}^{n} Score_k$ of merged sub-region which have n voxels. However, the neighbor voxels are not merged and remain to split if the $Score_{i,j}$ are over the threshold.

STEP2 : Iteratively merge and split the subregions until there is no subregions whose the $Score_{sub}$ is less than threshold.

The extracted end points, branch points, and joint points are shown in Figure 4.8. In Figure 4.8-(b), the segmented regions from skeleton are painted using various colors to see where it is segmented.

4.5 Experiments

We setup our proposed methodology with a Pentium 4 1.2 GHz CPU and a CUDA which is a technology for GPU computing from NVIDIA, Geforce 8200. It exposes the hardware as a set of SIMD multiprocessors, each of which consists of a number of multiprocessors. Our system is implemented in the GPU architecture which is specialized for parallel computing task. The graphic hardware consists of a set of processing grouped together in a common multi-processing block. CUDA is a technology for GPU computing from NVIDIA. These multiprocessors have arbitrary read/write access to a global memory region called the device memory, but also share a memory region known as shared memory. The implementation strategy has a great impact on the overall performance of the implementation. It deals with the allocation of threads to the problem and the usage of the different types of onboard memory. Figure 4.9 shows the structure of CUDA for parallel implementation of complex calculation.

For experiments, we tested the HumanEva database which have 7 calibrated cameras, 3D models of Princeton university and images in our environment using calibrated 4 cameras. In our environment, we used 4 Firewire cameras that have 640x480 resolution and extract the silhouettes using kernel density estimation based background subtraction [90]. Initial 10 frames are recognized as background image and then we continuously update the background to reduce the effect of illumination change and noise. The 3D object is reconstructed in volume dimension 128x128x128.

4.5.1 Camera calibration

In the experiments for camera calibration, we used the laser points to easily extract the corresponding points from multiple cameras. We extract 100 corresponding points from corresponding stereo cameras. We calculate the 3x1 translation matrix, T, 3x3 rotation matrix, and focal length, f to know how much the cameras are moved from the origin of the world coordinate system. Figure 4.10 shows the camera position in the world coordinate system.

(a) Skeletal features from unlabeld 3D model

(b) Skeleton splitting from unlabled skeletal features

Figure 4.8: 3D skeleton extraction and splitting using tensor based similarity measure.

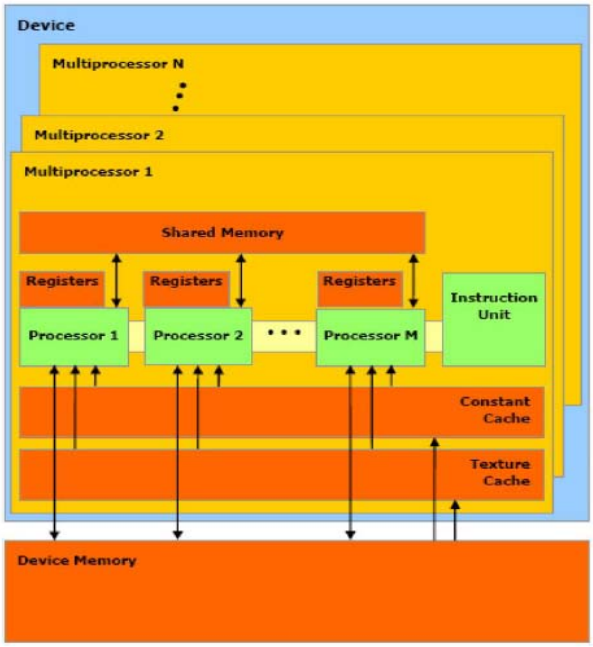


Figure 4.9: The structure of CUDA which is a technology for GPU computing from NVIDIA which is based on CUDA tutorial.

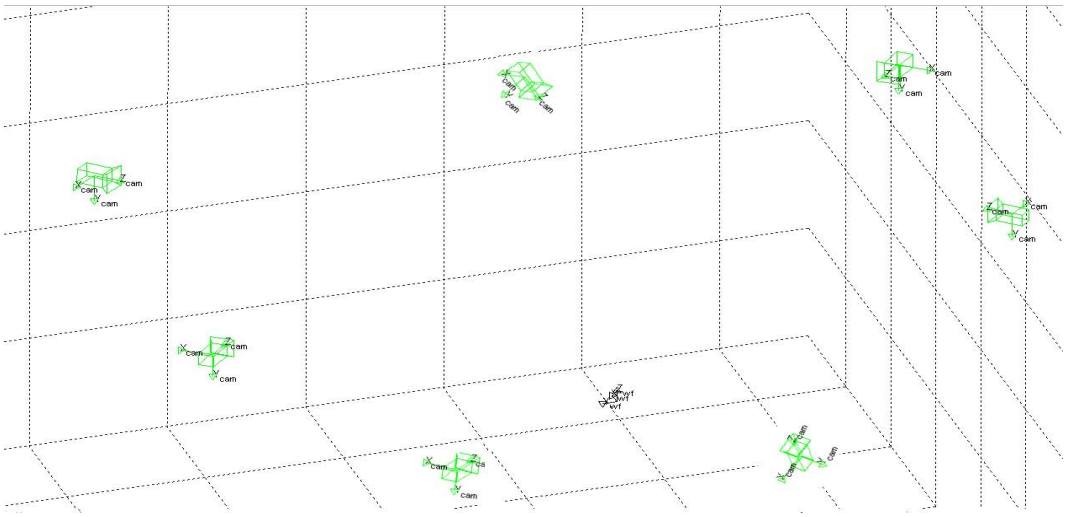


Figure 4.10: Camera calibration and its camera position.

Table 4.1: Running time for 3D action recognition from multiple images using 128x128x128 dimensional human body model

Category	module	total time (ms)	time(ms)
3D Reconstruction		153	
	Segmentation of 7 images		38
	Tracking of 7 images		15
	3D lattice construction		64
	Voxel carving in 3D lattice		72
3D Segmentation		1286	
	Superquadric decomposion		28
	Segmentation		1258
3D Action classification		28	
	MK-SVM classification		28

4.5.2 3D reconstruction

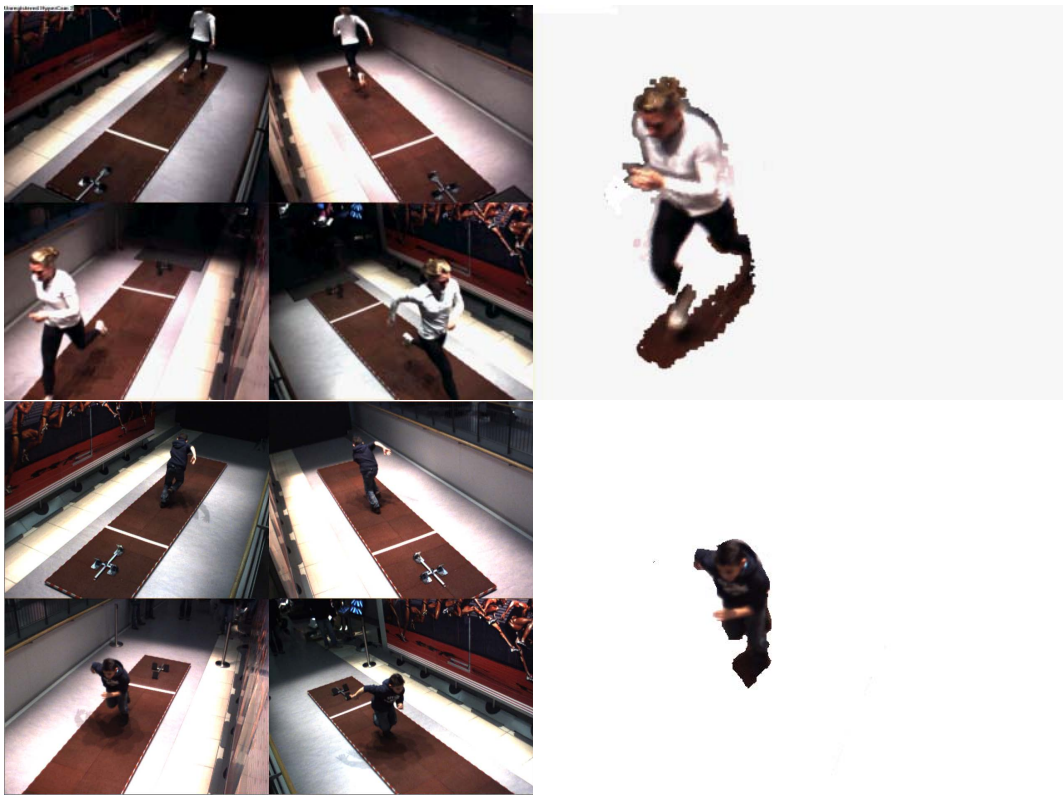
We conducted some experiments to analyze the athletes motion in real-time from a set of 7 images from HumanEva database and 4 images from our experimental environment. Figure 4.11-(a) shows the 3D reconstruction from 4 multiple images in our experimental environment, and Figure 4.11-(b) is the 3D reconstruction using HumanEva dataset. The 3D reconstructed object using HumanEva is not painted in color because some input images are gray images.

Our 3D reconstruction methodology is implemented in GPGPU environment to render its shape in real-time. Table 4.1 shows the average running time of 3D reconstruction. Its running time is analyzed by every module.

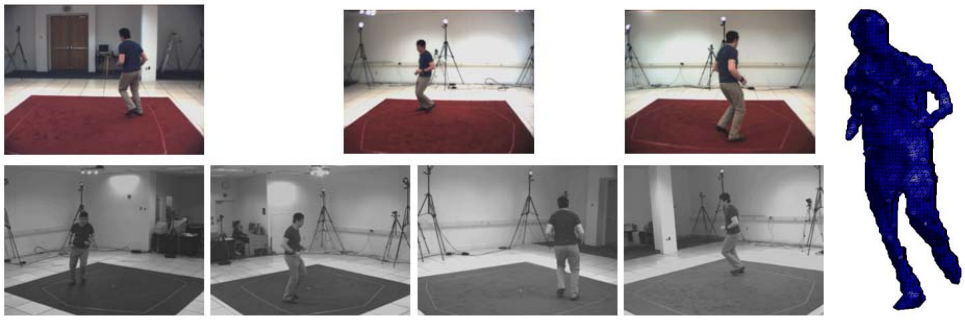
We also compare the accuracy of 3D reconstruction with our approach and original voxel coloring method. Figure 4.12-(a) shows the 3D reconstructed model using our approach and 4.12-(b) is the original voxel coloring approach in the dimension of 128x128x128. As shown in Figure 4.12, our proposed method is more detail in arms and legs of the deformable object because we discrize the 3D scene within the intersection of convex cone of each camera, but the original voxel coloring method discrizes the whole 3D scene. In particular, our proposed method is more accurate to represent the characteristics of real object in the arms and legs of 3D reconstructed human body models.

4.5.3 Skeleton Extraction

The skeleton feature extraction and splitting method from the unlabeled volume data is explained in this section.



(a) 3D reconstruction from our environment



(b) 3D reconstruction using HumanEva dataset

Figure 4.11: Photo-realistic 3D reconstruction of target object in real-time.

(a) Original voxel coloring

(b) Our proposed 3D reconstruction method

Figure 4.12: Comparison of 3D reconstruction between our proposed and original voxel carving method.

We have conducted skeleton extraction and splitting from a set of the HumanEva database and 3D models without 3D reconstruction procedure. We also conduct the skeleton extraction and splitting from the 3D model data of Princeton university. Figure 4.13 shows skeleton extraction from 3D deformable model and split skeletal features are painted by different colors.

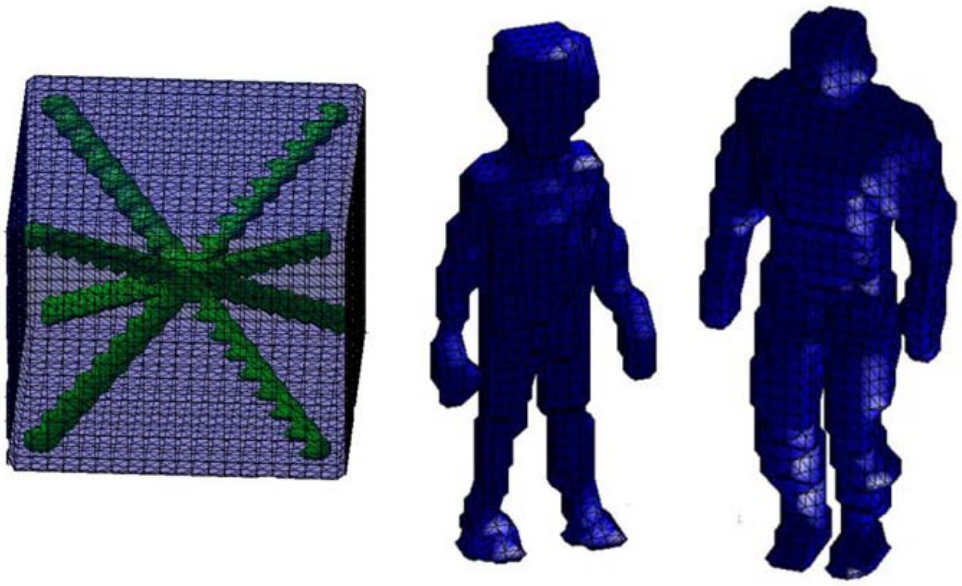
Our proposed methodology can be applied to medical volume visualization and segmentation. Figure 4.14 is a 3D volume data of a human brain whose dimension is 512x512x512. This Figure 4.14 proves that our proposed methodology can be used for various volume data with have high dimension. Especially, the medical volume data segmentation and rendering system from CT or MRI provides an important information to medical doctor. With support of this information, they can easily separate the complex 3D medical volume data which have similar character and find non-normal areas within our human body part.

We also lead the experiments to compare our approach with previous skeleton extraction methodologies which use the gradient and divergence distance within 3D volume data. Figure 4.15 is the extracted skeleton from simple 3D cubic volume data to easily visualize their difference. Figure 4.14-(a) is our proposed method and 4.15-(b) and 4.15-(c) are based on the distance using gradient and divergence from the surface of 3D volume data. We can see that the skeleton using our approach is sharper than other approaches shown in 4.15-(b) and 4.15-(c).

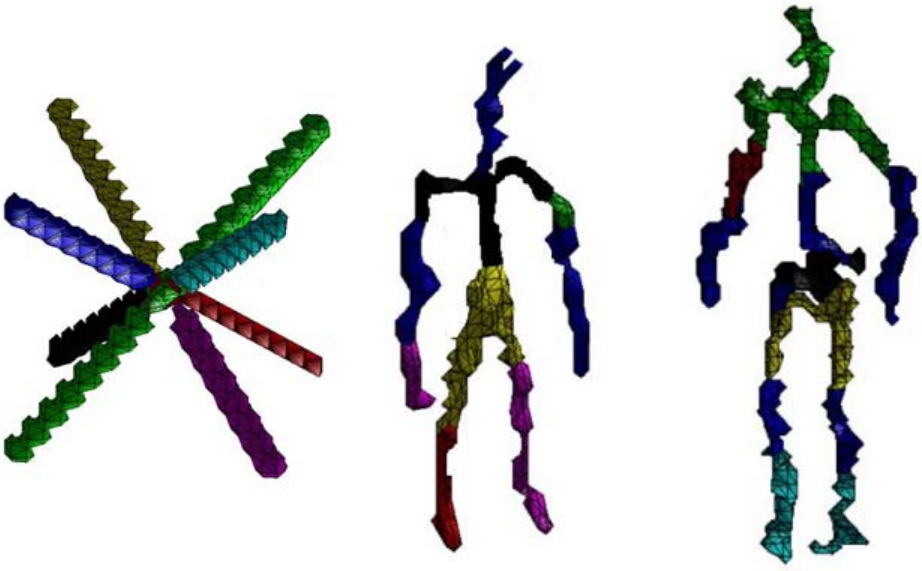
4.5.4 Comparison of our approach and Pseudo-Zernike Moment based approach

In this section, we will compare to our skeletal feature based motion analysis and model based approach. In markerless motion capture and analysis, 3D motion analysis method can be classified into model based methods and non-model based method. In this section, our proposed skeletal based motion analysis will be compared to Zernike model based motion analysis method.

The efficient algorithms for a fast computation of Pseudo-Zernike moments (PZM) especially for image processing tasks have been proposed in various publications such as [18, 120, 170, 193] and are the basis for the 3D reconstruction of full body motion. The reconstruction of body moments using the Pseudo-Zernike Moment is based on the minimization of the difference between artificial and real silhouettes within all camera images. The most important constraint of the Pseudo-Zernike Moment algorithm in order to achieve real time capabilities, is a discretization of possible movements by the user. The intension of the visu-

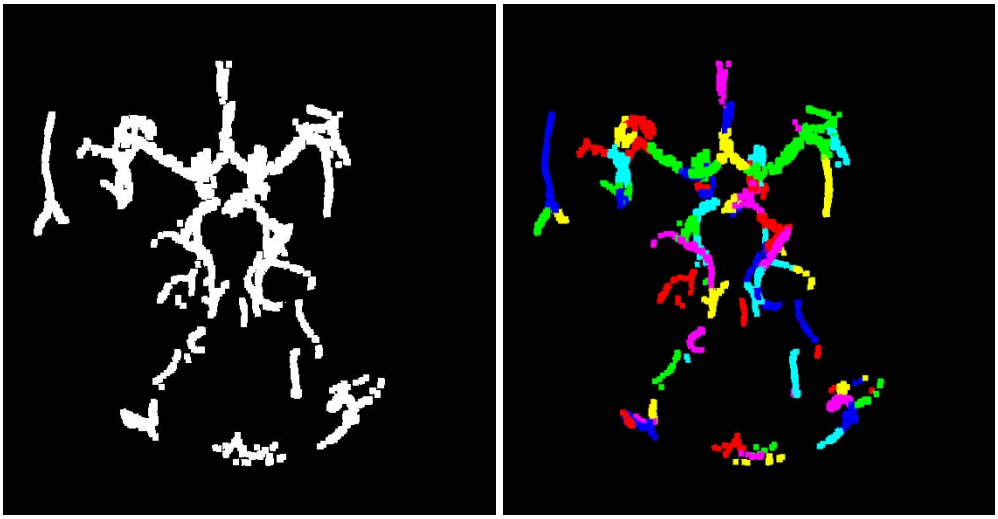


(a) Unlabeld 3D model



(b) Extracted skeleton and its splitting

Figure 4.13: 3D skeleton extraction and its splitting from Princeton 3D model dataset.



(a) Unlabeled volume data (b)3D segmentation of volume data using our proposed method

Figure 4.14: 3D volume segmentation for medical volume visualization.

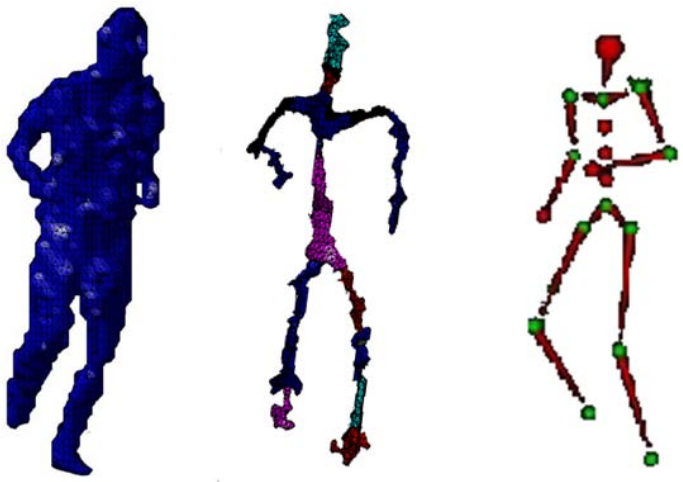
(a) Our approach (b) Gradient distance based skeleton extraction (c) divergence from the surface of 3D volume data

Figure 4.15: Comparison of 3D skeleton extraction from 3D cubic volume data.

alization is not a precise reconstruction of the user's body poses to analyze the slight changes of joint rotations from frame to frame, but to capture the overall movement of the user. Therefore, motion capture data, which has been recorded prior to the reconstruction phase is used as the basis for the evaluation of possible best-fit body movement. Due to the fact that motion capture data consist of skeleton poses with up to 120 frames per second, these sequences are reduced to fewer poses using ten frames per second only. For the rendering of the avatar and the creation of the silhouettes the Scenegraph system OpenSG is used. After loading the model into the memory the Scenegraph is parsed and all color nodes are automatically set to white, which ensures a white silhouette in front of a black background equivalent to a segmented and binarized silhouette of a real image calculated by the segmentation procedure. Figure 4.16 shows the comparison of our proposed skeleton extraction from 3D reconstructed object and Pseudo-Zernike Moment based motion analysis. Our proposed method has more noise than model based approach in the arms and foot, but our approach split the skeleton according to its movement in detail comparing to model based approach. For example, the foot of right leg is segmented into four regions, but the Pseudo-Zernike Moment approach only segments two regions. Due to the fact that for the calculation/generation of real and artificial silhouettes the same camera parameters are used, silhouettes should match in several images. Nevertheless, the rotation of the avatar in 3D space plays an important role. It is obvious, that small changes of the rotation of the virtual model lead to different silhouette representations. Therefore, a bunch of silhouettes for each predefined body pose with different rotation parameters for the root node of the avatar is generated. For the sprint starting sequence the real orientation of the user is approximately known, because the sprinting direction of the user and his/her movements are known by the setup of the environment.

4.6 Summary

In this chapter, we explained the detail on how to extract the skeleton. We then merge and split the extracted skeleton from the 3D reconstructed object by using multiple images. Basically, the 3D skeleton extraction and splitting is an extension of 2D image which was explained in Chapter 3. However, 3D reconstruction methods are added a similarity measure for deformable 3D model is developed in this chapter. Our proposed 3D motion analysis using skeletal features from complex 3D volume data was very efficient and effective in visualizing and rendering the target object in real-time.



(a) Unlabeled volume data (b) Our proposed method (c) Pseudo-Zernike Moment based approach

Figure 4.16: 3D motion analysis comparison between our approach and Pseudo-Zernike Moment based approach.

We first provided an efficient 3D reconstruction methodology from calibrated multiple images by tracking its 3D position of target object and voxelized the 3D scene by intersecting the 3D convex cones which were built by the convex hulls and camera's center position. The 3D model was carved by checking its photo-consistency within the tracked the 3D volume area, not the whole 3D scene. 3D reconstruction of the target object by tracking the 3D position of the real object was better than original voxel coloring and image based visual hull methodologies in a large environment using few multiple cameras.
Within the 3D reconstructed object, we then analyzed the Normalized Gradient Vector Flow in the space of diffusion tensor fields to extract the skeletal features. We decomposed each voxel into 3D ellipsoidal model whose scale and rotation were determined by three dimensional second order diffusion tensor fields. Its eigen-features of each voxel were used for skeletal extraction and splitting within complex 3D model. We iteratively split and merged the skeletal features by measuring the dissimilarity between neighbor voxels.
As we showed in the experimental results, we have compared our proposed system to Pseudo-Zernike model based motion analysis approach . Proposed skeletal extraction and splitting methodology was also used for medical volume visualiza-

tion and analysis in large dimension. In particular, the tracking based 3D model reconstruction method was very efficient to voxelize the very large 3D scene and carving the voxels which had high quality than original voxel coloring and image based visual hulls. Lastly, we compared our markerless motion capture and analysis to a model based motion analysis method using a Pseudo-Zernike Moment technique.

Our system did not require any prior information to analyze the complex objects which were reconstructed from multiple camera. So it can be applied to recognize and retrieve the 3D objects like chair, bicycle, and tools in media database as we will show in the next Chapter.

Chapter 5

Applications

5.1 Introduction

In this chapter, we will discuss how we use the skeletal features and its characteristics from 2D image and 3D volume data for various real life applications. As we surveyed the related works in the areas of markerless motion capture and analysis in Chapter 2, there are various applications in the area of computer vision and computer graphics. From several applications, we will apply our basic principles to human action recognition and sketch-based object detection and retrieval system. Human action recognition and sketch-based image retrieval are an important issues in Human Computer Interaction, because these topics will be very useful to understand human's attention and intention. For example, human action recognition and behavior understanding can be applied for video surveillance and monitoring, user pose estimation for sport scene analysis and computer animation, and user's attention analysis in virtual reality.
Furthermore, sketch based target object detection and retrieval will be useful in interaction with various tools like tablet PCs, touch-based cell phones, and multi-touch screens. These devices are invented for comfortable interaction. By drawing a simple sketched image of a target object, we can extract the key features from the sketched image and retrieve the images from a large database.
This Chapter discussing applications using our proposed methodology is separated into two topics, namely human action recognition from 2D images or 3D reconstructed models, and sketch-based image retrieval. In section 5.2, we will explain the details of our proposed methodology for human action classification and recognition using multiple-kernel Support Vector Machines (MK-SVM). Then,

sketch-based image retrieval using an hierarchical clustering technique is shown in section 5.3. The experimental results for human action recognition and sketch-based image retrieval are evaluated at the end of each section.

5.2 2D/3D Human action recognition

Human action analysis and recognition [2, 13, 24, 44, 45, 68, 69, 85, 86, 138, 139, 162, 186, 198, 228–230, 267, 270, 272, 273], defined to understand the basic human actions such as jogging, walking and boxing from images, has a long history in the area of computer vision and computer graphics. It gives rise to a great deal of applications such as automated surveillance [13, 138, 139], smart home applications [198], video indexing and browsing [2], virtual reality [44], human-computer interaction [45], and analysis of sports events [69, 85, 270]. Human action recognition from a single 2D image is heavily studied by numerous researchers, but still a challenging issue due to partial occlusion, clutter, dependence of viewpoint, and pose ambiguity within a 2D image. In multiple camera environments, the number of observables which can be used for human action recognition are extended. They are more reliable than single image based methods and independent on viewpoint, but the system is more complex and difficult in a high-dimensional and multi-modal space.
On the other hand, the problems of human action recognition, which can be interpreted as one part of object recognition and retrieval [10, 78, 103, 104, 118, 126, 131, 171, 174, 178, 181, 185, 231, 260, 276, 279], are how to define the appropriate similarity measure from the reliable features of 3D deformable models and to automatically assess the similarity between any pair of 3D human motions based on a suitable notion of similarity.
We propose to solve these problems of 2D/3D human action recognition by focusing on adequate feature extraction and separating human body model into several human body parts like head, torso, and limbs. In particular, the 3D human action recognition is based on 3D reconstruction using calibrated multiple images and appearance model of a target object, and similarity measure from the features which are extracted from three-dimensional second order diffusion tensor fields. We will explain the details of our proposed methodology for 3D reconstruction and segmentation from multiple images and the human action classification and recognition using multiple-kernel Support Vector Machine in the following Section.

5.2.1 Human action classification using Multiple-Kernel Support Vector Machine

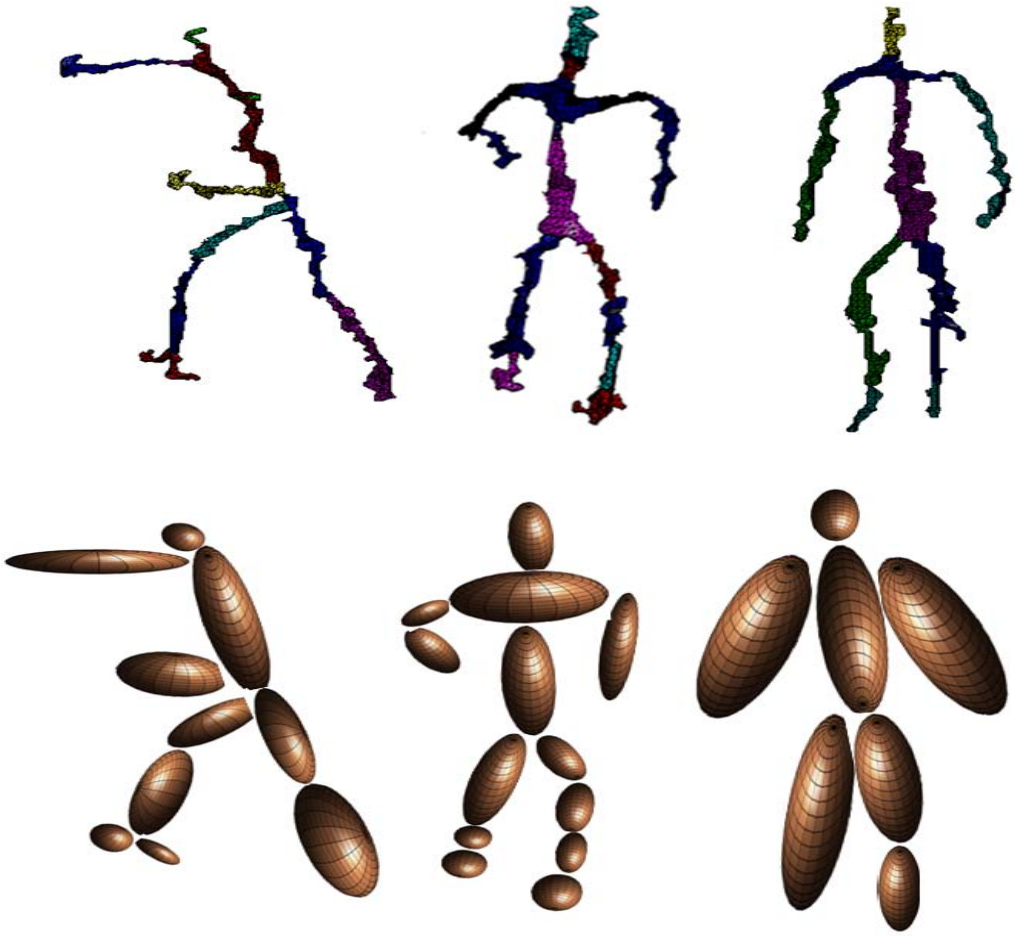
In previous Chapter 3 and 4, we already explained how we extract the eigen-features from segmented areas of the deformable 2D/3D model. Using the ellipsoidal decomposition of the 2D/3D human model, we segmented the human model according to its tensorial characteristics. Figure 5.1 shows the splitting ares of the target object and its ellipsoidal representation of the segmented regions from the 2D image or the 3D model. The eigen-features of the segmented skeletons from the 2D image or the 3D volume data will be used as the input data for human action recognition.

The eigen-features of the segmented object, $x_i = \Sigma_{s=1}^{S}\Sigma_{j=1}^{k}\lambda_{ij}$ represent the scale and rotation of the segmented skeleton. It is used for classification of human action recognition, where S is the number of the segmented human body parts and $x_i \in \chi$ (the input space χ). Figure 5.2 visualizes the scenario for human action recognition using the averaged eigen-features of the segmented regions. We first extract the skeletal features from image database and calculate the averaged eigenvalues and eigenvectors for each split region. The averaged eigen-features are used as input information to classify the various human actions.

The performance of different classifiers which are applied in object detection and recognition has been evaluated and compared in the area of pattern recognition and data learning. Bazzani concluded that the Support Vector Machine (SVM) performed better than the Multi-Layer Perception (MLP) for a small number of training data [17]. Papadopoulos [184] also showed that Support Vector Machine achieved a higher accuracy rate than Neural Network. Having evaluated the Support Vector Machine, Kernel Fisher Discriminant (KFD), Relevance Vector Machine (RVM), Feedfoward Neural Network (FNN), and committee machines, Wei. et al. [266] concluded that the Kernel based classification yielded the best performance. In this section, multiple-kernel based support vector machine is used for classifying the complex human action according to the proofs of previous researches.

In $\{x_i, y_i\}_{i=1}^{l}$ where l denotes the number of training features, each x in χ is then mapped to a $\Phi(x)$ and y_i is separated into human actions like boxing, jogging, and walking. The non-linear Support Vector Machine maps the training samples from the input space into a higher-dimensional feature space via a mapping function Φ

(a) Ellipsoidal representation from 2D skeleton extraction



(b) Ellipsoidal representation from 3D skeleton extraction

Figure 5.1: Ellipsoidal representation of segmented skeleton from our proposed method.

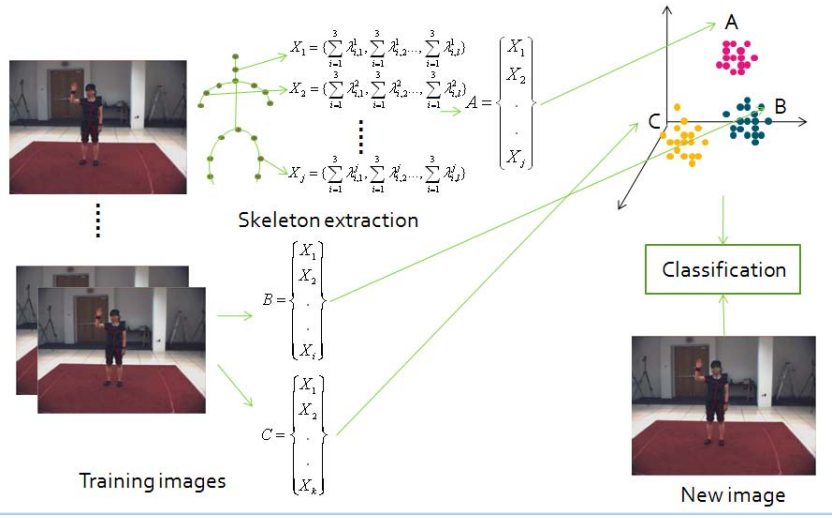


Figure 5.2: Scenario of skeletal feature based human action recognition.

and construct a hyperplane defined by $w\Phi(x)+b=0$ to separate examples from the classes. $\{x_i,y_i\}_{i=1}^{l}$ in the kernel-induced feature space is related to the kernel function K which intuitively computes the similarity between examples in Support Vector Machine. The standard Support Vector Machine [46,53,209,241,264] tries to find a hyperline $w^T\Phi(x)+b$ that has large margin and small training error. Mathematically, this follows from the:

$$min_{w,b,\xi}\frac{1}{2}||w||^2+C\sum_{i=1}^{l}\xi_i s.t. y_i(w^T\Phi(x_i)+b)\succeq 1-\xi_i,\xi_i\succeq 0, i=1,2,,...,l. \quad (5.1)$$

Here, $\xi=[\xi_1,\xi_2,...,\xi_l]^T$ is the vector of slack variables for the errors, and C is a user-defined regularization parameters that trades off the margin with error.
Instead of having a single kernel K, suppose that we have a set of M base kernels $K_1,K_2,...,K_M$ with corresponding kernel-induced feature maps $\Phi_1,...,\Phi_M$. The multi-kernel Support Vector Machine is extended from the single-kernel based Support Vector Machine and shown like that:

$$min_{w,b,\xi}\frac{1}{2}(\Sigma_{k=1}^{M}||w_k||)^2+C\sum_{i=1}^{l}\xi_i s.t. y_i(\Sigma_{k=1}^{M}w_k^T\Phi_k(x_i)+b)\succeq 1-\xi_i,\xi_i\succeq 0, i=1,2,,...,l. \quad (5.2)$$

where $w=\{w_1,w_2,...,w_M\}$, ξ is the non-negative slack variables, and w_k is the weight for component Φ_k.

Support Vector Machine has proven to be powerful for a wide range of different data analysis problems. They employ a so-called kernel function $k(x_i, x_j)$ which intuitively computes the similarity between two examples x_i, and x_j. The result of Support Vector Machine learning is a α-weighted linear combination of kernel elements and the bias b.

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b) \tag{5.3}$$

The regularization parameter C determines the trade-off between the maximization margin $\frac{1}{||w||^2}$ and the minimum experience risk.
The multiple kernel based Support Vector Machine (MK-SVM) [223] is extended to following equation from single kernel.

$$k(x_i, x_j) = \sum_{k=1}^{K} \beta_k k(x_i, x_j) \tag{5.4}$$

with $\beta_k \geqslant 0$ and $\sum_k \beta_k = 1$. where each kernel k_{x_k, y_k} uses only a distinct set of features of instance.
For our human action recognition, we used two kernels like Radial Basis Function (RBF) and quadric kernels for efficient human action recognition.

5.2.2 Experiments

We setup our proposed methodology with a Pentium 4 1.2 GHz CPU and a CUDA which is a technology for GPU computing from NVIDIA Geforce 8200. It exposes the hardware as a set of SIMD multiprocessors, each of which consists of a number of multiprocessors.
Our experiment is separated into 2D human action recognition and 3D model based human action recognition.

2D human action recognition

For 2D human action recognition, we used public image data, such as HumanEva database and KTH human action dataset to compare previous human action recognition approaches to ours. The HumanEva dataset is separated into boxing, walking, and jogging, and KTH human action database is more specifically separated into 6 human actions like hand-clapping, hand-waving, jogging, running, walking and boxing. The HumanEva database which contain 7 calibrated video sequences which is composed of synchronized 3 color images and 4 gray images

with 640x480 resolution. Table 5.1 shows the human action recognition ratio of HumanEva dataset for 7 different viewpoints. The variation of acceptance ratio is very dependent on camera's viewpoint. In the dataset, the actors/actrees are moving around the environment, so the variation of acceptance ratio for jogging and walking is less than for boxing actions. The variations of boxing action recognition ratio are larger than other human actions like walking and jogging because boxing action is more dependent on other human actions. In the boxing action, the acceptance ratio of the camera view 4 and 7 are less than those of other camera viewpoints because we could not extract the features which are closely related to boxing actions. Figure 5.3 shows the error images from view 4 and view 7. The skeleton from view 4 and 7 is very simple, so we can not extract the typical features of boxing action.

Using the HumanEva dataset, we have compared our Multiple Kernel Support Vector Machine classification methodology to K-Nearest Neighbor(KNN) [223] and Single Kernel Support Vector Machine (SK-SVM) [46, 53, 209, 241, 264]. This result which is shown in Table 5.2 displays that our proposed Multiple Kernel Support Vector Machine methodology [100] is better than other approaches. In Multiple Kernel Support Vector Machine methodology, various kernels like Radial Basis Function(RBF), quadric, and linear kernels are used for robust action recognition. On the other hand, only Radial Basis Function kernel is used in Single Kernel Support Vector Machine based human action recognition. As shown in Table 5.2, the Multiple Kernel Support Vector Machine based human action recognition acceptance ratio is better than K-Nearest Neighbor and Single Kernel Support Vector Machine based human action recognition.

The KTH human action dataset is less dependent on camera viewpoint than the HumanEva dataset, but its image resolution is less than the HumanEva dataset and it has noise in the background. Figure 5.4 shows the example images of human actions like boxing, jogging, running, walking, hand-clapping, and hand-waving in KTH human action dataset. We also compared our systems to other classification methodologies using K-Nearest Neighbor and Single Kernel Support Vector Machine methodologies. We have used 1000 training images and 2500 images per human actions including multiple persons. Table 5.3 is the matrix of human action recognition using KTH human action dataset.

Table 5.4 shows the matrix of acceptance ratio of each human actions. In Table 5.5, we have compared our approach to previous approaches using optical flow using SVM classification method [57], local feature based human action recognition [5, 210], spatial-temporal feature based human action recognition

Table 5.1: 2D human action recognition ratio of HumanEav Dataset for different viewpoint

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	95.1	3.4	1.5
Walking	300	800	3.6	87.5	8.9
Jogging	300	800	2.2	5.3	92.5

(a) View 1

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	97.3	1.6	1.1
Walking	300	800	6.2	83.4	10.4
Jogging	300	800	4.1	11.3	84.6

(b) View 2

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	93.4	4.1	2.5
Walking	300	800	1.3	91.5	7.2
Jogging	300	800	2.5	6.8	90.7

(c) View 3

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	86.3	7.3	6.4
Walking	300	800	2.7	88.2	9.1
Jogging	300	800	4.1	9.8	86.1

(d) View 4

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	97.3	2.1	0.6
Walking	300	800	3.2	86.4	10.4
Jogging	300	800	6.4	11.6	82.0

(e) View 5

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	93.4	4.7	1.9
Walking	300	800	4.7	88.1	7.2
Jogging	300	800	5.4	8.4	85.2

(f) View 6

Human action	Num. of traning images	Num. of testing image	Boxing	Walking	Jogging
Boxing	300	800	84.9	9.8	5.3
Walking	300	800	4.0	87.3	8.7
Jogging	300	800	5.7	9.5	84.8

(g) View 7

(a) Example images which are not correctly recognized images from view 4

(b) Example images which are not correctly recognized images from view 7

Figure 5.3: Example images which are not correctly recognized human action.

Table 5.2: 2D human action recognition ratio of HumanEav Dataset for different viewpoint and its comparison

Human action	View 1	View 2	View 3	View 4	View 5	View 6	View 7
Boxing	92.1	96.2	92.5	83.7	96.1	92.3	83.1
Walking	86.7	81.2	89.1	87.9	83.1	85.5	86.7
Jogging	91.5	81.9	89.1	84.3	80.7	83.1	85.5

(a) KNN based human action recognition ratio for HumanEva dataset

Human action	View 1	View 2	View 3	View 4	View 5	View 6	View 7
Boxing	95.1	97.3	93.4	86.3	97.3	93.4	84.9
Walking	87.5	83.4	91.5	88.2	86.4	88.1	87.3
Jogging	92.5	84.6	90.7	86.1	82.0	85.2	84.8

(b)SK-SVM based human action recognition ratio for HumanEva dataset

Table 5.3: 2D human action recognition ratio of KTH Dataset using different classification methods

Human action	Boxing	Walking	Jogging	Running	Clapping	Waving
Boxing	95	1	3	1	0	0
Walking	3	81	9	5	2	0
Jogging	1	3	89	7	0	0
Running	0	1	6	92	0	1
Clapping	1	1	2	3	87	6
Waving	0	1	3	4	2	90

Table 5.4: 2D human action recognition ratio of KTH Dataset using different classification methods and its comparison using K-Nearest Neighbor, Single Kernel Support Vector Machine, and Multiple Kernel Support Vector Machine.

Human action	Num. of training images	Num. of testing image	KNN	SK-SVM	MK-SVM
Boxing	1000	2500	91	94	95
Walking	1000	2500	78	76	81
Jogging	1000	2500	84	86	89
Running	1000	2500	61	89	92
Hand-clapping	1000	2500	79	85	87
Hand-waving	1000	2500	81	85	90

(a) Example images in KTH human action dataset

(b) Extracted skeletal features and its ellipsoidal representation

Figure 5.4: Example of human actions from KTH human action dataset.

methods [64, 173].

3D human action recognition

Within 3D model based human action recognition, our experiments are separated into three categories: (1) photo-realistic 3D reconstruction from multiple images, (2) 3D segmentation of human body model in the space of diffusion tensor fields, and (3) action recognition results from Multiple Kernel Support Vector Machine technique. The Table 5.6 shows the average running time of our procedure in a hardware accelerated environment using CUDA. In our pipelines, the 3D segmentation and feature extraction part consumes much more time than other parts in action recognition.

In this section, we conducted our proposed Multiple Kernel Support Vector Machine based 3D human action recognition and compared with a K-Nearest Neighbor classification, and a Single Kernel Support Vector Machine. The HumanEva dataset is used for testing our proposed methodology. It provided various human motions for four different people. We reconstructed 3D human model of boxing, jogging and walking actions and trained the tensorial features. We trained 600 human body models per human action and tested 1200 human models which are not included in training models. Table 5.7 is the human action recognition ratio using standardized dimension of 3D human body model at 64x64x64 and 128x128x128. Compared to 2D image based human action recognition, the acceptance ratio of each human action recognition from 3D volume data is higher than the lowest acceptance ratio from 2D image.

Figure 5.5 shows the example images and the 3D model which is not accepted in our methodology. The action is classified into ”boxing”, but is recognized to ”walking” category because the main human upper body part is very close to the two arms. We could not separate the arms and the torso.

In Table 5.8, we show the recognition ratio to K Nearest Neighbor and Single Kernel Support Vector Machine to proof the robustness of our proposed system. The acceptance ratios of 3D human actions of boxing, jogging, and walking in the dimension of 128x128x128 and 64x64x64 in Table 5.7 are higher than other classification ratios in Table 5.8.

We also tested a 3D motion action recognition from four image sequences which have illumination change and clutters to see the robustness of our proposed method in our experimental environment. We tested 500 3D human models (500 images from 4 cameras which have 640x480 resolution) in a large environ-

Table 5.5: Comparison of KTH human action recognition method

Human action	Boxing	Walking	Jogging	Running	Hand-clapping	Hand-waving
Boxing	86	0	0	0	14	0
Walking	0	89	0	11	0	0
Jogging	0	92	8	0	0	0
Running	0	0	8	92	0	0
Hand-clapping	22	0	0	0	78	0
Hand-waving	13	0	0	0	12	75

(a) Human action recognition using Optical flow and Support Vector Machine [57]

Human action	Boxing	Walking	Jogging	Running	Hand-clapping	Hand-waving
Boxing	82	0	0	0	0	18
Walking	0	100	0	0	0	0
Jogging	1	40	58	40	0	0
Running	0	9	0	91	0	0
Hand-clapping	7	0	0	0	89	4
Hand-waving	1	0	0	0	4	95

(b) Human action recognition in compressed domain [5].

Human action	Boxing	Walking	Jogging	Running	Hand-clapping	Hand-waving
Boxing	82	0	0	0	0	18
Walking	0	84	16	0	0	0
Jogging	0	23	60	0	0	0
Running	0	6	39	55	0	0
Hand-clapping	35	0	0	0	60	5
Hand-waving	13	0	0	0	16	74

(c) Human action recognition using local Support Vector Machine [210].

Human action	Boxing	Walking	Jogging	Running	Hand-clapping	Hand-waving
Boxing	93	0	0	0	6	1
Walking	1	89	4	5	1	0
Jogging	0	20	56	24	0	0
Running	0	3	13	84	0	0
Hand-clapping	22	0	0	1	76	1
Hand-waving	7	0	0	1	4	88

(d) Human action recognition using spatio-temporal features [64]

Human action	Boxing	Walking	Jogging	Running	Hand-clapping	Hand-waving
Boxing	100	0	0	0	0	0
Walking	1	79	14	1	0	0
Jogging	0	11	52	37	0	1
Running	0	1	11	88	0	0
Hand-clapping	6	0	0	0	93	1
Hand-waving	23	0	0	0	0	77

(e) Human action recognition using spatio-temporal word [173].

Human action	Boxing	Walking	Jogging	Running	Hand-clapping	Hand-waving
Boxing	95	1	3	1	0	0
Walking	3	81	9	5	2	0
Jogging	1	3	89	7	0	0
Running	0	1	6	92	0	1
Hand-clapping	1	1	2	3	87	6
Hand-waving	0	1	3	4	2	90

(f) Our approach

Table 5.6: Running time for 3D action recognition from multiple images in 128x128x128 dimension.

Category	module	total time (ms)	time(ms)
3D Reconstruction		153	
	Segmentation of 7 images		38
	Tracking of 7 images		15
	3D lattice construction		64
	Voxel carving in 3D lattice		72
3D Segmentation		1286	
	Ellipsoidal representation		28
	Segmentation		1258
3D Action classification		28	
	MK-SVM classification		28

Table 5.7: 3D human action recognition ratio using HumanEav Dataset

Human action	Number of Training	Number of Testing	Boxing	Walking	Jogging
Boxing	600	1200	94.2	1.9	3.9
Walking	600	1200	6.7	84.7	8.6
Jogging	600	1200	7.5	10.8	81.7

(a) 3D human action recognition using 3D human model of 128x128x128 dimension

Human action	Number of Training	Number of Testing	Boxing	Walking	Jogging
Boxing	600	1200	92.7	2.2	5.1
Walking	600	1200	5.8	83.8	10.4
Jogging	600	1200	5.1	13.5	81.4

(b) 3D human action recognition using 3D human model of 64x64x64 dimension

Table 5.8: Comparison of human actio recognition using K Nearest Neighbor and single-kernel Support Vector Machine to compare with our proposed MK-SVM based human action recognition

Human action	Training	Testing	Boxing	Walking	Jogging
Boxing	600	1200	86.2	6.8	8.4
Walking	600	1200	9.3	75.8	12.9
Jogging	600	1200	7.8	12.5	79.7

(a) Human action recognition ratio using K Nearest Neighbor classification method.

Human action	Training	Testing	Boxing	Walking	Jogging
Boxing	600	1200	87.4	4.4	7.2
Walking	600	1200	11.6	74.8	13.6
Jogging	600	1200	4.8	14.3	80.9

(b) Human action recognition ratio using Single Kernel Support Vector Machine classification method.

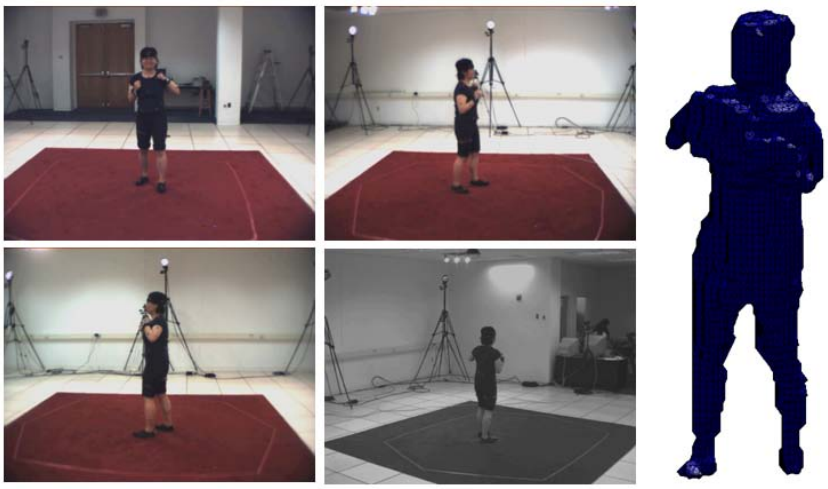


Figure 5.5: Example of error of our proposed human action recognition

Table 5.9: 3D human action recognition matrix in our environment

Human action	testing	Boxing	Walking	Jogging
Boxing	500	83.2	5.4	10.4
Walking	500	13.8	72.7	16.1
Jogging	500	9.5	16.3	74.2

ment whose illumination condition is worse than the HumanEva dataset to see the robustness of our proposed system. As shown in the Figure 5.6, the shadows are recognized the foreground and reconstructed in our procedure. Some human body parts are not also reconstructed because of partial occlusion of the camera. As shown in Table 5.9, the acceptance ratio in our experimental environment is lower than the HumanEva data based human action recognition because we only use 4 cameras and shadow effect in the floor.

Human action recognition ratios of walking and jogging are lower than the acceptance ratio of boxing because the eigen-features of arms and legs of walking and jogging is similar with each other.

Unfortunately, there is no public database for 3D human model based action recognition, so we could not directly compare our approach and others. The state of the art in human action recognition from 2D image based on KTH human action dataset is still not able to correctly recognize the jogging, walking, and running, while our proposed approach using the 3D reconstructed model has a balanced

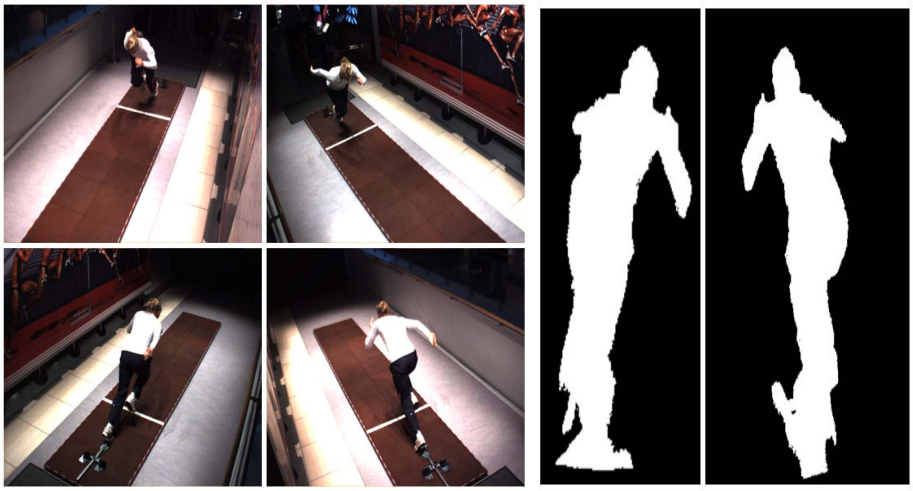


Figure 5.6: Example images for human action recognition in our experimental environment.

human action recognition ratio.

5.3 Query-by-Sketch based Image Retrieval

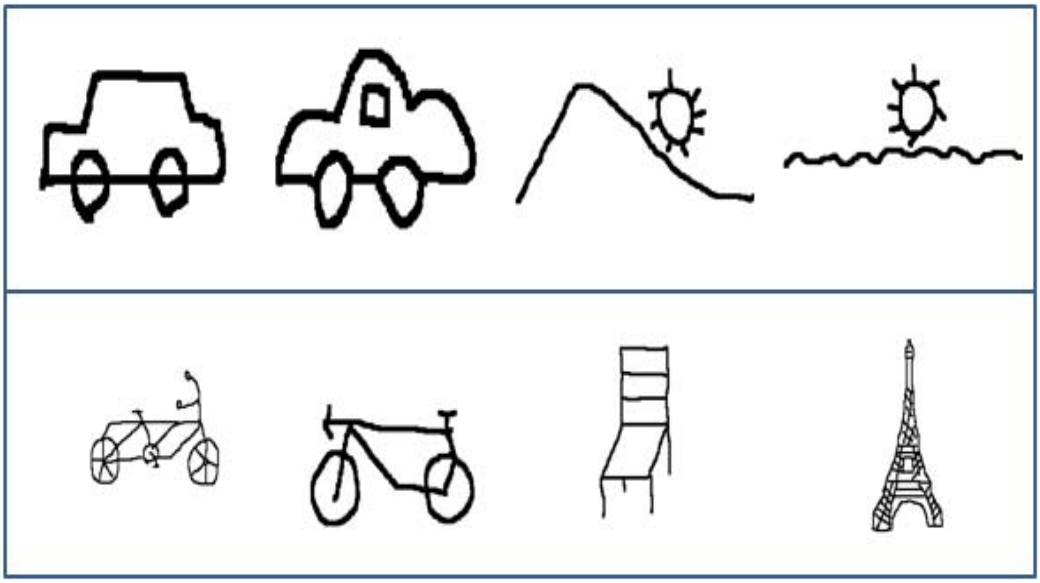
There is a growing interest in the new Human Computer Interaction techniques on tablet PCs, touch phones, and multi-touch screen. Numerous research projects in direct touch systems focus not only on interaction using one or two fingers simultaneously, but also on working together collaboratively in a multi-user scenario [108, 119, 135]. Thus, new ways of generating immersion are possible using simple handling. However, real HCI within these devices is still lacking in terms of understanding freehand input and low level interactive visualization of content. Among numerous interaction methodologies for detecting and retrieving various images in database, freehand sketch, the informal drawing of shape using lines and curves, is one of the most powerful and intuitive tools for Human Computer Interaction because of its familiar in the human visual perception and fast representation of important characteristics of target objects. Although a sketch is composed of few lines, it is a coarse but detailed picture including its key features [89]. For instance, if we ask the people to draw objects like a human, a car, or a cup without any further information, most people intend to sketch the objects as shown in Figure 5.7. These sketches are passive and cannot be directly simulated

or analyzed using computational engineering tools. Computers cannot understand the sketched images which have various viewpoints and non-textual information, even though users can easily understand their characteristics and classify the images into categories. Therefore, the sketched query image needs to be transformed into a computer-suited representation. Then we extract the features which are very close to user intention.

Nevertheless, image retrieval and browsing applications have been developed for users to efficiently search for browse, and order target images in large databases from a personal image library [235] or on the web [151]. Retrieving images that share some similar visual elements from a query image has been one of the most challenging areas in computer vision, named Content-based Image Retrieval (CBIR) [242] to efficiently find the photos which users want to see. CBIR techniques attempt to search for image content within media databases based on meaningful features as well as to measure the dissimilarity of visual objects by distance functions. The capacity of CBIR heavily relies on the definition of similarity measures and an adequate configuration of databases. Standard similarity measures or naive database configuration methodologies are often too generic forcing CBIR into time consuming image search and retrieval or leading to a complete failure. Searching techniques through visual features such as color [6], texture [21, 136], or shape information [106] could lead to more accurate CBIR. Simply sketched image can be used as a query to describe the users' intention in a short time and retrieve the images. Recently, Sketch based Image Retrieval (SBIR) [49, 156] has evolved as one part of Content based Image Retrieval. SBIR were developed to overcome the limitations of previously well known approaches such as keyword or query-by-example based image retrieval methods. Several Sketch based Image Retrieval techniques are based on feature detection and similarity measures. The obvious advantage over other existing methods is its ease of use and handle because the sketched images which are shown in Figure 5.7 are familiar with human visual perception and account for fast representations of important characteristics of the target objects. However, the boundary contours of each target object from different view directions or the information on shapes are needed and have to be prepared during a preprocessing phase.

In this Chapter, we present a query-by-sketch based image retrieval based on understanding of hand drawn sketches which serve as a query input. This system is based on an adequate extraction of features from query and database images as well as measuring their similarities in order to retrieve the most similar objects within the database. Figure 5.8 shows our methodology which we will explain in detail further on. First, the images which are available at different resolutions,

(a) Sketched query image in cell phone and tablet PC



(b) Examples of sketched query images

Figure 5.7: Some sketched images such as ”car”, ”sunset”, ”bicycle”, ”chair”, and ”Eiffel tower”, which are familiar with human visual perception in various tools

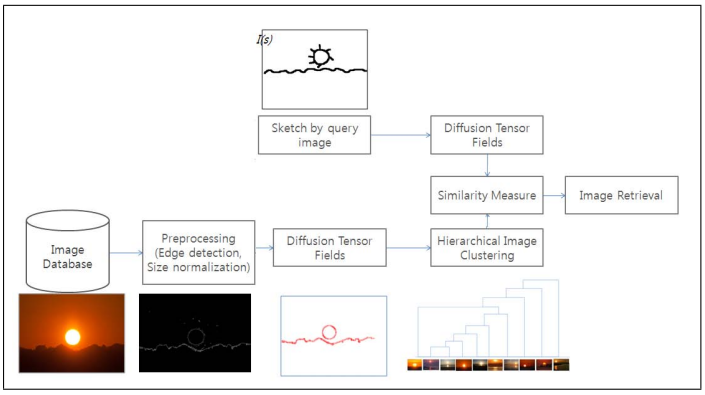


Figure 5.8: Total flowchart of our proposed query-by-sketch based image retrieval. Our proposed SBIR is composed of hierarchical image clustering, tensorial feature extraction, and similarity measure to retrieve the most similar image in database. Preprocessing step of image dataset by Canny edge detection and size normalization, tensorial feature extraction and its analysis are explained, and then similarity measure between a sketched query image and image dataset is described and hierarchical image clustering is explained in detail in the last section.

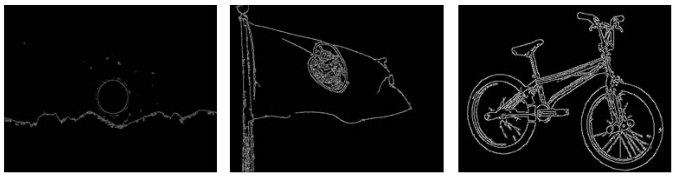
partial occlusions, and change of view within a database are transformed and classified. This will be based on an adequate extraction of edges and an analysis of their tensorial properties. We then extract the eigenvalues and eigenvectors using a novel topological analysis within a given query and the hierarchically classified image dataset. Last but not least, we retrieve the most similar image by introducing the similarity measure and analyzing the characteristics of tensorial features between a sketched query image and the dataset.
Our proposed methodology has the following advantages compared to previous vector fields based skeleton extraction and object retrieval technique:

(1) There is no need to determine a priori-information from images in the database or the sketch query image to detect the target objects.

(2) Our proposed methodology shows an improved skeleton extraction from a sketch image that includes a singular region.

(3) We can easily detect and retrieve the images in a database by its skeletal char-

(a) Original input images in database



(b) Edge extraction

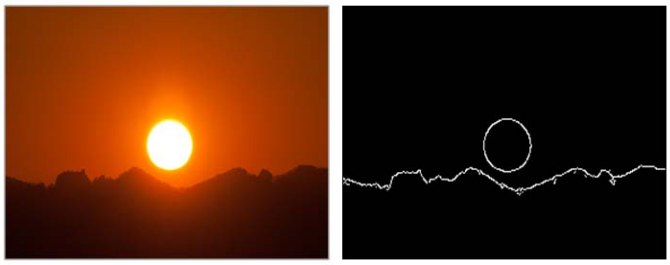
Figure 5.9: Edge extraction of images in database to easily extract the robust and efficient features which are similar to user drawn sketched query images. Canny edge information contains similar cues from user drawn sketch images.

acteristics.

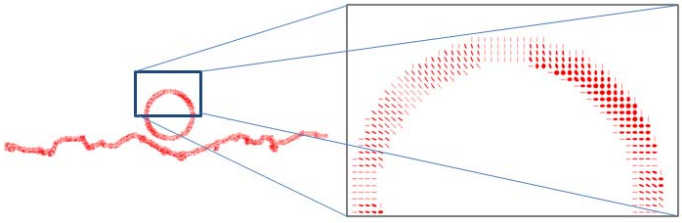
In next Section, we will explain our proposed approach for target object detection and retrieval using query-by-sketch image. It is composed of the three steps; i) sketched query based image transformation, ii) skeletal feature extraction using Normalized Gradient Vector Flow in the space of diffusion tensor fields, and iii) similarity measuring between a query image and an image dataset or target images.

5.3.1 Preprocessing of image of database

For sketch images which are drawn by users it is very difficult to understand their characteristics, because users sometimes omit important features or draw in detail with noisy lines. Before retrieving the image data from a sketched query image, we need to transform the image dataset to be easily understandable from a rough query image, classify the dataset to reduce the retrieval time and find the common features between a query and database images which can reduce the sensory distortion such as view change, illumination, and partial occlusion. Cole et al. tried to identify where people draw the lines of real 3D objects when they are requested to sketch the images. They found that the Canny edge [7] provided the strongest cues for people to draw the lines of the target objects comparing to other computer



(a) Input image in database (b)Edge image



(c) Ellipsoidal expression of each pixel using extracted eigenvalues and eigenvectors. In the right image which is the detail of the ellipsoidal representation using eigenvalues and eigenvectors, we can see how each ellipses' scale and angle are changed by its eigenvalues and eigenvectors.

Figure 5.10: Ellipsoidal expression of each pixel from the image of database using tensorial properties.

graphics line definitions such as apparent ridges, suggestive contours, and ridges and valleys.

This Chapter adopts the assumption that edge information is one of the strongest characteristics, which can provide a link between user drawn sketch and a target image within a database. Therefore, we first transform the database by using size normalization, edge extraction, and binarization methods. Figure 5.9 shows the binarized edge images within a dataset of images which serve as an input to a further preprocessing step trying to identify adequate features for the measurement of similarities.

Figure 5.10 shows the process to extract the tensorial features from a query image and image dataset. We also display each pixel as the ellipsoidal representation of the images using extracted eigenvalues and eigenvectors. The characteristics of tensorial properties is used as feature to retrieve the images in database and we will explain the detail of measuring the similarity within the next section.

5.3.2 Similarity Measure of a Query Image

In Sketch based Image Retrieval, the way to measure the similarity between a query and databases is significant to the overall system performance and might be a crucial bottleneck to retrieve the exact image in databases. Hence, we present a new similarity measure in order to avoid a similarity measure within a scalar space contributing to further computational overhead. We first label the edge image and ignore the labels which have few number of edge elements because it might be a noise in a similarity measure. The edge image, E can be expressed using eigenvalues λ for each label as:

$$E = \begin{bmatrix} \Sigma_{l=1}^{a}\Sigma_{k=1}^{2}\lambda_{lk} & & \\ \Sigma_{l=1}^{b}\Sigma_{k=1}^{2}\lambda_{lk} & & \\ \cdots & \cdots & \cdots \\ \Sigma_{l=1}^{n}\Sigma_{k=1}^{2}\lambda_{lk} & & \end{bmatrix} \tag{5.5}$$

where the number of matrix's column is the number of label, L within an edge image and $a,b,..,n$ are the number of elements within each label. The score matrix between an edge image of a query image, E_i which have L labels, and an edge from databases, E_j which have M labels can be shown like that:

$$Score = \frac{1}{N}\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ \vdots & \cdots & \cdots & \vdots \\ s_{L1} & s_{L2} & \cdots & s_{LM} \end{bmatrix} \tag{5.6}$$

where N is the normalization factor to make the score value between 0 and 1. Since the score using tensorial analysis leads to the proper choice of a similarity measure that can be used, several schemes within the space of tensor fields have been published. Those have been established according the nature of application domain aiming at solving the problems of medical image processing, color image segmentation and so on [61].
Given two tensorial elements, $\mathbf{T}_i$, and $\mathbf{T}_j$ the simplest comparison between two tensor quantities is shown by:

$$d_1(\mathbf{T}_i,\mathbf{T}_j) = \lambda_{1,i}\lambda_{1,j}(e_{1,i}\cdot e_{1,j})^2 + \lambda_{2,i}\lambda_{2,j}(e_{2,i}\cdot e_{2,j})^2 \tag{5.7}$$

where $e_{1,i}$ and $e_{1,j}$ are principal eigenvectors of tensors $\mathbf{T}_i$, and $\mathbf{T}_i$, respectively. The second example is the tensor Euclidean distance obtained by using the Frobe-

nius norm. Due to its simplicity, tensor Euclidean distance has been used extensively in DTI restoration:

$$d_2(\mathbf{T}_i, \mathbf{T}_j) = \sqrt{Trace((\mathbf{T}_i - \mathbf{T}_j)^2)} \tag{5.8}$$

We define our similarity measure as the multiplication of $d_1(\mathbf{T}_i, \mathbf{T}_j)$ and $d_3(\mathbf{T}_i, \mathbf{T}_j)$. We have found that this captures both the difference in scale and in angle of the elliptical tensorial elements [124].
The score matrix element, $s^{ij} = d_1 \cdot d_2$ using tensorial properties is merged to each labeled branch and we can recalculate the similarity measures between each label. The final scoring will be determined by combining the minimum similarity values from each branch.

5.3.3 Hierarchical configuration of image data

In this section, we construct a hierarchical order of layers by using our derived similarity measure based on tensorial features extracted from images within our database. To classify and retrieve an accurate image from a large number of images in database, it is required to reduce the local features from images and learning technique to efficiently training the images using transformed features.
In image clustering applications, unsupervised image clustering can be separated into non-hierarchical and hierarchical clustering algorithms [274]. In numerous non-hierarchical clustering methods, K-Means clustering is an algorithm to cluster n images based on attributes into k partitions, where k is less than n, the number of images, to form a k-block set partition of data. The final goal is to find good local minimum and have linear complexity O(kmin) with respect to the number of instances. Vice versa, hierarchical clustering algorithms are non-iterative and to create a *dendrogram* which is a tree structure containing a k-block set partition for each value of k between 1 and n, n is the number of images at the lowest level, allowing the user to choose a particular clustering granularity. The basic agglomerative hierarchical clustering algorithm begins with each image in its own cluster. Hierarchical clustering and summarization of the image database helps us to reduce the time to search the related images. On the lowest layer of the hierarchy, we extract the tensorial features of the edges images which build the foundation of our system and upper layers of the hierarchical structure are separated by a clustering algorithm, in which the established feature space consists of the tensorial features. Image data clustering and categorization are means for high-level description of image content. Most content based image clustering and summariza-

tion algorithms rely on feature extraction and similarity measures by comparing the visual features like color, texture, shape or text information of images. However, we analyze the tensorial feature of each edge pixel instead, in order to cluster and classify the images for sketched query image based image retrieval.
Let

$$S = \{Score_1, Score_2, ..., Score_{n-1}, Score_n\} \tag{5.9}$$

be the set of images to be clustered, where $Score_i$ is score of the similarity measure which are extracted in the previous section. Initially, the number of clusters is same as the number of images in each cluster, n, and C_i is represented as the cluster within a same category. Then we progressively join the closest clusters through equation 5.10 and 5.11 until k= 1.

$$sm(i,j) = D(C_i, C_j), \forall i, j; \tag{5.10}$$

$$C_l = Join(C_l, C_m); Remove(C_m) \tag{5.11}$$

where where *sm(i,j)* is the similarity measure between cluster C_i and C_j. In this paper, the similarity measure between clusters is calculated by the Euclidean distance, D between clusters.

The object of our hierarchical clustering algorithm is to extract a multi-level partitioning of images and visual features based on tensorial feature's similarity measure, i.e. a partitioning which groups images into a set of clusters and then, recursively, partitions them into smaller sub-clusters, until some break criteria is satisfied. Agglomerative hierarchical clustering algorithms start with several clusters containing only one object, and iteratively two clusters are chosen and merged to form one larger cluster. This process is repeated until only one large cluster is left, which contains all objects. We focused on the agglomerative hierarchical clustering algorithm because the feature basis are tensorial features from the edge image. Figure 5.11 shows the *dendrogram* of one of clusters in database. The dendrogram is constructed by the distance between images in database by using the tensorial similarity measure. Our proposed hierarchical image clustering method is also very efficient to add a new image in a local repository. The new image will traverse down to the lowest layer by measure the similarity measure and make a new hierarchical layers within the cluster.

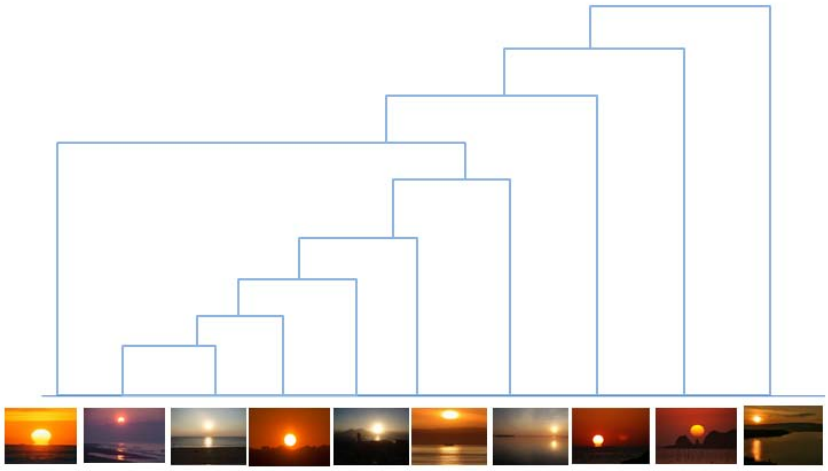


Figure 5.11: Database configuration using aggromerative hierarchical clustering by using tensorial similarity measure. This Figure is the configuration of one of categories as ”sunset” images.

5.3.4 Experiments

We conducted several experiments to retrieve the most similar image within our database by a roughly sketched query image. We have implemented our system on a Pentium 4 1.2GHz and downloaded 600 images from the web containing various illumination environments, view changes and image sizes. The image data is categorized manually into 60 classes such as sunset, Eiffel tower, bicycle, chair, etc. Our edge extraction is based on a realized Canny edge detection technique,

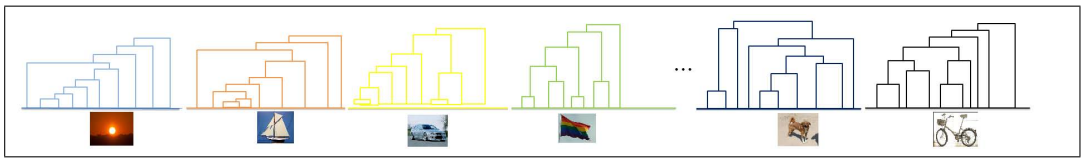


Figure 5.12: Hierarchical clustering method for some clusters in database using our tensorial feature based similarity measure. Images in database are downloaded on the web and separated with 60 clusters such as chair, sunset, cars, bicycle, and Eiffel tower, etc.

Table 5.10: Average running time for sketch-query based image retrieval

module	running time (ms)
Edge detection and size normalization	34
Tensorial feature extraction	26
Image retrieval from database	118

and we normalized the image dataset to 100x100. We also labelled the edge features and eliminated the labels which have a small number of edge pixels in order to reduce the performance errors. The computational time to retrieve the image in database is dependent on the similarity measure between a query image and image dataset because our approach is based on the number of labels of a query image. The average running time to retrieve the images in dataset is shown in the Table 5.10.

The classification of the images within our database was implemented using the presented agglomorative hierarchical clustering method. Some resulting clusters are shown in Figure 5.12.

Figure 5.13 shows the results of top ranked retrieved images based on a sketchy query image (left side). This type of query image is not contained within our database. To retrieve the image, the query image is ”similarity measured” from the top of the hierarchical layer tree, and traversed down to find the most similar image in database. As explained in last section, we also lead the experiments by adding a new image into the established cluster. A new image measures the similarity layer by layer and creates a new position by neighboring similar tensorial features.

We have also tested our proposed method from various users. Figure 5.14 shows the retrieved ”chair” images from 5 different users. When we asked them to draw the chair without any information, the drawn sketch image is different case by case. The score value and order is different according to its similarity measure. Especially, the scores of retrieved image which is used by user 4 are lower than other retrieved top ranked images because the database in the chair class does not have similar pictures.

As shown in Figure 5.13, the sketched images of ”sunset” or ”sunrise” are composed of multiple objects like mountain, see and sun while other query images use single object. Even though our is not dependent on the number of objects within a query image, we mainly find images related to the dominant features in Figure 5.15. From the query image which is composed of ”bike” and ”flag”, we only

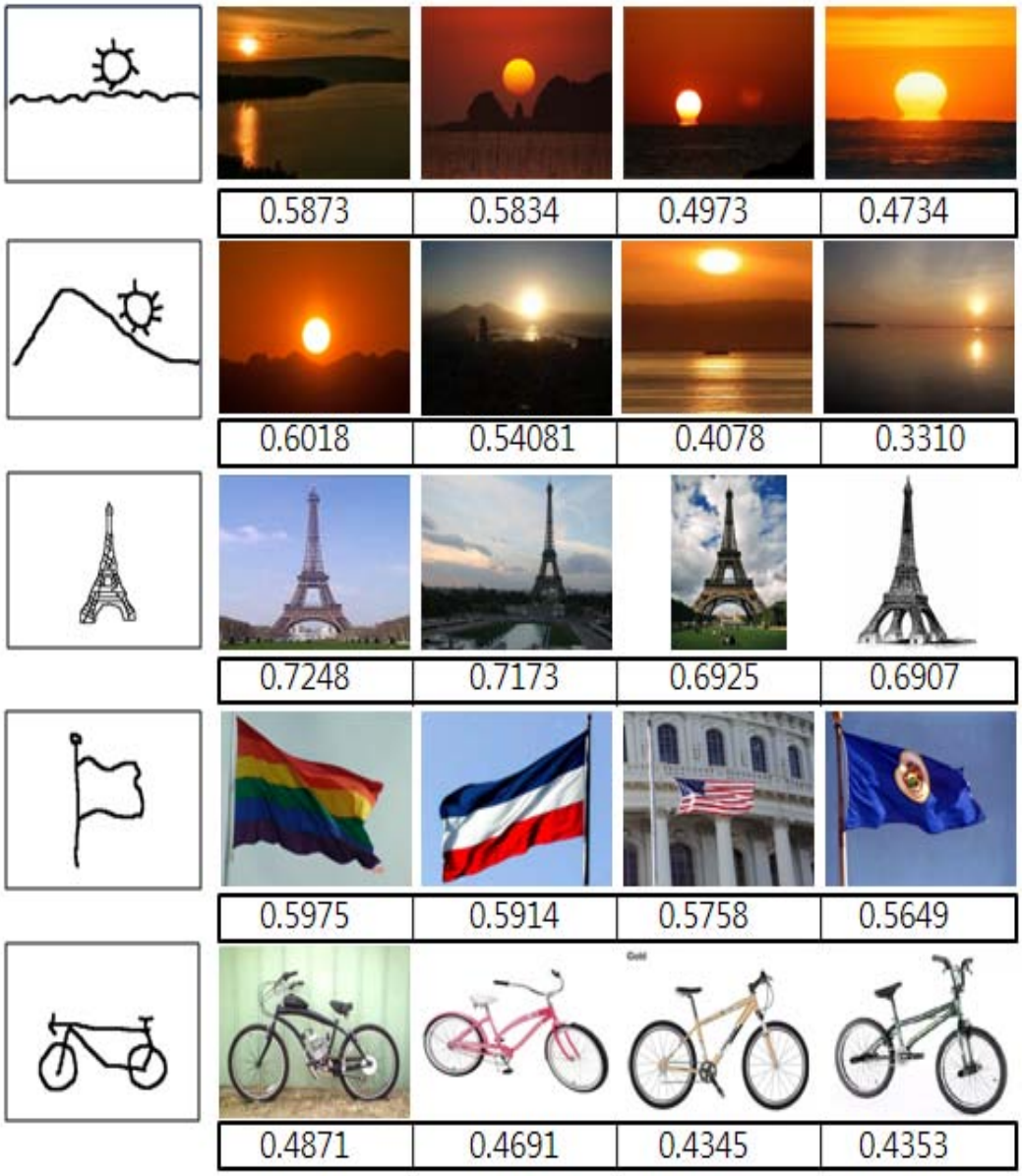


Figure 5.13: Some example of top ranked retrieved images from a query image and its similarity measure between a query image and image dataset.

Figure 5.14: ””Image retrieval from sketched ”chair” image from various users.

Figure 5.15: Top ranked images in databases from a query image which have multiple object.

retrieve and order the images within the category of ”bike”, and can not retrieve the images which contain the flag because the similarity measure between a query image and bike images is higher than the flag images.

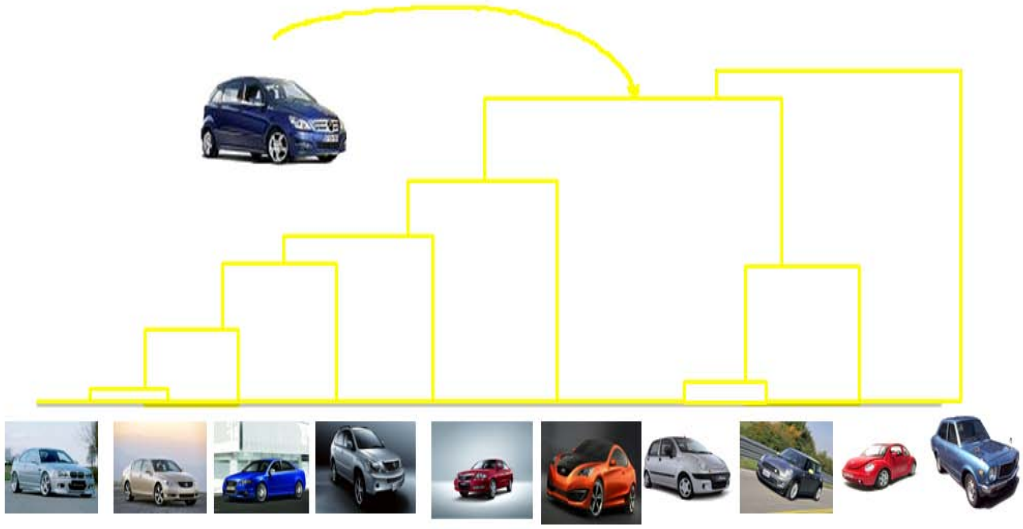
If users take new picture and add this image in the databases, the new image can easily find its location in the hierarchy using our proposed method. Figure 5.16 represents a new clustering configuration method when a new image is added. After adding the new image in our database, the hierarchical clustering structure of each object has new hierarchical layers and retrieve the images including added image.

5.4 Summary

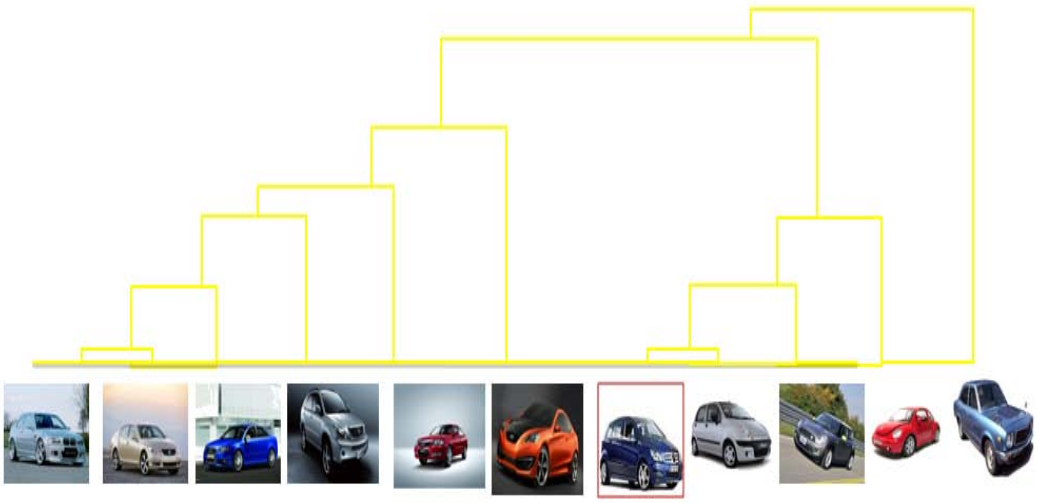
5.4.1 Human action recognition

As one of numerous applications using markerless motion capture and its analysis, human action recognition is a very important issue because we could understand the user's intention by analyzing their actions like boxing, walking, jogging, and so on.

In this chapter, we have presented 2D/3D human action recognition using properties of the segmented human body parts' eigenvalues and eigenvectors. Our system uses eigenvalues and eigenvectors within a 2D image or 3D volume data and is very efficient and robust in partial occlusion and clutter of 3D model as shown in experiments. The human action recognition in 2D images was still affected by camera viewpoint, because the skeletal features from back of users could not extract the typical features of target object, but the 3D reconstructed model from multiple images solved the problems of 2D image based human ac-



(a) New image adding within already established cluster in database



(b) Cluster reconstruction after adding a new image image. Red boxed image is the new position in a cluster within database.

Figure 5.16: Cluster reconstruction using hierarchical image clustering when a new image is added.

tion recognition. As shown in our experimental results, 2D image based on human action recognition was still dependent on viewpoint. However, 3D model based on human action recognition methodology overcame the dependency of camera viewpoint.
3D model segmentation part was very critical in our system in the view of performance related to running time and acceptance ratio. We could focus more on our effort to reducing the processing time for 3D model segmentation. Our system could also be extend to 3D model detection and retrieval.

5.4.2 Sketch-based image retrieval

The query-by-sketch based target object detection and image retrieval is a very efficient method to express users' intension for Human Computer Interaction. The performance of Sketch based Image Retrieval is very dependent on the nature of complex image data, on the extraction of meaningful features from complex images, and on the similarity measure determined by a roughly sketched image.
In this chapter, we have presented our approach for extraction of tensorial features and the measurement of similarities and enhanced hierarchical image classification techniques. The essential idea is based on the analysis of the tensor topology in order to extract the ellipsoidal characteristics of features. We have proven that our hierarchical image clustering methodology is very efficient to retrieve the most similar image from a large repository in a short time and scalable to the addition of new images into the database.
Our proposed sketch based image retrieval will not be limited to 2D image search and retrieval. We will extend our methodology to the understanding of 2D/3D motions of target objects. Especially, we will focus in particular on a 3D structural analysis in the space of tensor fields.

Chapter 6

Conclusion and Discussion

6.1 Conclusion

This dissertation presents a novel approach for markerless motion capture and its analysis for various applications in the field of computer vision and computer graphics. Our proposed markerless motion capture and analysis method did not request any prior information of the target objects which have high-degree of freedom. Even though there was a partial occlusion and a clutter in the background or motion ambiguities within the target objects, we successfully extracted and split the skeleton features by using the eigen-features which came from diffusion tensor fields.

First, we briefly surveyed the historical developments of motion capture and its analysis and heavily studied related works which were researched during the last twenty years in the area of markerless motion capture, photo-realistic 3D reconstruction to build the 3D model from multiple images and skeletal feature extraction and similarity measure using the properties of diffusion tensor fields in Chapters 1 and 2.

Skeletal feature extraction and its splitting methodology within the deformable objects was explained in Chapter 3. The skeleton which was analyzed by using Normalized Gradient Vector Flow and connecting the degenerate points in the space of diffusion tensor fields, was successfully split by measuring the similarity between neighboring skeletal features. Our proposed skeleton extraction and splitting method was very efficient and robust against the noise of the appearance models. We conducted the experiments to show the effectiveness of our proposed method by comparing to previous skeleton extraction methods, measuring the Eu-

clidean distance between the ground truth of joints and extracted features by using our method.
The 3D skeletal feature extraction and splitting method which was analyzed in Chapter 4 was an extension of two-dimensional Normalized Gradient Vector Flow and diffusion tensor fields, but the similarity measure method for splitting the neighboring voxels was different from 2D approach. We have compared our approach to model based human motion estimation using the Pseudo-Zernike Moment method. The Pseudo-Zernike model based human motion analysis method was faster than our proposed method, but it had limitations in visualizing and segmenting the deformable objects in detail. In particular, its technology in the field of computer vision and computer graphics was very dependent on the capacity of computers and cameras. With the help of the parallel calculation using hardware accelerated GPU and high speed cameras, we realized the real-time 3D reconstruction and motion analysis. The running time which was shown in Chapter 4 for photo-realistic 3D model reconstruction showed that our proposed method made it possible to real-time rendering and analysis in an arbitrary viewpoint.
In Chapter 5, we presented the human action recognition and sketch-based target object detection and retrieval to show that our motion analysis method could be applied to diverse areas in computer vision. The applications using our proposed method were very important in the view of understanding the user's intention in multiple camera environment. In human action recognition, the ellipsoidal representation of segmented regions showed how much the subregions were rotated and translated from the origin. These features were used for classifying and learning each human action.

6.2 Discussion

Our system to detect and retrieve the image or 3D model is in essence based on a single-query 2D image /3D model. However, natural phenomena are composed of multiple target objects. There are limitations in retrieving the multiple images using our proposed methodology because we need to separate the unknown 2D/3D model into several target objects and retrieve the models in database. Most 3D object retrieval are based on a single query image or 3D model. In a hierarchical clustering methodology, there are no problems to retrieve the most similar 3D models in database because we cannot access the other object classes when the hierarchy of target object is decided.

The various topics in computer graphics and computer vision are going to understand the user's behaviors and his/her intentions by analyzing their motion. Even though numerous novel techniques are invented by many researchers in this area, there is still a gap between an advanced visualization and analysis tool, and actual acceptance of engineers and scientists for whom the work is intended. Some pipelines of our proposed system are already displayed in the museum[1] to interactively analyze the user's motion. We need to complement our system for usage in diverse environments without the help of skilled engineers and scientists. In particular, the medical applications using motion analysis are very critical because they are very closely related to human's life. So we need to carefully analyze and visualize the motion not to miss the important information.

[1] Heinz Nixdorf Museum Form, Paderborn, GERMANY

Bibliography

[1] Marker based tracking systems, 2005.

[2] L. N. Abdullah and S. A. M. Noah. Integrating audio visual data for human action detection, 2008.

[3] A. Agarwal and B. B. Triggs. 3d human pose from silhouettes by relevance vector regression, 2004.

[4] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images, 2006.

[5] P. Ahammad, C. Yeo, S. S. Sastry, and K. Ramchandran. Compressed domain real-time action recognition, 2006.

[6] S. Aksoy and R. M. Haralick. Content based image database retrieval using variances of gray level spatial dependencies, 1998.

[7] D. Alexander, J. Gee, and R. Bajcsy. Similarity measures for matching diffusion tensor images, 1999.

[8] B. Allen, B. Curless, and Z. Popovic. The space of human body shapes: reconstruction and parameterization from range scans, 2005.

[9] D. Anguelov, P. Srinivsan, D D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people, 2005.

[10] M. Ankerst, G. Kastenmuller, H. P. Kriegel, and T. Seidl. 3d shape histograms for similarity search and classification in spatial databases, 1999.

[11] M. Attene, B. Falcidieno, and M. Spagnuolo. Hierarchical mesh segmentation based on fitting primitives, 2006.

[12] M. Attene, B. Falcidieno, and M. Spagnuolo. Hierarchical mesh segmentation bsed on fitting primitives, 2006.

[13] D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment, 2001.

[14] S. R. Aylward and E. Bullitt. Initialization, noise, singularities and scale in height ridge traversal for tubular object centerline extraction, 2002.

[15] A. H. Barr. Superquadrics and angle preserving transformations, 1981.

[16] P. Basser, J. Mattiello, and D. LeBihan. Estimation of the effective self-diffusion tensor from the nmr spin echo, 1994.

[17] A. Bazzani. An svm classifier to separate false signals from microcalcifications in digital mammograms, 2001.

[18] S. O. Belkasim, M. Ahmadi, and M. Shridhar. Efficient algorithm for fast computation of zernike moments, 1996.

[19] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts, 2002.

[20] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts, 2002.

[21] A. Del Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches, 1997.

[22] S. Bischoff and L. Kobbelt. Elliposoidal decomposition of 3d model, 2002.

[23] I. Bitter, A. E. Kaufman, and M. Sato. Penalized-distance volumetric skeleton algorithm, 2001.

[24] M. Blank, L. Gorelick, E. Shechtman, M M. Irani, and R. Basri. Actions as space-time shapes, 2005.

[25] H. Blum. A transformation for extracting new descritions of shape, 2003.

[26] A. Bobick and J. Davis. The recognition of human movement using temporal templates, 2001.

[27] G. Borgefors, I. Nystrom, and G. Sanniti di Baja. Computing skeletons in three dimensional, 1999.

[28] S. Bouix, K. Siddiqi, and A. Tannenbaum. Flux driven fly throughs, 2003.

[29] B. Boulay, F. Bremond, and M. Thonnat. Posture recognition with a 3d human model, 2005.

[30] R. Bowden, T. Mitchell, and M. Sarhadi. Reconstructing 3d pose and motion from a single camera view, 1998.

[31] M. Bray, P. Kohli, and P. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humand using dynamic graph-cuts, 2006.

[32] C. Bregler and J. Malik. Tracking people with twists and exponential maps, 1998.

[33] C. Bregler and J. Malik. Tracking people with twists and exponential maps, 1998.

[34] A. Brennecke and T. Isenberg. 3d shape matching using skeleton graph, 2004.

[35] C. Buehler, M. Bosse, L. McMillan, S. J. Gortler, and M. F. Cohen. Unstructured lumigraph rendering, 2001.

[36] T. W. Calvert, J. Chapman, and A. Patla. Aspects of the kinematic simulation of human movement, 1982.

[37] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Freeviewpoint video of human actors, 2003.

[38] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors, 2003.

[39] J. E. Chadwick, D. R. Haumann, and R. E. Parent. Layered construction for deformable animated characters, 1989.

[40] J. H. Challis. A procedure for determining rigid body transformation parameters, 1995.

[41] A. Charlechale, G. Naghdy, and A. Mertins. Sketch-based image matching using angular partitioning, 2005.

[42] D. Chen and A. A. Farag. Detecting critical points of skeletons using triangle decomposition of gradient vector flow field, 2005.

[43] L. Chevalier, F. Jaillet, and A. Baskurt. Segmentation and superquadric modeling of 3d objects export, 2003.

[44] J. Choi, Y.-I. Cho, K. Cho, S.-J. Bae, and H. Y. Yang. A view-based multiple objects tracking and human action recognition for interactive virtual environments, 2008.

[45] J. Choi, Y.-I. Cho, K. Cho, T. Han, and H. Y. Yang. A view-based real-time human action recognition system as an interface for human computer interaction, 2008.

[46] J. Christopher and C. Burge. A tutorial on support vector machines for pattern recognition, 1998.

[47] C.-W. Chu, O. J. Jenkins, and M. J. Mataric. Markerless kinematic model and motion capture from volume sequences, 2003.

[48] J.-H. Chuang, C.-H. Tsai, and M.-C. Ko. Skeletonization of three-dimensional object using generalized potential field, 2000.

[49] F. Cole, A. Golovinskiy, A. Limpaecher, H. Scoddart, S. Barros, A. Finkelstein, T. Funkhouser, and S. Rusinkiewicz. Where do people draw lines, 2008.

[50] S. Corazza, L. Mundermann, A. Chaudhari, T. Demattio, C. Cobelli, and T. Andriacchi. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach, 2006.

[51] N. Cornea, D. Silver, X. Yuan, and R. Balasubramanian. Computing hierarchical curve-skeletons of 3d objects, 2005.

[52] M. Cote, P. Payeur, and G. Comeau. Video segmentation for markerless motion capture in unconstrained environments, 2007.

[53] N. Cristianini and J. Shawe-Taylor. Kernel methods for pattern analysis, 2004.

[54] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams, 2003.

[55] Y. Cui and J. J. Weng. Hand segmentation using learning-based prediction and verification for hand sign recognition, 1997.

[56] D D. Ebert and C. Shaw. Minimally immersive flow visualization, 2000.

[57] S. Danafar, Gheissari, and Niloofar. Action recognition for surveillance applications using optic flow and svm, 2007.

[58] C. Davatzikos, F. Abraham, G. Biros, and R. Verma. Correspondence detection in diffusion tensor images, 2006.

[59] A. J. Davison, J. Deutscher, and I. D. Reid. Markerless motion capture of complex full-body movement for character animation, 2001.

[60] P. E. Debevec, G. Borshukov, and Y. Yu. Efficient view-dependent image-based rendering with projective texture-mapping, 1998.

[61] T. Delmarcelle and L. Hesselink. The topology of symmetric, second-order tensor fields, 1994.

[62] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering, 2000.

[63] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search, 2005.

[64] P. Dollar, V. Rabaud, G. Cottrel, and S. Belongie. Behavior recognition via spatio-temporal features, 2005.

[65] G. Doretto, D. Cremers, P. Favaro, and S. Soatto. Dynamic texture segmentation, 2003.

[66] T. W. Drummond and R. Cipolla. Real-time tracking of complex structures for visual servoing, 2000.

[67] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance, 2003.

[68] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing actions at a distance, 2003.

[69] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization, 2003.

[70] A. Elgammal, A. Elgammal, and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning, 2004.

[71] A. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction, 1999.

[72] S. Engel and J. Rubin. Detecting visual motion boundaries, 1986.

[73] O. Faugeras. Three-dimensional computer vision: A geometric approach, 1996.

[74] M. Federico, D. Giordani, and P. Coletti. Development and evaluation of an italian broadcast news corpu, 2000.

[75] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the qbic system, 1995.

[76] W. Freeman. Computer vision for computer games, 1996.

[77] P. Fua, R. Plankers, and D. Thalmann. Tracking and modeling people in video sequences, 2001.

[78] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3d models, 2003.

[79] T. Funkhouser, P. Min, M. Kazhdan, J J. Chen, A. Halderman, and D. Dobkin. A search engine for 3d models, 2003.

[80] D. Gavrila and L. Davis. Towards 3-d model-based tracking and recognition of human movement, 1995.

[81] D. M. Gavrila. The visual analysis of human movement: A survey, 1999.

[82] Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes, 1999.

[83] C. M. Ginsberg and D. Maxwell. Graphical marionette, 1983.

[84] A. Golovinsky and T. Funkhouser. Randomized cuts for 3d mesh analysis, 2008.

[85] Y. Gong, T. S. Lim, H. C. Chua, H. J. Zhang, and M. Sakauchi. Automatic parsing of tv soccer programs, 1995.

[86] L. Gorelick, M. Brank, E. Shechtman, M M. Irani, and R. Basri. Actions as space-time shapes for human motion analysis, 2004.

[87] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model, 2003.

[88] A. Gupta and R. Jain. Visual information retrieval, 1997.

[89] T. Hammond, B. Eoff, B. Paulson, A. Wolin, K. Dahmen, J. Johnston, and P. Rajan. Free-sketch recognition: Putting the chi in sketching, 2008.

[90] B. Han, D. Comaniciu, and L. S. Davis. Sequential kernel density approximation through mode propagation: applications to background modeling, 2004.

[91] R. Hartley and A. Zisserman. Multiple view geometry in computer vision, 2000.

[92] M. S. Hassouna and A. A. Farag. Robust centerline extraction framework using level sets, 2005.

[93] L. Herda, R. Urtasun, and P. Fua. Implicit surface joint limits to constrain video-based motion capture, 2004.

[94] J. Hoey and J. Little. Representation and recognition of complex human motion, 2000.

[95] S. Hou and K. Ramani. Sketch-based 3d engineering part class browsing and retrieval, 2006.

[96] S. Hou and K. Ramani. Classifier combination for sketch-based 3d part retrieval, 2007.

[97] Y. B. Hou and Y. Xiao. Active snake algorithm on the edge detection for gallstone ultrasound images, 2008.

[98] N. R. Howe, M. E. Leventon, and W. T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video, 2000.

[99] E. W. Hsu, A. L. Muzikant, S. A. Matulevicius, R. C. Penland, and C. S. Henriquez. Magnetic resonance myocardial fiber-orientation mapping with direct histological correlation, 1998.

[100] M. Hu, Y. Chen, and J. T.-Y. Kwok. Building sparse multiple-kernel svm classifiers, 2009.

[101] J. Huang, S. Ravi Kumar, M. Mitra ans W.-J. Zhu, and R. Zabih. Content-based image indexing and searching using daubechies wavelets, 1999.

[102] S. Intille and A. Bobick. Representation and visual recognition of complex, multi-agent actions using belief networks. Technical Report 454, Perceptual Computing Section, MIT Media Lab, 1998.

[103] C. Y. Ip, D. Lapadat, L. Sieger, and W. C. Regli. Using shape distributions to compare solid models, 2002.

[104] C. Y. Ip, L. Sieger, W. C. Regli, and A. Shokoufandeh. Automated learning of model classifications, proceedingss of solid modeling, 2003.

[105] H. H. S. Ip, A. K. Y. Cheng, W. Y. F. Wong, and J. Feng. Affine invariant sketch based retrieval of images, 2001.

[106] H. H. S. Ip, A. K. Y. Cheng, W. Y. F. Wong, and J. Feng. Affine invariant sketch based retrieval of images, 2001.

[107] M. Irani, P. Anandan, and M. Cohen. Direct recovery of planar-parallax from multiple frames, 2002.

[108] J.Y. J. Y. Han. Multi-touch interaction wall, 2006.

[109] B. Jaehne and H. Hauecker. Computer vision and applications, a guide for students and practitioners, 2000.

[110] A. Jain and F. Farrokhia. Unsupervised texture segmentation using gabor filters, 1990.

[111] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction, 2007.

[112] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo, 2001.

[113] L. B. Kara and K. Shimada. Construction and modification of 3d geometry using a sketch-based interface, 2006.

[114] M. Kass, A. Witkin, and D. Terzopoulos. Snakes : active contour models, 1988.

[115] S. Katz, G. Leifman, and A. Tal. Mesh segmentation using feature point and core extraction from 3d model.

[116] S. Katz, G. Leifman, and A. Tal. Mesh segmentation using feature point and core extraction, 2005.

[117] J. Kautz and H.-P. Seidel. Hardware accelerated displacement mapping for image based rendering, 2001.

[118] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors, 2003.

[119] L. Kim, J. Park, H. Kim, and C. Lee. Hci using multi-touch tabletop display, 2007.

[120] E. C. Kintner. The mathematical properties of the zernike polynomials, 1976.

[121] J. Kniss, G. Kindlmann, and C. Hansen. Multidimensional transfer functions for interactive volume rendering, 2002.

[122] M. Kohle, D. Merkl, and J. Kastner. Human walking: tracking and analysis, 1999.

[123] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russel. Towards robust automatic traffic scene analysis in real-time, 1994.

[124] S. Krishnamachari and M. Abdel-Mottaleb. Hierarchical clustering algorithm for fast image retrieval, 1999.

[125] Y.-K. Lai, S.-M. Hu, R. R. Martin, and P. L. Rosin. Fast mesh segmentation using random walks, 2008.

[126] C.-H. Lee, J.-L. Shih, and J. T. Wang. Object retrieval system based on grid information, 2005.

[127] M.-W. Lee and I. Cohen. Human upper body pose estimation in static images, 2003.

[128] Y. C. Lee, J. R. Parent, and R. Machiraju. Markerless monocular motion capture using image features and physical constraints, 2005.

[129] A. Leonrdis, A. Jaklic, and F. Solina. Superquadrics for segmenting and modeling range data, 1997.

[130] F. Leymarie and M. D. Levine. Tracking deformable objects in the plane using an active contour model, 1993.

[131] B. Li and H. Johan. View context: A 3d model feature for retrieval, 2009.

[132] M. Li, M. Magnor, and H.-P. Seidel. Hardware-accelerated visual hull reconstruction and rendering, 2003.

[133] Y. Li, S. Ma, and H. Lu. Human posture recognition using multi-scale morphological method and kalman motion estimation, 1998.

[134] J. M. Lien, J. Kkeyser, and N. M. Amato. Simultanious shape decomposition and skeletonization, 2006.

[135] J. H. Lim, J. Li, P. Mulher, and Q. Tian. Content-based summarization for personal image library, 2002.

[136] H.-C. Lin, L.-L. Wang, and S.-N. Yang. Regular-texture image retrieval based on texture-primitive extraction, 1999.

[137] H. Y.S. Lin, H.-Y. M. Liao, and J.-C. Lin. Visual salience-guided mesh decomposition, 2007.

[138] W. Lin, M. T. Sun, R. Poovandran, and Z. Zhang. Activity recognition using a combination of category components and local models for video surveillance, 2008.

[139] W. Lin, M. T. Sun, Poovendran, and Z. Zhang. Human activity recognition for video surveillance, 2008.

[140] C. Liu, R. Bammer, B. Acar, and M. Moseley. Characterizing non-gaussian diffusion by using generalized diffusion tensors, 2004.

[141] P. C. Liu, P. C. Wu, W. C. Ma, R. H. Liang, and M. Ouhyoung. Automatic animation skeleton construction using repulsive force field, 2003.

[142] R. Liu and H. Zhang. Segmentation of 3d meshes through spectral clustering, 2004.

[143] T. Liu and D. Geiger. Approximate tree matching and shape similarity, 1999.

[144] B. P. L. Lo and S. A. Velastin. Automatic congestion detection system for underground platforms, 2000.

[145] B. Lok. Online model reconstruction for interactive virtual environments, 2001.

[146] D. G. Lowe. Object recognition from local scale-invariant features, 1999.

[147] W. Ma and B. Manjunath. A toolbox for navigating large image databases, 1997.

[148] W.-C. Ma, F.-C. Wu, and M. Ouhyoung. Skeleton extraction of 3d objects with radial basis functions, 2003.

[149] N. Magnenat-Thalmann, H. Seo, and F. Cordier. Automatic modeling of virtual humans and body clothing, 2004.

[150] P. Maragos and R. Schafer. Morphological skeleton representation and coding of binary images, 1986.

[151] R. Maree, P. Geurts, and L. Wehenkel. Content-based image retrieval by indexing random subwindows with randomized tree, 2007.

[152] D.-C. Marr and T. Poggio. A computational theory of human stereo vision, 1979.

[153] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, 2001.

[154] A. Martinez-Uso, F. Pla Pedro, and Garcia-Sevilla. Unsupervised image segmentation using a hierarchical clustering selection process, 2006.

[155] S. Matusiak, M. Daoudi, T. Blu, and O. Avaro. Sketch based images database retrieval, 1998.

[156] S. Matusiak, M. Daoudi, T. Blu, and O. Avaro. Sketch based images database retrieval, 1998.

[157] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image based visual hulls, 2000.

[158] S. Maybank and T. Tan. Introduction to special section on visual surveillance, 2000.

[159] D. Meyer, J. Denzler, and H. Niemann. Model based extraction of articulated objects in image sequences for gait analysis, 1997.

[160] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data, 2003.

[161] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel datas, 2003.

[162] D. Minnen, I. Essa, and T. Starner. Expectation grammars: Leveraging high-level expectations for activity recognition, 2003.

[163] A. Mittal, L. Zhao, and L. S. Davis. Human body pose estimation using silhouette shape analysis, 2003.

[164] H. M.Lakany, G. M.Haycs, M. Hazlewood, and S. J. Hillman. Clinical gait analysis by neural networks: issues and experiences, 1997.

[165] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture, 2001.

[166] S. Moezzi, A. Katkere, Don Y. Kuramura, and R. Jain. Reality modeling and visualization from multiple video sequences, 1996.

[167] J. M. Morel and S. Solimini. Variational methods for image segmentation, 1995.

[168] M. Mortara, G. Patane, M. Spagnuolo, B. Falcidieno, and J. Rossignac. Plumber: a method for a multi-scale decomposition of 3d shapes into tubular primitives and bodies, 2004.

[169] S. Mukherjea, K. Hirata, and Y. Hara. Amore: A world wide web image retrieval engine, 1999.

[170] R. Mukundan and K. R. Ramakrishnan. Fast computation of legendre and zernike momentss, 1995.

[171] A. Murching, E. Paquet, T. Naveen, A. Tabatabai, and M. Rioux. Description of shape information for 3-d objects, 2000.

[172] L. Nan, S. Dettmer, and M. Shah. Visually recognizing speech using eigen sequences, 1997.

[173] J. C. Niebles, H. Wang, and F.-F. Li. Unsupervised learning of human action categories using spatial-temporalwords, 2006.

[174] M. Novotni and R. Klein. Shape retrieval using 3d zernike descriptors, 2003.

[175] J. F. Obrien, R. E. Bodenheimer, G. J. Brostow, and J. K. Hodgins. Automatic joint parameter estimation from magnetic motion capture data, 2000.

[176] K. Ogawara, X. Li, and K. Ikeuchi. Marker-less human motion estimation using articulated deformable model, 2007.

[177] R. Ogniewicz and M Ilg. Voronoi skeletons: Theory and applications, 1992.

[178] R. Ohbuchi, T. Minamitani, and T. Takei. Shape-similarity search of 3d models by using enhanced shape functions, 2003.

[179] M. Okutomi and T. Kanade. A multiple-baseline stereo, 1993.

[180] N. M. Oliver, B. Rosario, and A. P. Pentland. A bayesian computer vision system for modeling human interactions, 2000.

[181] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions, 2002.

[182] P. Over, C. Leung, H. IP, and M. Grubinger. Multimedia retrieval benchmarks, 2004.

[183] S. Pajevic and C. Pierpaoli. Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: Application to white matter fiber track mapping in the human brain, 1999.

[184] A. Papadopoulos, D. I. Fotiadis, and A. Likas. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines, 2004.

[185] E. Paquet, A. Murching, T. Naveen, A. Tabatabai, and M. Rioux. Description of shape information for 2-d and 3-d objects, 2000.

[186] Parameswaran and V. R. Chellappa. View invariants for human action recognition, 2003.

[187] M. J. Park, M. G. Choi, and S. Y. Shin. Human motion reconstruction from inter-frame feature correspondences of a single video stream using a motion library, 2002.

[188] M. Pelillo, K. Siddiqi, and S. W. Zucker. Matching hierarchical structures using association graphs, 1999.

[189] A. Pendland. Recognition by parts, 1987.

[190] A. Pendland, R. Picard, and S. Scaroff. Photobook: Tools for content-based manipulation of image databases, 1994.

[191] R. Plankers and P. Fua. Articulated soft objects for multiview shape and motion capture, 2003.

[192] R. Polana and R. Nelson. Recognition of motion from temporal texture, 1992.

[193] A. Prata and W. V. T. Rusch. Algorithm for computation of zernike polynomials expansion coefficients, 1989.

[194] C. Pudney. Distance-ordered homotopic thinning: a skeletonization algorithm for 3d digital images, 1998.

[195] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions, 2002.

[196] F. Remondino, N. DApuzzo, G. Schrotter, and A. Roditakis. Markerless motion capture from single or multicamera video sequence, 2004.

[197] F. Remondino, N. Dpuzzo, G. Schrotter, and A. Roditakis. Markerless motion capture from single or multicamera video sequence, 2004.

[198] H. Ren and G. Xu. Human action recognition in smart classroom, 2002.

[199] L. Reveret, L. Favreau, C. Depraz, and M. P. Cani. Morphable model of quadrupeds skeletons for animating 3d animals, 2005.

[200] M. Ringer and J. Lasenby. A procedure for automatically estimating model parameters in optical motion capture, 2002.

[201] B. Robertson. Mike, the talking head, 1988.

[202] B. Robertson. Moving pictures, 1992.

[203] K. Rohr. Human movement analysis based on explicit motion models, 1997.

[204] B. Rosenhahn, T. Brox, U. Kersting, D. Smith, J. Gurney, and R. Klette. A system for marker-less human motion estimation. Kunstliche Intelligenz, 2006.

[205] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking, 2007.

[206] B. Rosenhahn and R. Klette. Geometric algebra for pose estimation and surface morphing in human motion estimation, 2004.

[207] B. Rosenhahn, R.Klette, and D. Metaxas, editors. *Human Motion: Understanding, Modelling, Capture, and Animation*, volume 36 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin Heidelberg, 2008.

[208] J. Rubin and W. Richards. Boundaries of visualization, 1985.

[209] B. Scholkopf and A. Smola. Learning with kernels: Support vector machines, regularization, optimization, and beyond (adaptive computation and machine learning, 2002.

[210] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach, 2004.

[211] D. F. Scollan, A. Holmes, R. Winslow, and J. Forder. Histological validation of myocardial microstructure obtained from diffusion tensor magnetic resonance imaging, 1998.

[212] T. B. Sebastian and B. B. Kimia. Curves vs skeletons in object recognition, 2001.

[213] T. B. Sebastian and B. B. Kimia. Curves vs. skeletons in object recognition, 2005.

[214] T. B. Sebastian, P. N. Klein, and B. B. Kimia. Alignment-based recognition of shape outlines, 2001.

[215] T. B. Sebastian, P. N. Klein, and B. B. Kimia. Recognition of shapes by editing shock graphs, 2001.

[216] M. Segal, C. Korobkin, R. van Widenfelt, J. Foran, and P. Haeberli. Fast shadows and lighting effects using texture mapping, 1992.

[217] J. Segen and S. Kumar. Shadow gestures: 3d hand pose estimation using a single camera, 1999.

[218] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring, 1997.

[219] M. Seki, T. Wada, H. Fujiwara, and K. Sumi. Background detection based on the cooccurrence of image variations, 2003.

[220] H. Seo and N. Magnenat-Thalmann. An examplebased approach to human body manipulation, 2004.

[221] H. Seo and N. Magnenat Thalmann. An automatic modeling of human bodies from sizing parameters, 2003.

[222] H. Seo and N. Magnenat Thalmann. An example-based approach to human body manipulation, 2004.

[223] S. Shakhnarovish, Darrell, and Indyk. Nearest-neighbor methods in learning and vision, 2005.

[224] A. Sharmir. Segmentation and shape extraction of 3d boundary meshes, 2006.

[225] D. Sharvit, J. Chan, H. Tek, and B. B. Kimia. Symmetry-based indexing of image databases, 1998.

[226] C. Shaw, D. Ebert, J. Kukla, A. Zwa, I. Soboroff, and D. Roberts. Data visualization using automatic, perceptually-motivated shapes, 1998.

[227] C. Shaw, C. Hall, C. Blahut, D. Ebert, and D. Roberts. Using shape to visualize multivariate data, 1999.

[228] E. Shechtman and M. Irani. Space-time behavioral correlation, 2005.

[229] Y. Sheikh and M. Shah. Exploring the space of an action for human action recognition, 2005.

[230] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa. Propagation networks for recognition of partially ordered sequential actions, 2004.

[231] J.-L. Shih, C.-H. Lee, and J. T. Wang. 3d object retrieval system based on grid d2, 2005.

[232] S. Shlafman, A. Tal, and S. Katz, editors. *Metamorphosis of polyhedral surfaces using decomposition*.

[233] K. Siddiqi, A. Shokoufandeh, S. J. Dickinson, and S. W. Zucker. Shock graphs and shape matching, 1999.

[234] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion, 2000.

[235] I. Simon, N. Snavely, and S. M. Saitz. Scene summarization for online image collections, 2007.

[236] C. Sminchisescu and Triggs B. Kinematic jump processes for monocular 3d human tracking, 2003.

[237] J. Smith and S.-F. Chang. Visualseek: A fully automated content-based image query system, 1997.

[238] J. J. Spiegelman and S. L.-Y. Woo. A rigid-body method for finding centers of rotation and angular displacements of planar joint motion, 1987.

[239] C. Stauffer and W. E. L. Grimson. Adaptive background mixture modelsfor real-time tracking, 1999.

[240] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking, 2000.

[241] I. Steinwart and A. Christmann. Support vector machines, 2008.

[242] M. Stricker and M. Orengo. Similarity of color images, 1995.

[243] J. Sullivan and S. Carlsson. Recognizing and tracking human action, 2002.

[244] H. Sunder, D. Silver, N. Gagnavi, and S. Dickinson. Skeleton based shape matching and retrieval?, 2003.

[245] R. Szeliski. Rapid octree construction from image sequences, 1993.

[246] M. Takashi and T. Takeshi. Generation, visualization, and editing of 3d video, 2002.

[247] S. Tari and J. Shah. Local symmetries of shapes in arbitrary dimension, 1998.

[248] D. G. Taylor and M. C. Bushell. The spatial mapping of translational diffusion coefficients by the nmr imaging technique, 1985.

[249] A. Telea and A. Vilanova. A robust level-set algorithm for centerline extraction, 2003.

[250] D. Terzopoulos and K. Fleischer. Deformable models, 1988.

[251] D. Terzopoulos and R. Szeliski. Tracking with kalman snakes, 1992.

[252] J. Tierny, J.-P. Vandeborre, and M. Daoudi. Fast and precise kinematic skeleton extraction of 3d dynamic meshesn, 2008.

[253] M. Tonga, Y. Liua, and T. S. Huangb. 3d human model and joint parameter estimation from monocular image, 2007.

[254] R. Y. Tsai. A versatile camera calibration technique for 3d machine vision, 1987.

[255] R. Urtasun and P. Fua. 3d human body tracking using deterministic temporal motion models, 2004.

[256] L. Vacchetti, V. Lepetit, G. Papagiannakis, M. Ponder, P. Fua, N. Magnenat-Thalmann, and D. Thalmann. Stable real-time interaction between virtual humans and real scenes, 2003.

[257] F. E. Veldpaus, H. J. Woltring, and L. J. M. G. Dortmans. A least-squares algorithm for the equiform transformation from spatial marker co-ordinates, 1988.

[258] F. E. Veldpaus, H. J. Woltring, and L. J. M. G. Dortmans. Skeletal parameter estimation from optical motion capture data, 2005.

[259] E. Vendrovsky and I. Neulander. Markerless facial motion capture using texture extraction and nonlinear optimization, 2006.

[260] D. V. Vranic, D. Saupe, and J. Richter. Tools for 3d-object retrieval: Karhunen?loeve transform and spherical harmonics, 2001.

[261] L. Wade and R. E. Parent. Automaticed gneration of control skeleton for use in animation, 2002.

[262] J. Wang, G. Wiederhold, O. Firschin, and S. R. Wei. Content-based image indexing and searching using daubechies wavelets, 1998.

[263] R. Y. Wang, K. Pulli, and J. Popovic. Real-time enveloping with rotational regression, 2007.

[264] W. Wapnik. Statistical learning theory, 1998.

[265] O. Weber, O. Sorkine, Y. Lipman, and C. Gotsman. Context-aware skeletal shape deformation, 2007.

[266] L. Wei and Y. Yang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications, 2005.

[267] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes, 2006.

[268] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder:real-time tracking of the human body, 1997.

[269] J. Wu and M. Levine. Structure recovery via hybrid variational surface approximation, 2005.

[270] L. Xie, P. Xu, S. F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden markov models, 2004.

[271] M.-H. Yang and N. Ahuja. Recognizing hand gesture using motion trajectories, 1999.

[272] A. Yilmaz and M. Shah. Action sketch: A novel action representation, 2005.

[273] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras, 2005.

[274] S. M. Yoon and H. Graf. Hierarchical online image representation based on 3d camera geometry, 2009.

[275] L. Younes. Computable elastic distance between shapes, 1998.

[276] M. Yu, I. Atmosukarto, W. K. Leow, Z Z. Huang, and R. Xu. 3d model retrieval with morphing-based geometric and topological feature maps, 2003.

[277] Z Z. Yu and R. Bajaj. Normalized gradient vector diffusion and image segmentation, 2002.

[278] L. Zelnik-Manor and M. Irani. Event-based analysis of video, 2001.

[279] C. Zhang and T. Chen. Efficient feature extraction for 2d/3d objects in mesh representation, 2001.

[280] D. Zhang and G. Lu. Shape based image retrieval using generic fourier descriptor, 2002.

[281] E. Zhang, K. Mischaiow, and G. Turk. Feature-based surface parameterization and texture mapping, 2005.

[282] Y. Zhou and A. W. Toga. Efficient skeletonization of volumetric objects, 1999.

[283] S. C. Zhu and A. L. Yuille. Forms: A flexible object recognition and modeling system, 1996.

[284] J. Zimmerman, A. Nealen, and M. Alexa. Silsketch: Automated sketch-based editing of surface meshes, 2007.

Appendix

Publications

This dissertation is based on the papers which are published or submitted on the international conferences.

S. M. Yoon, and A. Kuijper 3D Volume Data Segmentation from Superquadric Tensor Analysis. *In proceeding of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications(GRAPP)*, accepted, 9 pages, 2010.

S. M. Yoon, and A. Kuijper Object Retrieval based on User-Drawn Sketches. *In proceeding of international Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications(VISAPP)*, accepted, 7 pages, 2010.

S. M. Yoon, and A. Kuijper 3D Human Action Recognition using Model Segmentation. *In proceeding of International Conference on Image Analysis and Recognition (ICIAR)*, accepted, 11 pages, 2010.

S. M. Yoon, and A. Kuijper Human Action Recognition using Segmented Skeletal Features. *In proceeding of International Conference on Pattern Recognition (ICPR)*, accepted, 4 pages, 2010.

S. M. Yoon, and A. Kuijper Query-by-Sketch based Image Retrieval using Diffusion Tensor Fields. *In proceeding of IEEE International Conference on Image Processing, Theory, Tools and Applications (IPTA)*, accepted, 6 pages, 2010.

S. M. Yoon, and H. Graf Automatic Skeleton extraction and splitting of Target Objects. *In proceeding of IEEE conference on Image Processing (ICIP)*, pages 2421–2424, 2009.

S. M. Yoon, C. Malerczyk, and H. Graf 3D Skeleton Extraction from Volume Data Based on Normalized Gradient Vector Flow. *In proceeding of International*

Conference on Computer Graphics, Visualization and Computer Vision (WSCG), pages 37–42, 2009.

S. M. Yoon, C. Malerczyk, and H. Graf Skill Measurement through Real-Time 3D Reconstruction and 3D Pose Estimate. *In proceeding of International Conference on Multimodal Interfaces for Skills Transfer*, pages 59–66, 2009.

S. M. Yoon and H. Graf Hierarchical online image representation based on 3D camera geometry. *In proceeding of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pages 54–59, 2009.

S. M. Yoon and H. Graf Automatic Trimap Extraction for Efficient Alpha Matting. *In proceeding of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (GRAPP)*, pages 52–57, 2009.

S. M. Yoon and H. Graf Similarity Measure of the Visual Features using Constrained Hierarchical Clustering for Content Based Image Retrieval. *In proceeding of International Symposium of Visual Computing (ISVC)*, pages 860–868, 2008.

S. M. Yoon and H. Graf Eye tracking based Interaction with 3D Reconstructed Objects. *In proceeding of ACM Multimedia*, pages 881–884, 2008.

S. M. Yoon and H. Graf 3D Eye Position based Interaction within Hierarchically Represented Images. *In proceeding of IEEE 3DTV conference*, pages 389–392, 2008.

Other Publications

S. M. Yoon, S. H. Lee, and J. H. Jeong A novel face and hands tracking in a complex background. *In proceeding of 5th WSEAS International Conference on Computational Intelligence, Man. Machine Systems and Cybernetics*, pages 19–23, 2006.

S. M. Yoon, and S. H. Lee Face tracking using particle filter. *In proceeding of IEEE 14th International Conference on Artificial Reality and Telexistence (ICAT)*, 2004.

S. M. Yoon, and S. C. Kee Speaker detection and tracking at mobile home robot platform. *In proceeding of IEEE International Symposium on Intelligent Signal Processing and Communication System(ISPCAS)*, pages 596–600, 2004.

S. M. Yoon and H. Kim Real-time multiple people detection using skin color, motion, and appearance information. *In proceeding of IEEE 13th International workshop on Robot and Human Interactive Communication(Ro-Man)*, pages 331–334, 2004.

C. Choi, S. M. Yoon, D. Kong, and H. G. Lee Separation of multiple concurrent speeches using audio-visual speaker localization and minimum variance beamforming. *In proceeding of IEEE 8th International Conference on Spoken Language Processing (ICSLP)*, pages 2301–2304, 2004.

S. M. Yoon and S. C. Kee Detection of partially occluded face using Support Vector Machines. *In proceeding of IAPR workshop on Machine Vision and Applications (MVA)*, pages 13–21, 2002.

S. M. Yoon, I.J. Kim, H. Ko, and H. G. Kim Stereo Vision based 3D Input Device. *In proceeding of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2129–2132, 2002.

SANG MIN YOON

Fraunhoferstrasse 5

GRIS, TU Darmstadt

64283, Darmstadt, Germany

Phone: +49-6151-155-618 Email: sangmin.yoon@gris.tu-darmstadt.de

EDUCATION

PHD Degree	**TU Darmstadt, Darmstadt, Germany** Graphisch-Interaktive Systeme(GRIS) Department of Computer Science May 2007- June 2010
M.E. Degree	**Korea University, Seoul, South Korea** Computer vision in **Electronics Engineering** February 2002
B.E. Degree	**Korea University, Seoul, South Korea** **Electronics Engineering** February 2000

Ph.D RESEARCH

THESIS TITLE:	**Markerless 3D motion analysis in the space of diffusion tensor fields and its applications**
ADVISER:	Professor Dr.-Ing. Jose. L. Encarnacao. Professor Dr.techn. Dieter Fellner.
REFEREE	Professor Stefan Roth, Ph.D. Professor Dr. Konrad Schindler. Professor Dr. Sorin Huss

MASTERS RESEARCH

THESIS TITLE:	**Stereo Vision Based 3D Input Device**
ADVISER:	Professor Hanseok Ko Ph.D. Professor Hyung-Gon Kim Ph.D.

EXPERIENCE

GRIS, TU Darmstadt, Germany.

Scientific Researcher 2009 Feb.- Present

ZGDV, Computer Graphics Center, Darmstadt, Germany

Scientific Researcher 2007 May.- 2009, January

Korea National Agent of Information, Seoul, Korea.

Researcher 2006 -2006.

Korea Institute of Science and Technology, Seoul, Korea.

Researcher 2005 -2006.

Samsung Advanced Institute of Technology, Giheung, South Korea.
Computing Lab.,

Researcher 2002 - 2005

Korea Institute of Science and Technology, Seoul, Korea.

Student Researcher 2000 -2001.