Preface

Vision-based motion analysis aims to detect, track and identify objects, and more generally, to understand their behaviors, from image sequences. With the ubiquitous presence of video data and its increasing importance in a wide range of real-world applications such as visual surveillance, human-machine interfaces and sport event interpretation, it is becoming increasingly important to automatically analyze and understand object motions from large amount of video footage.

Not surprisingly, this exciting research area has received growing interest in recent years. Although there has been much progress in the past decades, many challenging problems remain unsolved, e.g., robust object detection and tracking, unconstrained object activity recognition, etc. Recently, statistical machine learning algorithms, such as manifold learning, probabilistic graphical models and kernel machines, have been successfully applied in this area for object tracking, activity modeling and recognition. It is fully believed that novel statistical learning technologies have strong potential to further contribute to the development of robust yet flexible vision systems. The process of improving the performance of vision systems has also brought new challenges to the field of machine learning, e.g., learning from partial or limited annotations, online and incremental learning, and learning with very large datasets. Solving the problems involved in object motion analysis will naturally lead to the development of new machine learning algorithms. In return, new machine learning algorithms are able to address more realistic problems in object motion analysis and understanding.

This edited book highlights the development of robust and effective vision-based motion understanding algorithms and systems from a machine learning perspective. Major contributions of this book are as follows: (1) it provides new researchers with a comprehensive review of the recent development in this field, and presents a variety of study cases where the state-of-the-art learning algorithms have been devised to address specific tasks in human motion understanding; (2) it gives the readers a clear picture of the most active research forefronts and discussions of challenges and future directions, which different levels of researchers might find to be useful for guiding their future research; (3) it draws great strength from the research communities of object motion understanding and machine learning and demonstrates the benefits from the interaction and collaboration of both fields.

The targeted audiences of this edited book are mainly researchers, engineers as well as graduate students in the areas of computer vision, pattern recognition and machine learning. The book is also intend to be accessible to a broader audience including practicing professionals working with specific vision applications such as surveillance, sport event analysis, healthcare, video conferencing, and motion video indexing and retrieval.

The origin of this book stems from the great success of the first and second International Workshop on Machine Learning for Vision-Based Motion Analysis (MLVMA'08 and MLVMA'09), held respectively in conjunction with the European Conference on Computer Vision 2008 (ECCV'08) and the IEEE International Conference on Computer Vision 2009 (ICCV'09). These workshops gathered experts from different fields working on machine learning, computer vision, pattern recognition, and related areas.

The book comprises both theoretical advances and practical applications. The organization of the book reflects the combination of analytical and practical topics addressed throughout the book. We have divided the book chapters into four parts as follows; each addresses a specific theme.

Part I: Manifold Learning and Clustering/Segmentation

Chapter "Practical Algorithms of Spectral Clustering: Toward Large-Scale Vision-Based Motion Analysis", presents some practical algorithms of spectral clustering for large-scale data sets. Spectral clustering is a kernel-based method of grouping data on separate nonlinear manifolds. Reducing its computational expense without critical loss of accuracy contributes to its practical use. The presented algorithms exploit random projection and subsampling techniques for reducing dimensionality and the cost for evaluating pairwise similarities of data. The resulting computation time is quasilinear with respect to the data cardinality, and it can be independent of data dimensionality in some appearance-based applications. The efficiency of the algorithms is extensively demonstrated in appearance-based image/video segmentation.

Chapter "Riemannian Manifold Clustering and Dimensionality Reduction for Vision-Based Analysis" focuses on the topic of segmentation, one fundamental aspect of vision-based motion analysis. The goal of segmentation is to group the data into clusters based upon image properties such as intensity, color, texture or motion. Most existing segmentation algorithms proceed by associating a feature vector to each pixel in the image or video and then segmenting the data by clustering these feature vectors. This process can be phrased as a manifold learning and clustering problem, where the objective is to learn a low-dimensional representation of the underlying data structure and to segment the data points into different groups. Over the past few years, various techniques have been developed for learning a low-dimensional representation of a nonlinear manifold embedded in a high-dimensional space. Unfortunately, most of these techniques are limited to the analysis of a single connected nonlinear manifold and suffer from degeneracies when applied to linear

manifolds. To address this problem, algorithms for performing simultaneous non-linear dimensionality reduction and clustering of data sampled from multiple linear and nonlinear manifolds have been recently proposed. In this chapter, a summary of these newly developed algorithms are given and their applications to vision-based motion analysis are demonstrated.

Chapter "Manifold Learning for Multi-dimensional Auto-regressive Dynamical Models", presents a general differential-geometric framework for learning distance functions for dynamical models. Given a training set of models, the optimal metric is selected among a family of pullback metrics induced by the Fisher information tensor through a parameterized automorphism. The problem of classifying motions, encoded as dynamical models of a certain class, can then be posed on the learnt manifold. In particular, the class of multidimensional autoregressive models of order 2 is considered. Experimental results concerning identity recognition are shown that prove how such optimal pullback Fisher metrics greatly improve classification performances.

Part II: Tracking

When analyzing motion observations extracted from image sequences one notes that the histogram of the velocity magnitude at each pixel shows a large probability mass at zero velocity, while the rest of the motion values may be appropriately modeled with a continuous distribution. This suggests the introduction of mixed-state random variables that have probability mass concentrated in discrete states, while they have a probability density over a continuous range of values. In the first part of chapter "Mixed-State Markov Models in Image Motion Analysis", a comprehensive description of the theory behind mixed-state statistical models, in particular the development of mixed-state Markov models that permits to take into account spatial and temporal interaction, is given. The presentation generalizes the case of simultaneous modeling of continuous values and any type of discrete symbolic states. For the second part, the application of mixed-state models to motion texture analysis is presented. Motion textures correspond to the instantaneous apparent motion maps extracted from dynamic textures. They depict mixed-state motion values with a discrete state at zero and a Gaussian distribution for the rest. Mixed-state Markov random fields and mixed-state Markov chains are defined and applied to motion texture recognition and tracking.

Chapter "Learning to Track Objects in Surveillance Image Streams at Very Low Frame Rate", studies on the problem of object tracking. Some camera surveillance systems are designed to be autonomous—both from the energy and memory point of view. Autonomy allows operation in environments where wiring cameras for power and data transmission is neither feasible nor desirable. In these contexts, for cameras to work unattended over long periods requires choosing an adequately low frame rate to match the speed of the process to be supervised while minimizing energy and memory consumption. The result of surveillance is then a large stream of images acquired sparsely over time with limited visual continuity from one frame to the

other. Reviewing these images to detect events of interest requires techniques that do not assume traceability of objects by visual similarity. If the process to be surveyed shows recurrent patterns of events over time, as it is often the case in industrial settings, other possibilities open up. Since images are time-stamped, this suggests techniques which use temporal data to help detecting relevant events. This contribution presents an image review tool that combines in cascade a scene change detector (SCD) with a temporal filter. The temporal filter learns to recognize relevant SCD events by their time distribution on the image stream. The learning phase is supported by image annotations provided by end-users during past reviews. The concept is tested on a benchmark of real surveillance images stemming from a nuclear safeguards context. Experimental results show that the combined SCD-temporal filter significantly reduces the workload necessary to detect safeguards-relevant events in large image streams.

In chapter "Discriminative Multiple Target Tracking", a metric learning framework is introduced to learn a single discriminative appearance model for robust visual tracking of multiple targets. The single appearance model effectively captures the discriminative visual information among the different visual targets as well as the background. The appearance modeling and the tracking of the multiple targets are all cast in a discriminative metric learning framework. An implicit exclusive principle is naturally reinforced in the proposed framework, which renders the tracker to be robust to cross occlusions among the multiple targets. The efficacy of the proposed multi-target tracker is demonstrated on benchmark visual tracking sequences and real-world video sequences as well.

Guidewire tracking in fluoroscopy is important to image guided interventions. In chapter "Applications of Wire Tracking in Image Guided Interventions", a semantic guidewire model, is introduced, based on which a probabilistic method is presented to integrate measurements of three guidewire parts, i.e., a catheter tip, a guidewire body and a guidewire tip, in a Bayesian framework to track a whole guidewire. This tracking framework is robust to measurement noises at individual guidewire parts. Learning based measurement models are used to track the guidewire. The learning-based measurement models are trained from a database of guidewires, to detect guidewire parts in low-quality images. The method further incorporates online measurement models, which are based on guidewire appearances, as a complementary to learning based measurements to improve the tracking robustness. A hierarchical and multi-resolution scheme is developed to track a deforming guidewire. By decomposing the guidewire motion into two major components, the hierarchical tracking starts from a rigid alignment, followed by a refined nonrigid tracking. At each stage, a multi-resolution searching strategy is applied by using variable bandwidths in a kernel-based measurement smoothing method, to effectively and efficiently track the deforming guidewire. The guidewire tracking framework is validated on a test set containing 47 sequences that are captured in real-life interventional scenario. Quantitative evaluation results show that the mean tracking error on guidewires is less than 2 pixels, i.e., 0.4 mm.

Part III: Motion Analysis and Behavior Modeling

Chapter "An Integrated Approach to Visual Attention Modeling for Saliency Detection in Videos", presents a framework to learn and predict regions of interest in videos, based on human eye movements. The eye gaze information of several users are recorded as they are watching videos that belong to a similar application domain. This information is used to train a classifier to learn low-level video features from regions that attracted the visual attention of users. Such a classifier is combined with vision-based approaches to provide an integrated framework to detect salient regions in videos. To date, saliency prediction has been viewed from two different perspectives, namely visual attention modeling and spatiotemporal interest point detection. These approaches have largely been pure-vision based. They detect regions having a predefined set of characteristics such as complex motion or high contrast, for all kinds of videos. However, what is 'interesting' varies from one application to another. By learning features of regions that capture the attention of viewers while watching a video, this chapter aims to distinguish those that are actually salient in the given context, from the rest. The integrated approach ensures that both regions with anticipated content (top–down attention) and unanticipated content (bottom–up attention) are predicted by the proposed framework as salient. In the experiments with news videos of popular channels, the results show a significant improvement in the identification of relevant salient regions in such videos, when compared with existing approaches.

Chapter "Video-Based Human Motion Estimation by Part-Whole Gait Manifold Learning", presents a general gait representation framework for video-based human motion estimation that involves gait modeling at both the whole and part levels. The goal is to estimate the kinematics of an unknown gait from image sequences taken by a single camera. This approach involves two generative models, called the kinematic gait generative model (KGGM) and the visual gait generative model (VGGM), which represent the kinematics and appearances of a gait by a few latent variables, respectively. Particularly, the concept of gait manifold is proposed to capture the gait variability among different individuals by which KGGM and VGGM can be integrated together for gait estimation, so that a new gait with unknown kinematics can be inferred from gait appearances via KGGM and VGGM. A key issue in generating a gait manifold is the definition of the distance function that reflects the dissimilarity between two individual gaits. Specifically, three distance functions each of which leads to a specific gait manifold are investigated and compared. Moreover, this gait modeling framework from the whole level to the part level has been extended by decomposing a gait into two parts, an upper-body gait and a lower-body gait, each of which is associated with a specific gait manifold for part level gait modeling. Also, a two-stage inference algorithm is employed for whole-part gait estimation. The proposed algorithms were trained on the CMU Mocap data and tested on the HumanEva data, and the experiment results show promising results compared with the state-of-the-art algorithms with similar experimental settings.

Extremely crowded scenes present unique challenges to motion-based video analysis due to the excessive quantity of pedestrians and the large number of occlusions they produce. The interactions between pedestrians, however, collectively

form a crowd that exhibits a spatially and temporally structured motion pattern within the scene. In chapter "Spatio-Temporal Motion Pattern Models of Extremely Crowded Scenes", a novel statistical framework is presented for modeling this structured motion pattern behavior, or steady state, of extremely crowded scenes. The key insight is to model the crowd by the spatial and temporal variations of the local non-uniform motion patterns generated by pedestrian interactions. The pedestrian activity is represented by modeling the rich motion information in local space-time volumes of the video. In order to capture the motion variations of the scene, a novel distribution-based hidden Markov model that encodes the temporal variations of local motion pattern is introduced. It is demonstrated that by capturing the steady-state behavior of a scene, the proposed method can naturally detect unusual activities as statistical deviations in videos with complex activities that are hard for even human observers to analyze.

Chapter "Learning Behavioral Patterns of Time Series for Video-Surveillance", deals with the problem of learning behaviors of people activities from (possibly big) sets of visual dynamic data, with a specific reference to video-surveillance applications. The study focuses mainly on devising meaningful data abstractions able to capture the intrinsic nature of the available data, and applying similarity measures appropriate to the specific representations. The methods are selected among the most promising techniques available in the literature and include classical curve fitting, string-based approaches, and hidden Markov models. The analysis considers both supervised and unsupervised settings and is based on a set of loosely labeled data acquired by a real video-surveillance system. The experiments highlight different peculiarities of the methods taken into consideration, and the final discussion guides the reader towards the most appropriate choice for a given scenario.

Part IV: Gesture and Action Recognition

Chapter "Recognition of Spatiotemporal Gestures in Sign Language Using Gesture Threshold HMMs", proposes a framework for automatic recognition of spatiotemporal gestures in sign language. An extension to the standard HMM model to develop a gesture threshold HMM (GT-HMM) framework is implemented which is specifically designed to identify inter gesture transitions. The performance of this system, and different CRF systems, is evaluated, when recognizing gestures and identifying inter gesture transitions. The evaluation of the system included testing the performance of conditional random fields (CRF), hidden CRF (HCRF) and latent-dynamic CRF (LDCRF) based systems and comparing these to the presented GT-HMM based system when recognizing motion gestures and identifying inter gesture transitions.

Learning-based approaches for human action recognition often rely on large training sets. Most of these approaches do not perform well when only a few training samples are available. Chapter "Learning Transferable Distance Functions for Human Action Recognition", considers the problem of human action recognition from a single clip per action. Each clip contains at most 25 frames. Using a patch based

motion descriptor and matching scheme, promising results on three different action datasets with a single clip as the template can be achieved. The results are comparable to previously published results using much larger training sets. A method for learning a transferable distance function is also presented for these patches. The transferable distance function learning extracts generic knowledge of patch weighting from previous training sets, and can be applied to videos of new actions without further learning. Experimental results show that the transferable distance function learning not only improves the recognition accuracy of the single clip action recognition, but also significantly enhances the efficiency of the matching scheme.

Bath, UK   Liang Wang
Oulu, Finland   Guoying Zhao
Singapore, Singapore   Li Cheng
Oulu, Finland   Matti Pietikäinen