Preface

The goal of vision-based motion analysis is to provide computers with intelligent perception capacities so they can sense the objects and understand their behaviors from video sequences. With the ubiquitous presence of video data and the increasing importance in a wide range of applications such as visual surveillance, human-machine interfaces, and sport event interpretation, it is becoming increasingly demanding 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 significant progress in the past decades, many challenging problems remain unsolved, e.g., robust object detection and tracking, unconstrained object activity recognition, etc. The field of machine learning, on the other hand, is driven by the idea that the essential rules or patterns behind data can be *automatically* learned by a computer or a system. Statistical learning approach is one major frontier for computer vision research. We have evidenced in recent years a growing number of successes of machine learning applications to certain vision problems. It is fully believed that machine learning technologies is going to significantly contribute to the development of practical systems for vision-based motion analysis.

This edited book presents and highlights a collection of recent developments along this direction. A brief summary of each chapter is presented as follow:

Chapter 1, *Human Motion Tracking in Video: A Practical Approach*, presents a new formulation for the problem of human motion tracking in video. Tracking is still a challenging problem when strong appearance changes occur as in videos of humans in motion. A solution is to use an online method that updates iteratively a subspace of reference target models, integrating color and motion cues in a particle filter framework to track human body parts. The algorithm process consists of two modes, switching between detection and tracking. The detection steps involve trained classifiers to update estimated positions of the tracking windows, whereas tracking steps rely on an adaptive color-based particle filter coupled with optical flow estimations. The Earth Mover distance is used to compare color models in a global fashion, and constraints on flow features avoid drifting effects. The proposed method has revealed its efficiency to track body parts in motion and can cope with full appearance changes.

Chapter 2, *Learning to Recognise Spatio-Temporal Interest Points*, presents a generic classifier for detecting spatio-temporal interest points within video. The premise being that, given an interest point detector, a classifier is learnt that duplicates its functionality, which is both accurate and computationally efficient. This means that interest point detection can be achieved independent of the complexity of the original interest point formulation. The naive Bayesian classifier of Ferns is extended to the spatio-temporal domain and learn classifiers that duplicate the functionality of common spatio-temporal interest point detectors. Results demonstrate accurate reproduction of results with a classifier that can be applied exhaustively to video at frame-rate, without optimisation, in a scanning window approach.

Chapter 3, *Graphical Models for Representation and Recognition of Human Actions*, reviews graphical models that provide a natural framework for representing state transitions in events and also the spatio-temporal constraints between the actors and events. Hidden Markov Models (HMMs) have been widely used in several action recognition applications but the basic representation has three key deficiencies: These include unrealistic models for the duration of a sub-event, not encoding interactions among multiple agents directly and not modeling the inherent hierarchical organization of these activities. Several extensions have been proposed to address one or more of these issues and have been successfully applied in various gesture and action recognition domains. More recently, Conditional Random Fields (CRF) are becoming increasingly popular since they allow complex potential functions for modeling observations and state transitions, and also produce superior performance to HMMs when sufficient training data is available. This chapter first reviews the various extensions of these graphical models, then presents the theory of inference and learning in them and finally discusses their applications in various domains.

Chapter 4, *Common Spatial Patterns for Real-Time Classification of Human Actions*, presents a discriminative approach to human action recognition. At the heart of the approach is the use of common spatial patterns (CSP), a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. Such a transformation focuses on differences between classes, rather than on modeling each class individually. The most likely class is found by pairwise evaluation of all discriminate functions, which can be done in real-time. Image representations are silhouette boundary gradients, spatially binned into cells. The method achieves scores of approximately 96% on the Weizmann human action dataset, and shows that reasonable results can be obtained when training on only a single subject.

Chapter 5, *KSM Based Machine Learning for Markless Motion Capture*, proposes a marker-less motion capture system, based on machine learning. Pose information is inferred from images captured from multiple (as few as two) synchronized cameras. The central concept of which, they call: Kernel Subspace Mapping (KSM). The images-to-pose learning could be done with large numbers of images of a large variety of people (and with the ground truth poses accurately known). What makes machine learning viable for human motion capture is that a high percentage of human motion is coordinated. Indeed, it is now relatively well known that there is large redundancy in the set of possible images of a human (these images form some sort of relatively smooth lower dimensional manifold in the huge dimensional space of all possible images) and in the set of pose angles (again, a low dimensional and smooth sub-manifold of the moderately high dimensional space of all possible joint angles). KSM, is based on the KPCA (Kernel PCA) algorithm, which is costly. They show that the Greedy Kernel PCA (GKPCA) algorithm can be used to speed up KSM, with relatively minor modifications. At the core, then, is two KPCA's (or two GKPCA's) - one for the learning of pose manifold and one for the learning image manifold. Then they use a modification of Local Linear Embedding (LLE) to bridge between pose and image manifolds.

Chapter 6, *Multi-Scale People Detection and Motion Analysis for Video Surveillance*, addresses visual processing of people, including detection, tracking, recognition, and behavior interpretation, a key component of intelligent video surveillance systems. Computer vision algorithms with the capability of "looking at people" at multiple scales can be applied in different surveillance scenarios, such as far-field people detection for wide-area perimeter protection, midfield people detection for retail/banking applications or parking lot monitoring, and near-field people/face detection for facility security and access. In this chapter, they address the people detection problem in different scales as well as human tracking and motion analysis for real video surveillance applications including people search, retail loss prevention, people counting, and display effectiveness.

Chapter 7, *A Generic Framework for 2D and 3D Upper Body Tracking*, targets upper body tracking, a problem to track the pose of human body from video sequences. It is difficult due to such problems as the high dimensionality of the state space, the self-occlusion, the appearance changes, etc. In this chapter, they propose a generic framework that can be used for both 2D and 3D upper body tracking and can be easily parameterized without heavily depending on supervised training. They first construct a Bayesian Network (BN) to represent the human upper body structure and then incorporate into the BN various generic physical and anatomical constraints on the parts of the upper body. They also explicitly model part occlusion in the model, which allows to automatically detect the occurrence of self-occlusion and to minimize the effect of measurement errors on the tracking accuracy due to occlusion. Using the proposed model, upper body tracking can be performed through probabilistic inference over time. A series of experiments were performed on both monocular and stereo video sequences to demonstrate the effectiveness and capability of the model in improving upper body tracking accuracy and robustness.

Chapter 8, *Real-Time Recognition of Basic Human Actions*, describes a simple and computationally efficient, appearance-based approach for real-time recognition of basic human actions. They apply a technique that depicts the differences between two or more successive frames accompanied by a threshold filter to detect the regions of the video frames where some type of human motion is observed. From each frame difference, the algorithm extracts an incomplete and unformed human body shape and generates a skeleton model which represents it in an abstract way. Eventually, the recognition process is formulated as a time-series problem and handled by a very robust and accurate prediction method (Support Vector Regression). The proposed technique could be employed in applications such as vision-based autonomous robots and surveillance systems.

Chapter 9, *Fast Categorisation of Articulated Human Motion*, exploits the problem of visual categorisation of human motion in video clips. Most published methods either analyse an entire video and assign it a single category label, or use relatively large look-ahead to classify each frame. Contrary to these strategies, the human visual system proves that simple categories can be recognised almost instantaneously. Here they present a system for categorisation from very short sequences ("snippets") of 1–10 frames, and systematically evaluate it on several data sets. It turns out that even local shape and optic flow for a single frame are enough to achieve 80-90% correct classification, and snippets of 5-7 frames (0.2-0.3 seconds of video) yield results on par with the ones state-of-the-art methods obtain on entire video sequences.

Chapter 10, *Human Action Recognition with Expandable Graphical Models*, proposes an action recognition system that is independent of the subjects who perform the actions, independent of the speed at which the actions are performed, robust against noisy extraction of features used to characterize the actions, scalable to large number of actions and expandable with new actions. In this chapter, they describe a recently proposed expandable graphical model of human actions that has the promise to realize such a system. This chapter first presents a brief review of the recent development in human action recognition. Then, the expandable graphical model is presented in detail and a system that learns and recognizes human actions from sequences of silhouettes using the expandable graphical model is developed.

Chapter 11, *Detection and Classification of Interacting Persons*, presents a way to classify interactions between people. Examples of the interactions they investigate are; people meeting one another, walking together and fighting. A new feature set is proposed along with a corresponding classification method. Results are presented which show the new method performing significantly better than the previous state of the art method as proposed by Oliver et al.

Chapter 12, *Action Recognition*, first reviews the current action recognition methods from the following two aspects: action representation and recognition strategy. Then, a novel method for classifying

human actions from image sequences is investigated. In this method, the human action is represented by a set of shape context features of human silhouette, and a dominant sets-based approach is employed to classify the predefined actions. The comparison between the dominant sets-based approach with K-means, mean shift, and Fuzzy-Cmean is also discussed.

Chapter 13, *Distillation: A Super-Resolution Approach for the Selective Analysis of Noisy and Unconstrained Video Sequences*, argues that image super-resolution is one of the most appealing applications of image processing, capable of retrieving a high resolution image by fusing several registered low resolution images depicting an object of interest. However, employing super-resolution in video data is challenging: a video sequence generally contains a lot of scattered information regarding several objects of interest in cluttered scenes. The objective of this chapter is to demonstrate why standard image super-resolution fails in video data, which are the problems that arise, and how they can overcome these problems. They propose a novel Bayesian framework for super-resolution of persistent objects of interest in video sequences, called *Distillation*. With Distillation, they extend and generalize the image super-resolution task, embedding it in a structured framework that accurately distills all the informative bits of an object of interest. They also extend the Distillation process to deal with objects of interest whose transformations in the appearance are not (only) rigid. The ultimate product of the overall process is a strip of images that describe at high resolution the dynamics of the video, switching between alternative local descriptions in response to visual changes. The approach is first tested on synthetic data, obtaining encouraging comparative results with respect to known super-resolution techniques, and a good robustness against noise. Second, real data coming from different videos are considered, trying to solve the major details of the objects in motion.

In summary, this book contains an excellent collection of theoretical and technical chapters written by different authors who are worldwide-recognized researchers on various aspects of human motion understanding using machine learning methods. The targeted audiences are mainly researchers, engineers as well as graduate students in the areas of computer vision 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 video surveillance, sport event analysis, healthcare, video conferencing, motion video indexing and retrieval. We wish this book would help toward the development of robust yet flexible vision systems.

Liang Wang
University of Bath, UK

Li Cheng
TTI-Chicago, USA

Guoying Zhao
University of Oulu, Finland

May 20, 2009