Preface

This book presents an extended version of a few selected papers originally submitted to the 11th International Workshop on Image Analysis for Multimedia Interactive Services, which took place in April 2010 in Desenzano del Garda, Brescia, Italy. This workshop is one of the main international events for the presentation and discussion of the latest technological advances in interactive multimedia services. The objective of the workshop is to bring together researchers and developers from academia and industry working in the areas of image, video, and audio applications, with a special focus on analysis.

The book is organized into five main sections, considering Multimedia Content Analysis, Motion and Activity Analysis, High-Level Descriptors and Video Retrieval, 3D and Multi-View, and Multimedia Delivery.

Part 1: Multimedia Content Analysis

Multimedia Content Analysis is of great relevance in the scenario of image analysis for multimedia interactive services. In this respect, it is very important to consider also the audio signal and caption text eventually superimposed on the considered images. Also, the objects displayed in the images could be very helpful in content analysis.

Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso, in the book chapter "*On the use of audio events for improving video scene segmentation*" deal with the problem of automatic temporal segmentation of a video into elementary semantic units known as scenes. The novelty lies in the use of high-level audio information, in the form of audio events, for the improvement of scene segmentation performance. More specifically, the proposed technique is built upon a recently proposed audio-visual scene segmentation approach that involves the construction of multiple scene transition graphs (STGs) that separately exploit information coming from different modalities. In the extension of the latter approach presented in this chapter, audio event detection results are introduced to the definition of an audio-based scene transition graph, while a visual-based scene transition graph is also defined independently. The results of these two types of STGs are subsequently combined. The results of the application of the proposed technique to broadcast videos demonstrate the usefulness of audio events for scene segmentation and highlight the importance of introducing additional high-level information to the scene segmentation algorithms.

The important problem of caption text extraction is addressed in the chapter "*Region-based caption text extraction*", authored by Miriam León, Veronica Vilaplana, Antoni Gasull, and Ferran Marques. The authors present a method for caption text detection that takes advantage of texture and geometric features to detect the caption text. Texture features are estimated using wavelet analysis and mainly applied for *text candidate spotting*. In turn, *text characteristics verification* relies on geometric features, which are estimated exploiting the region-based image model. Analysis of the region hierarchy provides the final caption text objects. The final step of *consistency analysis for output* is performed by a binarization algorithm that robustly estimates the thresholds on the caption text area of support.

Image classification is a challenging task in computer vision. For e.g., fully understanding real-world images may involve both scene and object recognition. Many approaches have been proposed to extract meaningful descriptors from images and classifying them in a supervised learning framework. In the chapter "*K-nn boosting prototype learning for object classification*", Paolo Piro, Michel Barlaud, Richard Nock, and Frank Nielsen, revisit the classic k-nearest neighbors classification rule, which has shown to be very effective when dealing with local image descriptors. However, *k-nn* still features some major drawbacks, mainly due to the uniform voting among the nearest prototypes in the feature space. In this chapter, the authors propose therefore a generalization of the classic knn rule in a supervised learning (boosting) framework. Namely, they redefine the voting rule as a strong classifier that linearly combines predictions from the k closest prototypes. In order to induce this classifier, they propose a novel learning algorithm, *MLNN* (Multiclass Leveraged Nearest Neighbors), which gives a simple procedure for performing prototype selection very efficiently. Experiments carried out first on object classification using 12 categories of objects, then on scene recognition, using 15 real-world categories, show significant improvement over classic *K-nn* in terms of classification performances.

Part 2: Motion and Activity Analysis

Motion and activity information plays certainly a crucial role in content-based video analysis and retrieval. In this context the problem of automatic tracking of moving object in a video have been extensively studied in the literature and also in this book.

In the book chapter titled "*Semi-automatic object tracking in video sequences by extension of the MRSST algorithm*", Marko Esche, Mustafa Karaman, and Thomas Sikora investigate a new approach for segmentation of real-world objects in video sequences. While some amount of user interaction is still necessary for most algorithms in this field, in order for them to produce adequate results, these can be reduced making use of certain properties of graph-based image segmentation algorithms. Based on one of these algorithms a framework is proposed that tracks individual foreground objects through arbitrary video sequences and partly automates the necessary corrections required from the user. Experimental results suggest that the proposed algorithm performs well on both low- and high-resolution video sequences and can even, to a certain extent, cope with motion blur and gradual object deformations.

The problem of tracking a non-rigid object in an uncalibrated static multi-camera environment is considered in "*A multi-resolution particle filter tracking with a dual consistency check for model update in a multi-camera environment*", where Yifan Zhou, Jenny Benois-Pineau, and Henri Nicolas present a novel tracking method with a multi-resolution approach and a dual model check. The proposed method is based on particle filtering using color features. The major contributions of the method are: multi-resolution tracking to handle strong and non-biased object motion by short-term particle filters; stratified model consistency check by Kolmogorov-Smirnov test, and object trajectory-based view corresponding deformation in a multi-camera environment.

An interesting application of trajectories analysis in a surveillance scenario is proposed by Mattia Daldoss, Nicola Piotto, Nicola Conci, and Francesco G. B. De Natale in the book chapter "*Activity detection using regular expressions*". The authors propose a novel method to analyze trajectories in surveillance scenarios by means of Context-Free Grammars (CFGs). Given a training corpus of trajectories associated to a set of actions, a preliminary processing phase is carried out to characterize the paths as sequences of symbols. This representation turns the numerical representation of the coordinates into a syntactical description of the activity structure, which is successively adopted to identify different behaviors through the CFG models. Such a modeling is the basis for the classification and matching of new trajectories versus the learned templates and it is carried out through a parsing engine that enables the online recognition of human activities. An additional module is provided to recover parsing errors (i.e., insertion, deletion, or substitution of symbols) and update the activity models previously learned. The proposed system has been validated in indoor, in an assisted living context, demonstrating good capabilities in recognizing activity patterns in different configurations, and in particular in presence of noise in the acquired trajectories, or in case of concatenated and nested actions.

Katharina Quast, and André Kaup, in "*Shape adaptive mean shift object tracking using gaussian mixture models*" propose a new object tracking algorithm based on a combination of the mean shift and Gaussian mixture models (GMMs), named GMM-SAMT. GMM-SAMT stands for Gaussian mixture model-based shape adaptive mean shift tracking. Instead of a symmetrical kernel like in traditional mean shift tracking, GMM-SAMT uses an asymmetric shape adapted kernel which is retrieved from an object mask. During the mean shift iterations the kernel scale is altered according to the object scale, providing an initial adaptation of the object shape. The final shape of the kernel is then obtained by segmenting the area inside and around the adapted kernel into object and non-object segments using Gaussian mixture models.

Part 3: High-Level Descriptors and Video Retrieval

In the context of content-based video retrieval the high-level descriptors are clearly of great relevance. This topic is covered in this part of the book.

Seunghan Han, Bonjung Koo, Andreas Hutter, and Walter Stechele in "*Forensic reasoning upon pre-obtained surveillance metadata using uncertain spatiotemporal rules and subjective logic*" present an approach to modeling uncertain contextual rules using subjective logic for forensic visual surveillance. Unlike traditional real-time visual surveillance, forensic analysis of visual surveillance data requires matching of high level contextual cues with observed evidential metadata where both the specification of the context and the metadata suffer from uncertainties. To address this aspect, there has been work on the use of declarative logic formalisms to represent and reason about contextual knowledge, and on the use of different uncertainty handling formalisms. In such approaches, uncertainty attachment to logical rules and facts are crucial. However, there are often cases that the truth value of rule itself is also uncertain thereby, uncertainty attachment to rule itself should be rather functional. '*The more X then the more Y*' type of knowledge is one of the examples. To enable such type of rule modeling, in this chapter, the authors propose a reputational subjective opinion function upon logic programming, which is similar to fuzzy membership function but can also take into account uncertainty of membership value itself. Then they further adopt subjective logic's fusion operator to accumulate the acquired opinions over time. To verify the proposed approach, the authors present a preliminary experimental case study on reasoning likelihood of being a good witness that uses metadata extracted by a person tracker and evaluates the relationship between the tracked persons. The case study is further extended to demonstrate more complex forensic reasoning by considering additional contextual rules.

Nowadays, multimedia data is produced and consumed at an ever-increasing rate. Similar to this trend, diverse storage approaches for multimedia data have been introduced. These observations lead to the fact that distributed and heterogeneous multimedia repositories exist, whereas an easy and unified access to the stored multimedia data is not given. In this respect, Florian Stegmaier, Mario Döller, Harald Kosch, Andreas Hutter, and Thomas Riegel in "*AIR: architecture for interoperable retrieval on distributed and heterogeneous multimedia repositories*" present an architecture, named AIR, that offers the aforementioned retrieval possibilities. To ensure interoperability, AIR makes use of recently issued standards, namely the MPEG Query Format (multimedia query language) and the JPSearch transformation rules (metadata interoperability).

In the final chapter of this section, the detection of high-level concepts in video is considered. More specifically, Vasileios Mezaris, Anastasios Dimou, and Ioannis Kompatsiaris propose in "*Local invariant feature tracks for high-level video feature extraction*" the use of feature tracks for the detection of high-level features (concepts) in video. Extending previous work on local interest point detection and description in images, feature tracks are defined as sets of local interest points that are found in different frames of a video shot and exhibit spatio-temporal and visual continuity, thus defining a trajectory in the 2D + Time space. These tracks jointly capture the spatial attributes of 2D local regions and their corresponding long-term motion. The extraction of feature tracks and the selection and representation of an appropriate subset of them allow the generation of a Bag-of-Spatiotemporal-Words model for the shot, which facilitates capturing the dynamics of video content. Experimental evaluation of the proposed approach on two challenging datasets (TRECVID 2007, TRECVID 2010) highlights how the selection, representation, and use of such feature tracks enhance the results of traditional keyframe-based concept detection techniques.

Part 4: 3D and Multi-View

Among the various audio-visual descriptors useful for image and video analysis and coding there are the descriptors related to 3D structure and multi-view. In this section of the book we cover this topic, considering both the issue of 3D stereo correspondences and 3DTV video coding.

The problem of 3D stereo correspondences is considered in "*A new evaluation criterion for point correspondences in stereo images*" by Aleksandar Stojanovic, and Michael Unger. In this chapter, the authors present a new criterion to evaluate point correspondences within a stereo setup. Many applications such as stereo matching, triangulation, lens distortion correction, and camera calibration require an evaluation criterion for point correspondences. The common criterion used is the epipolar distance. The uncertainty of the epipolar geometry provides additional information, and the proposed method uses this information for a new distance measure. The basic idea behind this criterion is to determine the most probable epipolar geometry that explains the point correspondence in the two views. This criterion considers the fact that the uncertainty increases for point correspondences induced by world points that are located at a different depth-level compared to those that were used for the fundamental matrix computation. Furthermore, the authors show that by using Lagrange multipliers, this constrained minimization problem can be reduced to solving a set of three linear equations with a computational complexity practically equal to the complexity of the epipolar distance.

A novel learning-based approach used to estimate local homography of points belonging to a given surface is proposed in "*Local homography estimation using* *keypoint descriptors*" by Alberto Del Bimbo, Fernando Franco, and Federico Pernici. In this chapter the authors present a new learning-based approach used to estimate local homography of points belonging to a given surface and show that it is more accurate than specific affine region detection methods. While other works attempt to do this task by using iterative algorithms developed for template matching, this method introduces a direct estimation of the transformation. In more details, it performs the following steps. First, a training set of features captures the geometry and appearance information about keypoints taken from multiple views of the surface. Then, incoming keypoints are matched against the training set in order to retrieve a cluster of features representing their identity. Finally the retrieved clusters are used to estimate the local homography of the regions around keypoints. Thanks to the high accuracy, outliers and bad estimates are filtered out by multiscale Summed Square Difference test.

The problem of 3DTV multiple description coding is addressed by Simone Milani and Giancarlo Calvagno in the book chapter titled "*A cognitive source coding scheme for multiple description 3DTV transmission*". In this framework, Multiple Description Coding has recently proved to be an effective solution for the robust transmission of 3D video sequences over unreliable channels. However, adapting the characteristics of the source coding strategy (Cognitive Source Coding) permits improving the quality of 3D visualization experienced by the end-user. This strategy has been successfully employed for standard video signals, but it can be applied to Multiple Description video coding for an effective transmission of 3D signals. The chapter presents a novel Cognitive Source Coding scheme that improves the performance of traditional Multiple Description Coding approaches by adaptively combining traditional predictive and Wyner-Ziv coders according to the characteristics of the video sequence and to the channel conditions. The approach is employed for video + depth 3D transmissions improving the average PSNR value up to 2.5 dB with respect to traditional MDC schemes.

Part 5: Multimedia Delivery

In the final section of the book we consider the important aspects related to the problem of multimedia documents delivery, focusing the attention on both images and video.

In "*An efficient prefetching strategy for remote browsing of JPEG 2000 image sequences*", Juan Pablo GarcÍa Ortiz, Vicente González Ruiz, Inmaculada Garcıa, Daniel Müller, and George Dimitoglou propose an efficient prefetching strategy for interactive remote browsing of sequences of high resolution JPEG 2000 images. As a result of the inherent latency of client–server communication, the experiments of this study prove that a significant benefit can be achieved, in terms of both quality and responsiveness, by anticipating certain data from the rest of the sequence while an image is being explored. In this work a model based on the quality progression of the image is proposed in order to estimate which percentage of the bandwidth will be dedicated to prefetching. This solution can be easily implemented on top of any existing remote browsing architecture.

Matteo Naccari and Fernando Pereira in "*Comparing spatial masking modelling in just noticeable distortion controlled H.264/AVC video coding*" study the integration of a just noticeable distortion model in the H.264/AVC standard video codec to improve the final rate-distortion performance. Three masking aspects related to lossy transform coding and natural video contents are considered: frequency band decomposition, luminance component variations and pattern masking. For the latter aspect, three alternative models are considered, namely the Foley-Boynton, Foley-Boynton adaptive, and Wei-Ngan models. Their performance, measured for high definition video contents, and reported in terms of bitrate improvement and objective quality loss, reveals that the Foley-Boynton and its adaptive version provide the best performance with up to 35.6 % bitrate reduction at the cost of at most 1.4 % objective quality loss.

In traditional motion compensated predictive video coding, both the motion vector and the prediction residue are encoded and stored or sent for every predicted block. The motion vector brings displacement information with respect to a reference frame while the residue represents what we really consider to be the innovation of the current block with respect to that reference frame. This encoding scheme has proved to be extremely effective in terms of rate distortion performance. Nevertheless, one may argue that full description of motion and residue could be avoided if the decoder could be made able to exploit a proper a priori model for the signal to be reconstructed. In particular, it was recently shown that a smart enough decoder could exploit such an a priori model to partially infer motion information for a single block given only neighboring blocks and the innovation of that block. The last contribution, given by Claudia Tonoli and Marco Dalai presents an improvement over the single-block method. In the book chapter "*Coherent video reconstruction with motion estimation at the decoder*" the authors show that higher performance can be achieved by simultaneously reconstructing a frame region composed of several blocks, rather than reconstructing those blocks separately. A trellis-based algorithm is developed in order to make a global decision on many motion vectors at a time instead of many single separate decisions on different vectors.

Brescia, Italy
London, UK

Nicola Adami
Andrea Cavallaro
Riccardo Leonardi
Pierangelo Migliorati