SECTION 1

IMAGE EXTRACTION

This section addresses the extraction of elements from images, including objects, visual concepts, shape, emotional faces, and social media. With the explosion of digital photography and exponential growth of image sharing sites such as Flickr, the need for image extraction to support retrieval, summarization, and analysis is increasing. Challenges include scaling up to large collections, learning and relating linguistic descriptions and visual or semantic features, improving precision and recall, and intuitive interaction.

In the first chapter, Madirakshi Das, Alexander Loui, and Andrew Blose from Eastman Kodak address the challenge of searching large collections of digital photographs. The authors report novel methods for retrieving consumer images with the same object, typically taken at some location. High precision matching between two photos coupled with searches on date, events, and people in images provides a powerful consumer ability to help tag events with location. This approach enables search by location (e.g., "find all pictures taken in my living room") or for objects in a location (e.g., "look for picture of a friend at party in my backyard"). The authors' system uses SIFT (Scale Invariant Feature Transform) features to match unique keypoints in a scene, next it clusters matched points into spatial groups, and then it removes false matches by constraining clusters to those that are correlated, have consistent trajectories, and are compact. The authors evaluated their system by creating a consumer application for retrieving and tagging images from the personal image collections of 18 subjects each having 1–2 thousand images from a 1–2 year time frame. Ninety groups of images with two to five images each containing objects were carefully selected, which could not be matched using standard color histograms or texture cues and which included occlusion, partial visibility, varying viewpoints, and variable objects. The authors' algorithm provides 85% recall and

Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring, First Edition. Edited by Mark T. Maybury.

63% precision. Using an additional step of pruning retrieved images with a match score lower than a threshold, the precision improves to 85%. In contrast, logo detection in consumer images is particularly challenging on deformable three-dimensional objects, such as containers, billboards, or clothing. The authors tested logo detection using 32 images from a stock car race event where corporate sponsorship appears in many forms (e.g., car logos and logo billboards). Images were captured by spectators using consumer digital cameras. False positive rates (FPR) decreased as feature resolution increased. Applications of this research include personal and social media retrieval, as well as content-based targeted advertising.

The second chapter by Keiji Yanai and Hidetoshi Kawakubo at the University of Electro-Communications in Tokyo and Kobus Barnard at the University of Arizona turns to the analysis of annotations or tags on photos in collections such as Flickr or Picasa. The authors use entropy to analyze two types of image tags: those about image visual features and those about image geolocation. Using a 40 thousand-image collection from the World Wide Web using Google Image search on 150 adjectives, the authors assess 150 adjectives with respect to visual features and relations between image features and 230 nouns with respect to geotags. Using entropy to analyze the distribution of features, the authors discovered that concepts with low image feature entropy tend to have high geolocation entropy and vice versa. For example, sky-related concepts such as "sun," "moon," and "rainbow" have low image region entropy and high geolocation entropy, whereas concepts related to places such as "Rome," "Deutschland," and "Egypt" have high image region entropy and low geolocation entropy. The authors developed two methods to compute image region entropy, one using Gaussian mixture models and simple image features, and an alternative method using probabilistic latent semantic analysis (PLSA) and the bag-of-features (BoF) representation that is regarded to have more semantically discriminative power than other representations. The authors represent regions using color, texture, and shape features, and then probabilistically associate regions with concepts. A generic statistical model for image region features is based on about 50 thousand regions randomly picked up from the gathered Web images. Also, the authors create a generic distribution of image features from about 10 thousand randomly picked web images using probabilistic latent semantic analysis (PLSA). Geolocation entropy is computed by dividing the world map into 36×36 grids (or 1296 bins) and making a probability distribution of geolocations of a given concept. The authors plan to explore cultural and regional differences, for example, how concept usage differs based on location, such as how Western-style houses are different from Asian-style ones and African-style ones.

Whereas the first two chapters focus on improving image retrieval, which is essential for consumer photo collections, the third chapter turns toward automated extraction and generation of semantically enriched models from three-dimensional scans. Sven Havemann from Graz Technical University and Torsten Ullrich and Dieter Fellner from Fraunhofer IGD and GRIS depart from three-dimensional scanning, similar to taking a photograph but with added depth dimension. The authors describe the most important techniques for retrieving semantics from such acquired three-dimensional data, including parametric, procedural, and generative 3D modeling. The authors illustrate these concepts with two active and challenging domains: urban reconstruction and cultural heritage. For example, their ambitious CityFIT project aims to automatically reconstruct 80% of the facades in the archi-

tecturally diverse city of Graz, Austria by statistically inferring facade templates from LIDAR sensor data. The authors' digital sampling of the world coupled with the augmentation of shapes with semantics is an essential step toward representing reality within a computer in a meaningful, ideally even editable, form. The authors' contribution is twofold: A scalable solution for model-based production of new models, and information extraction from existing models by means of automated fitting procedures.

Just as we need improved models for extracting meaningful shape from human created edifices, so too we want to extract meaning from human faces. In the fourth chapter, Nicolas Stoiber and Gaspard Breton from Orange Labs in France and Renaud Seguier from Supelec, France report reliable and accurate emotional facial feature extraction. Features can be used for identity or pattern recognition (e.g., a frown or smile). To bridge the gap between conceptual models of emotion and actual facial deformations, the authors present a high-level representation of emotional facial expressions based on empirical facial expression data. By identifying the basic emotions (e.g., sadness, joy, surprise, fear, anger, and disgust) on a simple two-dimensional colored disc interface, the facial representation remains true to the real world yet becomes intuitive, interpretable, and manipulatable. Accordingly, it has been successfully applied to facial analysis and synthesis tasks. In the former case of facial analysis, even unseen, mixed facial expressions not included in the original database are recovered. In the latter case, the representation is used for facial expression synthesis on a virtual character. While applied primarily to emotional facial expressions, the representation space could apply the analysis method to other types of expressions, like speech-related facial configurations (visemes).

Taken together, the image extraction chapters illustrate the range of important applications enabled by image extraction, from improved organization and retrieval in consumer collections to extraction of 3D models from city or historic buildings to improved facial emotion extraction and synthesis. While the authors make a number of important data, algorithmic, and method contributions, they also outline a number of remaining challenges including image query context, results presentation, and representation and reasoning about visual content.