CHAPTER 1

Introduction to Multimodal Scene Understanding

Michael Ying Yang[*], **Bodo Rosenhahn**[†], **Vittorio Murino**[‡]

[*] *University of Twente, Enschede, The Netherlands* [†] *Leibniz University Hannover, Hannover, Germany*
[‡] *Istituto Italiano di Tecnologia, Genova, Italy*

Contents

1.1 Introduction

While humans constantly extract meaningful information from visual data almost effortlessly, it turns out that simple visual tasks such as recognizing, detecting and tracking objects, or, more difficult, understanding what is going on in the scene, are extremely challenging problems for machines. To design artificial vision systems that can reliably process information as humans do has many potential applications in fields such as robotics, medical imaging, surveillance, remote sensing, entertainment or sports science, to name a few. It is therefore our ultimate goal to be able to emulate the human visual system and processing capabilities with computational algorithms.

Computer vision has contributed to a broad range of tasks to the field of artificial intelligence, such as estimating physical properties from an image, e.g., depth and motion, as well as estimating semantic properties, e.g., labeling each pixel with a semantic class. A fundamental goal of computer vision is to discover the semantic information within a given scene, namely, *understanding* a scene, which is the basis for many applications: surveillance, autonomous driving, traffic safety, robot navigation, vision-guided mobile navigation systems, or activity recognition. Understanding a scene from an image or a video requires much more than recording and extracting some features. Apart from visual information, humans make use of further sensor data, e.g. from audio signals, or acceleration. The net goal is to find a mapping

Multimodal Scene Understanding
https://doi.org/10.1016/B978-0-12-817358-9.00007-X

to derive semantic information from sensor data, which is an extremely challenging task partially due to the ambiguities in the appearance of the data. These ambiguities may arise either due to the physical conditions such as the illumination and the pose of the scene components, or due to the intrinsic nature of the sensor data itself. Therefore, there is the need of capturing local, global or dynamic aspects of the acquired observations, which are to be utilized to interpret a scene. Besides, all information which is possible to extract from a scene must be considered in context in order to get a comprehensive representation, but this information, while it is easily captured by humans, is still difficult to extract by machines.

Using big data leads to a big step forward in many applications of computer vision. However, the majority of scene understanding tasks tackled so far involve visual modalities only. The main reason is the analogy to our human visual system, resulting in large multipurpose labeled image datasets. The unbalanced number of labeled samples available among different modalities result in a big gap in performance when algorithms are trained separately [1]. Recently, a few works have started to exploit the synchronization of multimodal streams to transfer semantic information from one modality to another, e.g. RGB/Lidar [2], RGB/depth [3,4], RGB/infrared [5,6], text/image [7], image/Inertial Measurement Units (IMU) data [8,9].

This book focuses on recent advances in algorithms and applications that involve multiple sources of information. Its aim is to generate momentum around this topic of growing interest, and to encourage interdisciplinary interactions and collaborations between computer vision, remote sensing, robotics and photogrammetry communities. The book will also be relevant to efforts on collecting and analyzing multisensory data corpora from different platforms, such as autonomous vehicles [10], surveillance cameras [11], unmanned aerial vehicles (UAVs) [12], airplanes [13] and satellites [14]. On the other side, it is undeniable that deep learning has transformed the field of computer vision, and now rivals human-level performance in tasks such as image recognition [15], object detection [16], and semantic segmentation [17]. In this context, there is a need for new discussions as regards the roles and approaches for multisensory and multimodal deep learning in the light of these new recognition frameworks.

In conclusion, the central aim of this book is to facilitate the exchange of ideas on how to develop algorithms and applications for multimodal scene understanding. The following are some of the scientific questions and challenges we hope to address:

- What are the general principles that help in the fusion of multimodal and multisensory data?
- How can multisensory information be used to enhance the performance of generic high-level vision tasks, such as object recognition, semantic segmentation, localization, and scene reconstruction, and empower new applications?
- What are the roles and approaches of multimodal deep learning?

To address these challenges, a number of peer-reviewed chapters from leading researchers in the fields of computer vision, remote sensing, and machine learning have been selected. These chapters provide an understanding of the state-of-the-art, open problems, and future directions related to multimodal scene understanding as a relevant scientific discipline.

The editors sincerely thank everyone who supported the process of preparing this book. In particular, we thank the authors, who are among the leading researchers in the field of multimodal scene understanding. Without their contributions in writing and peer-reviewing the chapters, this book would not have been possible. We are also thankful to Elsevier for the excellent support.

1.2 Organization of the Book

An overview of each of the book chapters is given in the following.

Chapter 2: Multimodal Deep Learning for Multisensory Data Fusion

This chapter investigates multimodal encoder–decoder networks to harness the multimodal nature of multitask scene recognition. In its position regarding the current state of the art, this work was distinguished by: (1) the use of the U-net architecture, (2) the application of translations between all modalities of the learning package and the use of monomodal data, which improves intra-modal self-encoding paths, (3) the independent mode of operation of the encoder–decoder, which is also useful in the case of missing modalities, and (4) the image-to-image translation application managed by more than two modalities. It also improves the multitasking reference network and automatic multimodal coding systems. The authors evaluate their method on two public datasets. The results of the tests illustrate the effectiveness of the proposed method in relation to other work.

Chapter 3: Multimodal Semantic Segmentation: Fusion of RGB and Depth Data in Convolutional Neural Networks

This chapter investigates the fusion of optical multispectral data (red-green-blue or near infrared-red-green) with 3D (and especially depth) information within a deep learning CNN framework. Two ways are proposed to use 3D information: either 3D information is directly introduced into the classification fusion as a depth measure or information about normals is estimated and provided as input to the fusion process. Several fusion solutions are considered and compared: (1) Early fusion: RGB and depth (or normals) are merged before being provided to the CNN. (2) RGB and depth (or normals) are simply concatenated and directly provided to common CNN architectures. (3) RGB and depth (or normals) are provided as two distinct inputs to a Siamese CNN dedicated to fusion. Such methods are tested on two benchmark datasets: an indoor terrestrial one (Stanford) and an aerial one (Vaihingen).

Chapter 4: Learning Convolutional Neural Networks for Object Detection with Very Little Training Data

This chapter addresses the problem of learning with very few labels. In recent years, convolutional neural networks have shown great success in various computer vision tasks, whenever they are trained on large datasets. The availability of sufficiently large labeled data, however, limits possible applications. The presented system for object detection is trained with very few training examples. To this end, the advantages of convolutional neural networks and random forests are combined to learn a patch-wise classifier. Then the random forest is mapped to a neural network and the classifier is transformed to a fully convolutional network. Thereby, the processing of full images is significantly accelerated and bounding boxes can be predicted. In comparison to the networks for object detection or algorithms for transfer learning, the required amount of labeled data is considerably reduced. Finally, the authors integrate GPS-data with visual images to localize the predictions on the map and multiple observations are merged to further improve the localization accuracy.

Chapter 5: Multimodal Fusion Architectures for Pedestrian Detection

In this chapter, a systematic evaluation of the performances of a number of multimodal feature fusion architectures is presented, in the attempt to identify the optimal solutions for pedestrian detection. Recently, multimodal pedestrian detection has received extensive attention since the fusion of complementary information captured by visible and infrared sensors enables robust human target detection under daytime and nighttime scenarios. Two important observations can be made: (1) it is useful to combine the most commonly used concatenation fusion scheme with a global scene-aware mechanism to learn both human-related features and correlation between visible and infrared feature maps; (2) the two-stream semantic segmentation without multimodal fusion provides the most effective scheme to infuse semantic information as supervision for learning human-related features. Based on these findings, a unified multimodal fusion framework for joint training of semantic segmentation and target detection is proposed, which achieves state-of-the-art multispectral pedestrian detection performance on the KAIST benchmark dataset.

Chapter 6: ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset

This chapter deals with color-thermal cross-modality person re-identification (Re-Id). This topic is still challenging, in particular for video surveillance applications. In this context, it is demonstrated that conditional generative adversarial networks are effective for cross-modality prediction of a person appearance in thermal image conditioned by a probe color image. Discriminative features can be extracted from real and synthesized thermal images for effective matching of thermal signatures. The main observation is that thermal cameras coupled with

generative adversarial network (GAN) Re-Id framework can significantly improve the Re-Id performance in low-light conditions. A ThermalGAN framework for cross-modality person Re-Id in the visible range and infrared images is so proposed. Furthermore, a large-scale multispectral ThermalWorld dataset is collected, acquired with FLIR ONE PRO cameras, usable both for Re-Id and visual objects in context recognition.

Chapter 7: A Review and Quantitative Evaluation of Direct Visual–Inertia Odometry

This chapter combines complementary features of visual and inertial sensors to solve direct sparse visual–inertial odometry problem in the field of simultaneous localization and mapping (SLAM). By introducing a novel optimization problem that minimizes camera geometry and motion sensor errors, the proposed algorithm estimates camera pose and sparse scene geometry precisely and robustly. As the initial scale can be very far from the optimum, a technique is proposed called dynamic marginalization, where multiple marginalization priors and constraints on the maximum scale difference are considered. Extensive quantitative evaluation on the EuRoC dataset demonstrates that the described visual–inertial odometry method outperforms other state-of-the-art methods, both the complete system as well as the IMU initialization procedure.

Chapter 8: Multimodal Localization for Embedded Systems: A Survey

This chapter presents a survey of systems, sensors, methods, and application domains of multimodal localization. The authors introduce the mechanisms of various sensors such as inertial measurement units (IMUs), global navigation satellite system (GNSS), RGB cameras (with global shutter and rolling shutter technology), IR and Event-based cameras, RGB-D cameras, and Lidar sensors. It leads the reader to other survey papers and thus covers the corresponding research areas exhaustively. Several types of sensor fusion methods are also illustrated. Moreover, various approaches and hardware configurations for specific applications (e.g. autonomous mobile robots) as well as real products (such as Microsoft Hololens and Magic Leap One) are described.

Chapter 9: Self-supervised Learning from Web Data for Multimodal Retrieval

This chapter addresses the problem of self-supervised learning from image and text data which is freely available from web and social media data. Thereby features of a convolutional neural network can be learned without requiring labeled data. Web and social media platforms provide a virtually unlimited amount of this multimodal data. This free available bunch of data is then exploited to learn a multimodal image and text embedding, aiming to leverage the semantic knowledge learned in the text domain and transfer it to a visual model for semantic image retrieval. A thorough analysis and performance comparisons of five different state-of-the-art text embeddings in three different benchmarks are reported.

Chapter 10: 3D Urban Scene Reconstruction and Interpretation from Multisensor Imagery

This chapter presents an approach for 3D urban scene reconstruction based on the fusion of airborne and terrestrial images. It is one step forward towards a complete and fully automatic pipeline for large-scale urban reconstruction. Fusion of images from different platforms (terrestrial, UAV) has been realized by means of pose estimation and 3D reconstruction of the observed scene. An automatic pipeline for level of detail 2 building model reconstruction is proposed, which combines a reliable scene and building decomposition with a subsequent primitive-based reconstruction and assembly. Level of detail 3 models are obtained by integrating the results of facade image interpretation with an adapted convolutional neural network (CNN), which employs the 3D point cloud as well as the terrestrial images.

Chapter 11: Decision Fusion of Remote Sensing Data for Land Cover Classification

This chapter presents a framework for land cover classification by late decision fusion of multimodal data. The data include imagery with different spatial as well as temporal resolution and spectral range. The main goal is to build a practical and flexible pipeline with proven techniques (i.e., CNN and random forest) for various data and appropriate fusion rules. The different remote sensing modalities are first classified independently. Class membership maps calculated for each of them are then merged at pixel level, using decision fusion rules, before the final label map is obtained from a global regularization. This global regularization aims at dealing with spatial uncertainties. It relies on a graphical model, involving a fit-to-data term related to merged class membership measures and an image-based contrast sensitive regularization term. Two use cases demonstrate the potential of the work and limitations of the proposed methods are discussed.

Chapter 12: Cross-modal Learning by Hallucinating Missing Modalities in RGB-D Vision

Diverse input data modalities can provide complementary cues for several tasks, usually leading to more robust algorithms and better performance. This chapter addresses the challenge of how to learn robust representations leveraging multimodal data in the training stage, while considering limitations at test time, such as noisy or missing modalities. In particular, the authors consider the case of learning representations from depth and RGB videos, while relying on RGB data only at test time. A new approach to training a hallucination network has been proposed that learns to distill depth features through multiplicative connections of spatio-temporal representations, leveraging soft labels and hard labels, as well as distance between feature maps. State-of-the-art results on the video action classification dataset are reported.

Note: The color figures will appear in color in all electronic versions of this book.

References

[1] A. Nagrani, S. Albanie, A. Zisserman, Seeing voices and hearing faces: cross-modal biometric matching, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[2] M.Y. Yang, Y. Cao, J. McDonald, Fusion of camera images and laser scans for wide baseline 3D scene alignment in urban environments, ISPRS Journal of Photogrammetry and Remote Sensing 66 (6S) (2011) 52–61.

[3] A. Krull, E. Brachmann, F. Michel, M.Y. Yang, S. Gumhold, C. Rother, Learning analysis-by-synthesis for 6d pose estimation in rgb-d images, in: IEEE International Conference on Computer Vision, ICCV, 2015.

[4] O. Hosseini, O. Groth, A. Kirillov, M.Y. Yang, C. Rother, Analyzing modular cnn architectures for joint depth prediction and semantic segmentation, in: International Conference on Robotics and Automation, ICRA, 2017.

[5] M.Y. Yang, Y. Qiang, B. Rosenhahn, A global-to-local framework for infrared and visible image sequence registration, in: IEEE Winter Conference on Applications of Computer Vision, 2015.

[6] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: IEEE International Conference on Computer Vision, ICCV, 2017.

[7] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: justifying decisions and pointing to the evidence, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[8] C. Reinders, H. Ackermann, M.Y. Yang, B. Rosenhahn, Object recognition from very few training examples for enhancing bicycle maps, in: IEEE Intelligent Vehicles Symposium, IV, 2018, pp. 1–8.

[9] T. von Marcard, R. Henschel, M.J. Black, B. Rosenhahn, G. Pons-Moll, Recovering accurate 3d human pose in the wild using imus and a moving camera, in: European Conference on Computer Vision, ECCV, 2018, pp. 614–631.

[10] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The kitti vision benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2012.

[11] S. Oh, A. Hoogs, A.G.A. Perera, N.P. Cuntoor, C. Chen, J.T. Lee, S. Mukherjee, J.K. Aggarwal, H. Lee, L.S. Davis, E. Swears, X. Wang, Q. Ji, K.K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A.K. Roy-Chowdhury, M. Desai, A large-scale benchmark dataset for event recognition in surveillance video, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 3153–3160.

[12] F. Nex, M. Gerke, F. Remondino, H. Przybilla, M. Baumker, A. Zurhorst, Isprs benchmark for multi-platform photogrammetry, in: Annals of the Photogrammetry, Remote Sensing and Spatial Information Science, 2015, pp. 135–142.

[13] Z. Zhang, M. Gerke, G. Vosselman, M.Y. Yang, A patch-based method for the evaluation of dense image matching quality, International Journal of Applied Earth Observation and Geoinformation 70 (2018) 25–34.

[14] X. Han, X. Huang, J. Li, Y. Li, M.Y. Yang, J. Gong, The edge-preservation multi-classifier relearning framework for the classification of high-resolution remotely sensed imagery, ISPRS Journal of Photogrammetry and Remote Sensing 138 (2018) 57–73.

[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, NIPS, 2012, pp. 1097–1105.

[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, NIPS, 2015, pp. 91–99.

[17] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.