Report from Dagstuhl Seminar 15081

Holistic Scene Understanding

Edited by

Jiří Matas[1], Vittorio Murino[2], Laura Leal-Taixé[3], and Bodo Rosenhahn[5]

1 **Czech Technical University, CZ,** matas@cmp.felk.cvut.cz
2 **Italian Institute of Technology – Genova, IT,** vittorio.murino@iit.it
3 **ETH Zürich, CH,** leal@geod.baug.ethz.ch
5 **Leibniz Universität Hannover, DE,** rosenhahn@tnt.uni-hannover.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15081 "Holistic Scene Understanding". During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Overall, the seminar was a great success, which is also reflected in the very positive feedback we received from the evaluation.

Seminar February 15–20, 2015 – http://www.dagstuhl.de/15081
1998 ACM Subject Classification I.2.6 Learning, I.2.10 Vision
Keywords and phrases Scene Analysis, Image Understanding, Crowd Analysis, People and Object Recognition
Digital Object Identifier 10.4230/DagRep.5.2.80
Edited in cooperation with Roberto Henschel

1 Executive Summary

Jiří Matas
Vittorio Murino
Laura Leal-Taixé
Bodo Rosenhahn

Motivations

To *understand* a scene in a given image or video is much more than to *simply* record and store it, extract some features and eventually recognize an object. The overall goal is to find a mapping to derive semantic information from sensor data. Purposive Scene understanding may require a different representation for different specific tasks. The task itself can be used as prior but we still require an in-depth understanding and balancing between local, global and dynamic aspects which can occur within a scene. For example, an observer might be interested to understand from an image if there is a person present or not, and beyond that, if it is possible to look for more information, e.g. if the person is sitting, walking or raising a hand, etc.

When people move in a scene, the specific time (e.g. 7:30 in the morning, workdays, weekend), the weather (e.g. rain), objects (cars, a bus approaching a bus stop, crossing bikes,

Holistic Scene Understanding, *Dagstuhl Reports*, Vol. 5, Issue 2, pp. 80–108
Editors: Jiří Matas, Vittorio Murino, and Bodo Rosenhahn
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

etc.) or surrounding people (crowded, fast moving people) yield to a mixture of low-level and high-level, as well as abstract cues, which need to be jointly analyzed to get an in-depth understanding of a scene. In other words, generally speaking, the so-called *context* is to be considered for a comprehensive scene understanding, but this information, while it is easily captured by human beings, is still difficult to obtain from a machine.

Holistic scene interpretation is crucial to design the next generation of recognition systems, which are important for several applications, e.g. driver assistance, city modeling and reconstruction, outdoor motion capture and surveillance.

With such topics in mind, the aim of this workshop was to discuss which are the sufficient and necessary elements for a complete scene understanding, i.e. what it really means to *understand* a scene. Specifically, in this workshop, we wanted to explore methods that are capable of representing a scene at different level of semantic granularity and modeling various degrees of interactions between objects, humans and 3D space. For instance, a scene-object interaction describes the way a scene type (e.g. a dining room or a bedroom) influences objects' presence, and vice versa. An object-3D-layout or human-3D-layout interaction describes the way the 3D layout (e.g. the 3D configuration of walls, floor and observer's pose) biases the placement of objects or humans in the image, and vice versa. An object-object or object-human interaction describes the way objects, humans and their pose affect each other (e.g. a dining table suggests that a set of chairs are to be found around it). In other words, the 3D configuration of the environment and the relative placements and poses of the objects and humans therein, the associated dynamics (relative distance, human body posture and gesture, gazing, etc.), as well as other contextual information (e.g., weather, temperature, etc.) support the holistic understanding of the observed scene.

As part of a larger system, understanding a scene semantically and functionally allows to make predictions about the presence and locations of unseen objects within the space, and thus predict behaviors and activities that are yet to be observed. Combining predictions at multiple levels into a global estimate can improve each individual prediction.

Since most scenes involve humans, we were also interested in discussing novel methods for analyzing group activities and human interactions at different levels of spatial and semantic resolution. As advocated in recent literature, it is beneficial to solve the problem of tracking individuals and understand their activities in a joint fashion by combining bottom-up evidence with top-down reasoning as opposed to attack these two problems in isolation.

Top-down constraints can provide critical contextual information for establishing accurate associations between detections across frames and, thus, for obtaining more robust tracking results. Bottom-up evidence can percolate upwards so as to automatically infer action labels for determining activities of individual actors, interactions among individuals and complex group activities. But of course there is more than this, it is indeed the cooperation of both data flows that makes the inference more manageable and reliable in order to improve the comprehension of a scene.

We gathered researchers which are not only well-known in Computer Vision areas such as object detection, classification, motion segmentation, crowd and group behavior analysis or 3D scene reconstruction, but also Computer Vision affiliated people from other communities in order to share each others point of view on the common topic of scene understanding.

Goals

Our main goals of the seminar can be summarized as follows:

- Address holistic scene understanding, a topic that has not been discussed before in detail at previous seminars, with special focus on a multidisciplinary perspective for sharing or competing the different views.
- Gather well-known researchers from the Computer Vision, Machine Learning, Social Sciences (e.g. Cognitive Psychology), Neuroscience, Robotics and Computer Graphics communities to compare approaches to representing scene geometry, dynamics, constraints as well as problems and task formulations adopted in these fields. The interdisciplinary scientific exchange is likely to enrich the communities involved.
- Create a platform for discussing and bridging topics like perception, detection, tracking, activity recognition, multi-people multi-object interaction and human motion analysis, which are surprisingly treated independently in the communities.
- Publication of an LNCS post-proceedings as previously done for the 2006, 2008 and 2010 seminars. These will include the scientific contributions of participants of the Seminar, focusing specially on the discussed topics presented at the Seminar.

Organization of the seminar

During the workshop we discussed different modeling techniques and experiences researchers have collected. We discussed sensitivity, time performance and e.g. numbers of parameters required for special algorithms and the possibilities for context-aware adaptive and interacting algorithms. Furthermore, we had extensive discussions on open questions in these fields.

On the first day, the organizers provided general information about Dagstuhl seminars, the philosophy behind Dagstuhl and the expectations to the participants. We also clarified the kitchen-rules and organized a running-group for the early mornings (5 people participated frequently!).

Social event. On Wednesday afternoon we organized two afternoon event: One group made a trip to Trier, and another group went on a 3h hike in the environment.

Working Groups. To strongly encourage discussions during the seminar, we organized a set of working groups on the first day (with size between 8–12 people). As topics we selected

- What does “Scene Understanding” mean ?
- Dynamic Scene: Humans.
- Recognition in static scenes (in 3D).

There were two afternoon slots reserved for these working groups and the outcome of the working groups has been presented in the Friday morning session.

LNCS Post-Proceedings. We will edit a Post-Proceeding and invite participants to submit articles. In contrast to standard conference articles, we allow for more space (typically 25 single-column pages) and allow to integrate open questions or preliminary results, ideas, etc. from the seminar into the proceedings. Additionally, we will enforce joint publications of participants who started to collaborate after the seminar. All articles will be reviewed by at least two reviewers and based on the evaluation, accepted papers will be published. We will publish the proceeding at the Lecture Notes in Computer Science (LNCS-Series) by Springer. The papers will be collected during the summer months.

Overall, it was a great seminar and we received very positive feedback from the participants. We would like to thank castle Dagstuhl for hosting the event and are looking forward to revisit Dagstuhl whenever possible.

2 Table of Contents

3 Overview of Talks

3.1 Superhuman-in-the-Loop Computer Vision

Gabriel Brostow (University College London, GB)

Joint work of Mac Aodha, Oisin; Stathopoulos, Vassilios; Terry, Michael; Jones, Kate E.; Brostow, Gabriel J.; Girolami, Mark

Main reference O. Mac Aodha, V. Stathopoulos, G. J. Brostov, M. Terry, M. A. Girolami, K. E. Jones, "Putting the Scientist in the Loop – Accelerating Scientific Progress with Interactive Machine Learning," in Proc. of the 22nd Int'l Conf. on Pattern Recognition, pp. 9–17, IEEE, 2014; pre-print available from author's webpage.

URL http://dx.doi.org/10.1109/ICPR.2014.12

URL http://web4.cs.ucl.ac.uk/staff/g.brostow/papers/engage_icpr_2014.pdf

Despite ubiquitous computing, most normal people are not benefiting from advancements in computer vision research. Equally, most vision systems do not improve with time or learn from their users' experience. This is a terrible waste, but is understandable: there are plenty of specific vision problems where progress a) can be made "offline" in labs trying to beat a recognized benchmark score, and b) the specific problem affects a big industry, like scene-flow for cars, or image-retrieval for search engines.

In this talk, I advocate that we should be aiming for responsive algorithms, and that these should be measured in terms of accuracy-improvement, and the user's ability to perform their specific tasks. This means we will need new benchmarks, and that we need to engage with real users for our models and experiments to be meaningful. While my group has started making software that adapts to specialist users, e.g. biologists/zoologists, the ageing population is just one mass-scale cohort that will require new computer vision models and interfaces.

3.2 From groups to crowds: a social signal processing perspective

Marco Cristani (University of Verona, IT)

After years of research on automated analysis of individuals, the computer vision community has shifted its attention towards the new issues of modeling groups and crowds. Within the scope of computer vision, groups are generally defined simply as two or more people moving at a similar velocity, spatially and temporally close to one another. However, things are a bit more complex: there are many kinds of groups, that differ in dimension, durability (ephemeral, ad hoc or stable groups), in/formality of organization, degree of "sense of belonging", level of physical dispersion etc.. Along the same lines, crowds are usually intended as a large number of persons gathered closely together; but, even in this case, the notion of crowd is much more complex and requires a more detailed account, which is basically missing in the computer vision community. In this talk, we build on concepts inherited from the sociological analysis and we offer a detailed taxonomy of groups and crowds. As we will see, this analysis individuates many typologies of social gatherings, each with its own characteristics and behavior dynamics. These differences are not only useful for a mere classification purpose, but are crucial when the need of automatic modeling comes into play, eliging particular computer vision techniques and models as the most appropriate to account for such differences. In this

talk, in particular, we will focus on a specific kind of group, i.e. free-standing conversational group, and one kind of crowd, i.e. spectator crowd, showing recent advancements in their automatic modeling.

3.3 Semantic Motion Segmentation and 3D Reconstruction

Alessio Del Bue (Italian Institute of Technology–Genova, IT)

Joint work of Del Bue, Alessio; Crocco, Marco; Rubino, Cosimo
Main reference C. Rubino, M. Crocco, A. Perina, V. Murino, A. Del Bue, "3D Structure from Detections," arXiv:1502.04754v2 [cs.CV], 2015.
URL http://arxiv.org/abs/1502.04754v2

In this talk I will present how to embed semantic information in classical multi-view geometry problems such as multi-body motion segmentation. The key feature is the explicit inclusion of geometrical priors given by general purpose object detectors that boost the segmentation of the moving objects. In the classical formulation of the problem, only 2D matched points between views are used to identify independently moving objects leveraging the principle that a set of points belonging to a moving object would satisfy some given multi- view relations (e.g. multi-body epipolar constraints). We improve and speedup motion segmentation by including the information that a set of 2D matches may belong to the same object given the output of a detector. As such, instead of sampling points uniformly with a RANSAC based strategy, the selection of the matches is driven by the position and score confidence of the object detectors. After some experimental evidence, conclusions will show that other problems, (e.g. 3D reconstruction) can be supported by the inclusion of semantic information extracted from a generic scene.

3.4 Feature Regression-based Pose Estimation for Object Categories

Michele Fenzi (Leibniz Universität Hannover, DE)

Joint work of Fenzi, Michele; Leal-Taixé, Laura; Rosenhahn, Bodo; Ostermann, Jörn
Main reference M. Fenzi, L. Leal-Taixé, B. Rosenhahn, J. Ostermann, "Class Generative Models based on Feature Regression for Pose Estimation of Object Categories," in 2013 IEEE Conf. on ", IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13), pp. 755-762, IEEE, 2013.
URL http://dx.doi.org/10.1109/CVPR.2013.103

I present an approach to pose estimation for object categories based on feature regression [1]. Pose estimation for object categories is becoming increasingly important and of interest, both as a fundamental part of larger tasks or as a standalone challenge. Among the many approaches proposed in literature, those based on local features have shown to work effectively. While some use explicit 3D information, others have shown that coupling feature regression and view labelling is enough to solve this task. I present a method for learning a class representation and a pose estimation algorithm that returns a continuous value for the pose of an unknown class instance using only 2D data and weak labelling information. Our method is based on generative feature models, i.e., regression functions learnt from local descriptors of the same patch collected under different viewpoints. The individual generative models are then clustered in order to create class generative models which form the class representation.

At run-time, geometric consistency is introduced in the matching step by means of a graph matching strategy [2]. Finally, the pose of the query image is estimated in a probabilistic fashion by combining the regression functions belonging to the matching clusters.

References

1 Michele Fenzi, Laura Leal-Taixé, Bodo Rosenhahn, Jörn Ostermann, "Class Generative Models based on Feature Regression for Pose Estimation of Object Categories", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, USA, June 2013

2 Michele Fenzi, Jörn Ostermann, "Embedding Geometry in Generative Models for Pose Estimation of Object Categories", British Machine Vision Conference (BMVC), Nottingham, United Kingdom, September 2014

3.5 Fish detection, tracking, recognition and analysis with the Fish4Knowledge Dataset

Bob Fisher (University of Edinburgh, GB)

Main reference B. J. Boom, J. He, S. Palazzo, P. X. Huang, C. Beyan, H.-M. Chou, F.-P. Lin, C. Spampinato, R. B. Fisher, "A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage," Ecological Informatics, Vol. 23, pp. 83–97, 2014.

URL http://dx.doi.org/10.1016/j.ecoinf.2013.10.006

The research presented here was based on the data collected by the EU funded Fish4Knowledge project. Altogether, 80 Tb of video was recorded from 9 cameras, resulting in 90,000 hours of video. From this data, 1.4 billion fish observations were detected and tracked, resulting in 145 million trajectories. The individuals from these trajectories were then classified into 23 species. Some of the interesting results observed in this scene data was:

1. Much of the data was degraded, due to water quality after storms, algae growing on the lenses and compression artifacts.
2. The distribution of fish species was greatly unbalanced, resulting in e.g. 50 times greater observation frequencies for the most common species relative to the 10th most commonly observed fish.
3. Because of similarities in appearance, a hierarchical classifier gave the best species recognition performance, with 97% recognition accuracy over all fish and 75% when averaged over the top 15 classes.
4. Although each observation varied by the individual behaviour and noise in the data, having millions of observations allowed conclusion to be made – eg. that there appear to be more fish observed at the ends of the day than in the middle, and the swimming speed of D. reticulatus increases as the water temperature increases.
5. Another interesting result presented concerned the ability to recognise individual fish. Using a collection of colour and texture properties, it was possible to strongly cluster A. clarkii (clownfish) observations. Although no ground truth was available, the fish in the clusters were highly similar in appearance and different from the fish in other clusters.

3.6 Human Pose: A Cue for Activities and Objects

Jürgen Gall (Universität Bonn, DE)

Joint work of Gall, Jürgen; Jhuang, Hueihan; Zuffi, Silvia; Schmid, Cordelia; Black, Michael J.; Srikantha, Abhilash
URL http://ps.is.tuebingen.mpg.de/project/JHMDB

In this talk, I discuss human pose as a cue for action recognition and object discovery. In order to allow a systematic performance evaluation of an action recognition pipeline, we annotated human joints for the HMDB dataset (J-HMDB). The annotation can be used to systematically replace the output of various algorithms in an existing pipeline with ground truth data to analyze the components with the highest potential for improving the recognition accuracy. For example, is it worth to invest more time on improving low-level algorithms like optical flow, is the image location of the human performing the action important, or would knowledge about human pose be helpful? Given pose and activities, we can also reason about objects. For example, small objects can be discovered from videos.

References

1 Hueihan Jhuang, Jürgen Gall, Silvia Zuffi, Cordelia Schmid and Michael J. Black. *Towards Understanding Action Recognition*. ICCV, 2013
2 Abhilash Srikantha and Jürgen Gall. *Discovering Object Classes from Activities*. ECCV, 2014

3.7 Explore Multi-source Information for Holistic Visual Search

Shaogang Gong (Queen Mary University of London, GB)

Joint work of Gong, Shaogang; Xiang, Tao; Hospedales, Tim; Loy, Chen Change; Zheng, Wei-Shi; Fu, Yanwei

For making sense of big visual data captured by large scale distributed multi-cameras in urban environments, understanding human activities and behaviour, detecting and searching their whereabouts in crowded spaces are required. In this talk, I will present some recent progress on exploring multiple information sources for modelling holistic context-aware object detection and activity profiling in large volumes of surveillance video data, addressing the problems of abnormal behaviour/action/event detection in public spaces and person re-identification in large scale distributed CCTV camera networks. In particular, I will discuss the needs and open-questions on a number of model learning challenges, including: learning visual context of activity for abnormal behaviour discovery and object association context for increasing detection robustness; exploring human-in-the-loop active learning for overcoming sparse labelling information in person re-identification and minimising false detection in anomaly detection; exploring crowd information for learning space-time camera network topology in disjoint multi-camera person tracking; and exploring semantic structure (e.g. the WordNet semantic space) for Zero-Shot-Learning in object recognition.

References

1 Gong, Xiang. *Visual Analysis of Behaviour: From Pixels to Semantics*, Springer, May 2011 http://dx.doi.org/10.1007/978-0-85729-670-2

2 Gong, Cristani, Yan, Loy. *Person Re-Identification*, Springer, January 2014 http://dx.doi.org/10.1007/978-1-4471-6296-4
3 Fu, Hospedales, Xiang, Gong. *Transductive Multi-View Zero-Shot Learning*. IEEE TPAMI, in press, 2015. http://dx.doi.org/10.1109/TPAMI.2015.2408354

3.8 VocMatch: Efficient Multiview Correspondence for Structure from Motion

Michal Havlena (ETH Zürich, CH)

Joint work of Havlena, Michal; Schindler, Konrad
Main reference M. Havlena, K. Schindler, "VocMatch: Efficient Multiview Correspondence for Structure from Motion," in Proc. of the 13th Europ. Conf. on Computer Vision (ECCV'14), LNCS, Vol. 8691, pp. 46–60, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-10578-9_4

Feature matching between pairs of images is a main bottleneck of structure-from-motion computation from large, unordered image sets. We propose an efficient way to establish point correspondences between all pairs of images in a dataset, without having to test each individual pair. The principal message is that, given a sufficiently large visual vocabulary, feature matching can be cast as image indexing, subject to the additional constraints that index words must be rare in the database and unique in each image. We demonstrate that the proposed matching method, in conjunction with a standard inverted file, is 2-3 orders of magnitude faster than conventional pairwise matching. The proposed vocabulary-based matching has been integrated into a standard SfM pipeline, and delivers results similar to those of the conventional method in much less time.

3.9 Solving Multiple People Tracking Using Minimum Cost Arborescences

Roberto Henschel (Leibniz Universität Hannover, DE)

Joint work of Henschel, Roberto; Leal-Taixé, Laura; Rosenhahn, Bodo
Main reference R. Henschel, L. Leal-Taixé, B. Rosenhahn, "Efficient Multiple People Tracking Using Minimum Cost Arborescences," in Proc. of the 36th German Conf. on Pattern Recognition (GCPR'14), LNCS, Vol. 8753, pp. 265–276, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-11752-2_21

Current state-of-the-art multiple people tracker that use a hierarchical tracklet framework are prone to error propagation due to wrong decisions in ambiguous situations.

I will thus present a new type of tracklets, which are called tree tracklets, that helps to solve this issue. A tree tracklet contains bifurcations to naturally model ambiguous situations. The optimal data association is derived from a minimum cost arborescence in an acyclic directed graph. Thereby a solution of the association problem is obtained in linear time. I will present experiments on six well-known multiple people tracking datasets showing the good performance compared to state-of-the art tracking algorithms.

3.10 Generic Object Detection in Video using Saliency and Tracking

Esther Horbert (RWTH Aachen, DE)

Joint work of Horbert, Esther; M. Garcia, German; Frintrop, Simone; Leibe, Bastian
Main reference E. Horbert, G. M. García, S. Frintrop, B. Leibe, "Sequence-Level Object Candidates Based on Saliency for Generic Object Recognition on Mobile Systems," in Proc. of the 2015 IEEE Int'l Conf. on Robotics and Automation (ICRA'15), to appear; pre-print available from author's webpage.
URL http://www.vision.rwth-aachen.de/projects/kod/horbert-icra15-preprint

We propose a novel approach for generating generic object candidates for object discovery and recognition in continuous monocular video. Such candidates have recently become a popular alternative to exhaustive window-based search as basis for classification. Contrary to previous approaches, we address the candidate generation problem at the level of entire video sequences instead of at the single image level. We propose a processing pipeline that starts from individual region candidates and tracks them over time. This enables us to group candidates for similar objects and to automatically filter out inconsistent regions. For generating the per-frame candidates, we introduce a novel multi-scale saliency approach that achieves a higher per-frame recall with fewer candidates than current state-of-the-art methods. Taken together, those two components result in a significant reduction of the number of object candidates compared to frame level methods, while keeping a consistently high recall.

3.11 A Spectral Perspective on Invariant Measures of Shapes

Ron Kimmel (Technion–Haifa, IL)

Joint work of Aflalo, Yonathan; Brezis, Haim; Kimmel, Ron
Main reference On the optimality of shape and data representation in the spectral domain
URL http://www.cs.technion.ac.il/~ron/publications.html

We explore the power of the Laplace Beltrami Operator (LBO) in processing and analyzing visual and geometric information. The decomposition of the LBO at one end, and the heat operator at the other end provide us with efficient tools for dealing with images and shapes. Denoising, segmenting, filtering, exaggerating are just few of the problems for which the LBO provides a solution. We will review the optimality of a truncated basis provided by the LBO, and a selection of relevant metrics by which such optimal bases are constructed. Specific example is the scale invariant metric for surfaces that we argue to be a natural selection for the study of articulated shapes and forms.

3.12 Visual Odometry and 3D Roadside Reconstruction

Reinhard Klette (University of Auckland, NZ)

Joint work of Klette, Reinhard; Johnny Chien; Haokun Geng; Simon Hermann; Waqar Khan; Sandino Morales; Radu Nicolescu

Future cars might possibly contribute to some kind of incremental 3D roadside reconstruction, supporting a spatio-temporal model of road environments. Besides accurate car trajectory calculation and provision of accurate depth data, it is also necessary to unify multiple runs in a uniform world coordinate system, to ignore data caused by dynamic or transient static objects, and to be forgiving when reconstructing secondary surfaces such as of plants.

The talk informed about data recording in a test vehicle (trinocular, 16 bit per pixel, 2046 x 1080, 30 Hz), what results into 126 GB in just about 5.5 minutes of recording. The generated 3D roadside data can be used to enhance 3D city scene reconstructions obtained by aerial mapping (illustrated by a 2007 example of reconstructing the Sony Centre in Berlin with 7 cm ground-sample accuracy; courtesy by H. Hirschmueller and K. Scheibe). Current iOS Maps do not yet have the accuracy required for roadside models. Stereo matchers such as iSGM (S. Hermann and R. Klette, winner of Robust Vision Challenge at ECCV 2012) or linBPM (W. Khan and R. Klette, IEEE IV 2013) may be considered to be satisfactory tools for providing 3D roadside depth data.

The talk suggested a definition for "robustness", considering the sum of accuracies for "challenging" scenarios times the probabilities of such scenarios. For defining the accuracy for one scenario, the third-eye technology (S. Morales and R. Klette, CAIP 2009) might be considered as an option, using an NCC measure defined by pixels closed to image discontinuities.

Examples (four short sequences of just 400 input stereo frames) illustrated that the performance of stereo matchers depends on the input data, characterised by complexity of scene geometry, weather or lighting conditions, traffic density, and so forth.

Generating visually satisfying roadside geometry requires dense depth maps with "visually accurate" occlusion edges; methods designed in (D. Liu and R. Klette, The Visual Computer, 2015) might be considered for depth map corrections.

The baseline algorithm (SfM: apply visual odometry for mapping clouds of points into a uniform world coordinate system) can be enhanced by various considerations. First, depth is more accurate closer to cameras, and road geometry further away can be considered later, when the car arrives there. However, far-away depth values already provide a useful approximation of the expected scene geometry. NCC values (third-eye technology) also provide weights for obtained depth data, to be considered when integrating into an already existing surface model.

Two options have been considered for visual odometry, sparse bundle adjustment (with simplifying 3D surface representations by depth-based representations only; see J. Chien, H. Geng, R. Klette, 2015, submitted) or feature-matching also using GPS data (with iconic Kalman filters, and an extended Kalman filter for the overall movement; see H. Geng, J. Chien, R. Nicolescu, and R. Klette, 2015, submitted). Tracked features have been studied for invariance properties, showing invariance issues for all the studied features (including SIFT and SURF).

The required accuracy for ego-motion detection, and the automated unification of noisy multi-run 3D data into one spatio-temporal roadside model appear to be the two most challenging subjects in this area.

3.13 Learning an image-based motion context for multiple people tracking

Laura Leal-Taixé (ETH Zürich, CH)

Joint work of Leal-Taixé, Laura; Fenzi, Michele; Kuznetsova, Alina; Savarese, Silvio; Rosenhahn, Bodo
Main reference L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, S. Savarese, "Learning an image-based motion context for multiple people tracking," in Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14), pp. 3542–3549, IEEE, 2014.
URL http://dx.doi.org/10.1109/CVPR.2014.453

In this talk, I presented a novel method for multiple people tracking that leverages a generalized model for capturing interactions among individuals. At the core of our model lies a learned dictionary of interaction feature strings which capture relationships between the motions of targets. These feature strings, created from low-level image features, lead to a much richer representation of the physical interactions between targets compared to hand-specified social force models that previous works have introduced for tracking. One disadvantage of using social forces is that all pedestrians must be detected in order for the forces to be applied, while our method is able to encode the effect of undetected targets, making the tracker more robust to partial occlusions. The interaction feature strings are used in a Random Forest framework to track targets according to the features surrounding them. Results on six publicly available sequences show that our method outperforms state-of-the-art approaches in multiple people tracking.

3.14 Image Annotation – From Machine Learning to Machine Teaching

Oisin Mac Aodha (University College London, GB)

Joint work of Mac Aodha, Oisin; Campbell, Neill DF; Johns, Edward; Kautz, Jan; Brostow, Gabriel J
Main reference O. Mac Aodha, N. D. F. Campbell, J. Kautz, G. J. Brostow, "Hierarchical subquery evaluation for active learning on a graph," in Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14), pp. 564–571, IEEE, 2014
URL http://visual.cs.ucl.ac.uk/pubs/graphActiveLearning/index.html

Current state-of-the classification and detection algorithms used in scene understanding require large quantities of labeled training data. Typically, these datasets are manually annotated–at a great expense, both in terms of time and money. The goal of Active Learning is to reduce this effort by only requesting labels for the most informative data. In practice however, many Active Learning strategies are computationally inefficient, rendering them useless for interactive labeling. Another problem arises if the human annotator does not possess the knowledge or experience required to correctly label the data. This problem exists in many domains, such as species classification and image geolocalization, where fine-grained visual discrimination is essential for effective image understanding.

In this talk I will present an algorithm for efficient Active Learning in the context of semi-supervised graph based learning. Inspired by this, I will also present very recent work that aims to improve the ability of annotators by turning the human into the learner, and the algorithm into the teacher.

3.15 Structured Models for Recognition: Towards Sub-Category and Interaction Discovery

Greg Mori (Simon Fraser University–Burnaby, CA)

Joint work of Mori, Greg; Khodabandeh, Mehran; Lan, Tian; Vahdat, Arash; Zhou, Guang-Tong; Kim, Ilseo; Oh, Sangmin; Sigal, Leonid

Visual recognition involves reasoning about structured relations between objects at multiple levels of detail. For example, human behaviour analysis requires a comprehensive labeling covering individual low-level actions to pair-wise interactions through to high-level events. Scene understanding can benefit from considering visual sub-categories and their relations. In this talk I will present structured models for scenes and group activity recognition, with holistic analysis of people interacting and taking different social roles. I will describe our work on latent max-margin clustering as a means to discover sub-categories of objects and types of interactions between people.

3.16 Towards Holistic Scene Understanding: low-level cue extraction, collective activity classification, and behavior recognition

Vittorio Murino (Italian Institute of Technology–Genova, IT)

Joint work of Murino, Vittorio; Cristani, Marco; Del Bue, Alessio; Nabi, Moin; Crocco, Marco; Zanotto, Matteo; et al, et al;

In this talk I'll address the scene understanding problem from the perspective of analyzing and to figure out the human behavior. Humans are essentially a social species and interactions among conspecifics is one essential characteristics of the life: understanding such interplays constitutes an interesting yet important issue which can lead to effective applications and to discover diverse basic insights. Considering the "classic" computer vision paradigm, understanding behavior implies the design of a data processing pipeline composed by a high-level reasoning process which needs of low-level cues or "features". Such cues are, in this cases, detection, tracking and other algorithms able to identify a person and track over time while extracting other important features like, for instance, face orientation or gaze. Actually, tackling the understanding problem from a "social" perspective, means to look for the so called "social signals" which are formed by a set of nonverbal cues which have a social meaning. They can be conveyed in many way, primarily by means of the voice, gesture, posture, face and gazing, in which also the environment (e.g., the spatial location of people wrt the scene geometry) has a role.

In this context, I will briefly describe a couple of algorithms developed for low level cue extraction, namely 1) a face detection and orientation technique based on a powerful statistical descriptor (covariance, entropy and mutual information) associated to a boosting-based classification method, and 2) a group tracking algorithm based on the decentralized particle filtering and dirichlet process mixture model able to manage any number of groups and (group) split and merge events.

I will also show a couple of methods for high-level reasoning process. First, I will show an algorithm based on a new time-based descriptor, the temporal poselet, which is based on

poselet detectors extended to time, and able to detect people assemblies and classify collective activities. Second, I will address the problem of classifying mouse behavior. Starting from the outcome of an automatic algorithm able to track mice in a cage and classify atomic action frame by frame, we proposed a technique based on a Bayesian Nonparametric approach able to spot higher-level patterns of behaviors, that is a latent sequence of atomic actions which are not visible to neuroscientists but potentially useful to "understand" the genetically modified mouse behavior wrt control mice.

Finally, I will conclude the talk presenting a different analysis of a scene based on the acoustics. We build an acoustic camera (coupled with an optical one) able to sense "acoustically" the environment and identify relevant acoustic sources, classify the sound received and track targets of interest. This constitutes a complementary approach wrt to vision, indeed necessary and useful to reach a full understanding of a scene.

3.17 Interactions Between Scene Elements

Caroline Pantofaru (Google Inc.–Mountain View, US)

The physical world consists of interrelated objects, spaces, scenes and activities, whose positions in 3D all influence each other. When we see a scene we naturally understand which object is in front of another, which objects are grouped together to form a functional unit, how these objects relate to their space and what kind of space it is. At a glance we also understand people's spatial interactions and use this information to determine groups, movement patterns and activities. In visual media we understand where the cameras are positioned and how it affects their representation of the world. In this talk I discussed methods for using 3D reasoning and the relationships between semantic scene elements to improve our understanding of the visual world.

References

1 W. Choi, Y. Chao, C. Pantofaru, and S. Savarese. *Indoor Scene Understanding with Geometric and Semantic Contexts*. International Journal of Computer Vision, 2014.
2 W. Choi, Y. Chao, C. Pantofaru, and S. Savarese. *Discovering Groups of People in Images*. European Conference on Computer Vision, 2014.

3.18 Temporally Consistent Superpixels

Matthias Reso (Leibniz Universität Hannover, DE)

Joint work of Reso, Matthias;Jachalsky, Jörn; Rosenhahn, Bodo; Ostermann, Jörn
Main reference M. Reso, J. Jachalsky, B. Rosenhahn, J. Ostermann, "Temporally Consistent Superpixels," in Proc. of the 2013 IEEE Int'l Conf. on Computer Vision (ICCV'13), pp. 385–392, IEEE, 2013.
URL http://dx.doi.org/10.1109/ICCV.2013.55

Superpixel algorithms represent a very useful and in- creasingly popular preprocessing step for a wide range of computer vision applications, as they offer the potential to boost efficiency and effectiveness. This talk presents a highly competitive approach for temporally consistent superpixels for video content. The approach is based on energy-minimizing clustering utilizing

a novel hybrid clustering strategy for a multi-dimensional feature space working in a global color space and local spatial spaces. For a thorough evaluation the proposed approach is compared to state of the art supervoxel and video superpixel algorithms using established benchmarks and shows a superior performance.

References

1 Matthias Reso, Jörn Jachalsky, Bodo Rosenhahn and Jörn Ostermann. *Superpixels for Video Content Using a Contour-based EM Optimization*. ACCV, 2014

3.19 Describing Videos with Natural Language

Anna Rohrbach (MPI für Informatik–Saarbrücken, DE)

Joint work of Rohrbach, Anna; Rohrbach, Marcus; Qiu, Wei; Friedrich, Annemarie; Pinkal, Manfred; Schiele, Bernt
Main reference A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, B. Schiele, "Coherent Multi-sentence Video Description with Variable Level of Detail," In Proc. of the German Conference on Pattern Recognition (GCPR'14), LNCS, Vol. 8753, pp. 184–195, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-11752-2_15

Generating descriptions for video is an interesting task requiring core techniques of computer vision and computational linguistics. Existing approaches for automatic video description focus on generating single sentences at a single level of detail. We address both of these limitations: for a variable level of detail we produce coherent multi-sentence descriptions of complex videos featuring cooking activities [1]. We follow a two-step approach where we first learn to predict a semantic representation (SR) from video and then generate natural language descriptions from it. For our multi-sentence descriptions we model across-sentence consistency at the level of the SR by enforcing a consistent topic. To understand the difference between detailed and short descriptions, we collect and analyze a video description corpus with three levels of detail.

To foster the research on automatic video description we propose a new MPII Movie Description Dataset [2], featuring movie snippets aligned to scripts and DVS (Descriptive video service). DVS is a linguistic description that allows visually impaired people to follow a movie. We benchmark state-of-the-art computer vision algorithms to recognize scenes, human activities, and participating objects and achieve encouraging results in video description on this new challenging dataset.

References

1 A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. *Coherent multi-sentence video description with variable level of detail*. In Proceedings of the German Conference on Pattern Recognition (GCPR), 2014.
2 A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. *A dataset for movie description*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

3.20 Towards 3D scene understanding: how to assemble the pieces?

Konrad Schindler (ETH Zürich, CH)

Computer vision has made great progress. Individual low- and mid-level components like object detection, 3D reconstruction etc. are working well– yet, we have not managed to connect them into convincing "scene understanding" systems, and it is unclear how to tackle that next challenge. I argue that this might be in part due to the fact that we do not know what lower-level input we need for scene understanding. I will look at different basic visual functions and ask what information they should deliver in order to be useful for high-level semantic understanding. The talk will not provide answers, but will hopefully trigger a discussion about what individual pieces of the vision pipeline need to deliver to enable "scene understanding".

3.21 "At-a-glance" Visualization of Highlights in Raw Personal Videos

Min Sun (National Tsing Hua University – Hsinchu, TW)

Joint work of Sun, Min; Farhadi, Ali; Seitz, Steve Seitz
Main reference M. Sun, A. Farhadi, S. Seitz, "Ranking Domain-specific Highlights by Analyzing Edited Videos," in 13th Europ. Conf. on Computer Vision (ECCV'14), LNCS, Vol. 8689, pp. 787–802, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-10590-1_51
Main reference M. Sun, A. Farhadi, B. Taskar, S. Seitz, "Salient Montages from Unconstrained Videos," in 13th Europ. Conf. on Computer Vision (ECCV'14), LNCS, Vol. 8689, pp. 472–488, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-10584-0_31

Nowadays, people share tons of images and videos online. However, raw personal videos typically will rarely be watched again, since they are long and boring most of the time. Our research focuses on generating an "at-a-glance" visualization of highlights in raw person videos. We have two related work for this project. Our first work, "Ranking Domain-specific Highlights by Analyzing Edited Videos" focuses on automatically finding highlights in raw personal videos. Our second work, "Salient Montages from Unconstrained Videos" focuses on automatically generating an at-a-glance visualization (referred to as salient montages) of highlights in raw personal videos.

3.22 Detecting conversational groups: a game-theoretic approach with sociological and biological constraints

Sebastiano Vascon (Italian Institute of Technology–Genova, IT)

Joint work of Vascon, Sebastiano; Mequanint Z., Eyasu; Cristani, Marco ; Hung, Hayley; Pelillo, Marcello; Murino, Vittorio
Main reference S. Vascon, E. Zemene Mequanint, M. Cristani, H. Hung, M. Pelillo, V. Murino, "A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups," in Proc. of the 12th Asian Conf. on Computer Vision (ACCV'14), LNCS, Vol. 9007, pp. 658–675, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-16814-2_43

In the last decade the problem of detecting groups of people in a scene is gaining increasing interest due to its importance in many fields like video surveillance, social signal processing, scene understanding and social robotics to cite a few. In this talk, I have presented a recently published work to detect groups of persons that are conversing. The modelling of such groups is not a trivial task because persons create groups respecting certain biological and sociological constraints. We proposed a game-theoretic framework to detect groups in which these constraints are both satisfied, outperforming the state of the art.

References

1 Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, Vittorio Murino. *A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups*. The 12th Asian Conference on Computer Vision (ACCV 2014), Singapore, 2014

3.23 Understanding scenes on mobile devices

Stefan Walk (Qualcomm Austria Research Center GmbH, AT)

Understanding a scene is an important step towards being able to interact with the real world on mobile devices. Seeing recognized objects in their scene context is a prominent example for this. In this talk I will show how problems like this are solved at Qualcomm. Solutions include being able to interact with unknown objects on a plane initialized from a known object, geometrically understanding the planes that a scene is composed of, and being able to track a scene when the object of interest temporarily leaves the view. I will also highlight problems that occur in real-world settings when bringing computer vision algorithms from research to commercial products.

3.24 Structured prediction in Remote Sensing: The key to automated cartographic mapping?

Jan Dirk Wegner (ETH Zürich, CH)

Joint work of Wegner, Jan Dirk; Montoya, Javier; L'ubor Ladický, Konrad Schindler
Main reference J. A. Montoya, J. D. Wegner, L. Ladický, K. Schindler, "Mind the gap: modeling local and global context in (road) networks," in Proc. of the 36th German Conf. on Pattern Recognition (GCPR), LNCS, Vol. 8753, pp. 212–223, Springer, 2014; pre-print available from author's webpage.
URL http://dx.doi.org/10.1007/978-3-319-11752-2_17
URL http://www.igp.ethz.ch/photogrammetry/publications/pdf_folder/montoya-etal-roads-gcpr14.pdf

The interpretation of remote sensing images is still done manually today, which is very costly in terms of time and money. Our long term goal is to completely automate this task. In a first step towards automated cartographic mapping we focus on the extraction of road networks that we view as the "structural backbone" of cities.

We propose a method to label roads in aerial images and extract a topologically correct road network. Three factors make road extraction difficult: (i) high intra-class variability due to clutter like cars, markings, shadows on the roads; (ii) low inter-class variability, because some non-road structures are made of similar materials; and (iii) most importantly, a complex structural prior: roads form a connected network of thin segments, with slowly changing width and curvature, often bordered by buildings, etc. We model this rich, but complicated contextual information at two levels. Locally, the context and layout of roads is learned implicitly, by including multi-scale appearance information from a large neighborhood in the per-pixel classifier. Globally, the network structure is enforced explicitly: we first detect promising stretches of road via shortest-path search on the per-pixel evidence, and then select pixels on an optimal subset of these paths by energy minimization in a CRF, where each putative path forms a higher-order clique. The model outperforms several baselines on two challenging data sets, both in terms of precision/recall and w.r.t. topological correctness.

3.25 Sparse Optimization for Motion Segmentation

Michael Yang (Leibniz Universität Hannover, DE)

Motion segmentation aims to decompose a video sequence into different moving objects that move throughout the sequence. In this talk, I will show some state-of-the-art motion segmentation methods and our new framework based on subspace clustering with sparse optimization. We combine two sparse representations to optimize both the global and local estimation. Sparse PCA is applied for the global optimization. The local subspace separation is achieved via automatically selecting the sparse nearest neighbours. In the end of this talk, I will show some experimental results on the Hopkins 155 Dataset and Freiburg- Berkeley Dataset.

3.26 Gesture Recognition Portfolios for Personalization

Angela Yao (Universität Bonn, DE)

Joint work of Yao, Angela; Van Gool, Luc; Kohli, Pushmeet
Main reference A. Yao, L. Van Gool, P. Kohli, "Gesture Recognition Portfolios for Personalization," in Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14), pp. 1923–1930, IEEE, 2014; pre-print available from author's webpage.
URL http://dx.doi.org/10.1109/CVPR.2014.247
URL http://www.vision.ee.ethz.ch/~yaoa/pdfs/yao_cvpr2014.pdf

Human gestures, similar to speech and handwriting, are often unique to the individual. Training a generic classifier applicable to everyone can be very difficult and as such, it has become a standard to use personalized classifiers in speech and handwriting recognition. In this paper, we address the problem of personalization in the context of gesture recognition, and propose a novel and extremely efficient way of doing personalization. Unlike conventional personalization methods which learn a single classifier that later gets adapted, our approach learns a set (portfolio) of classifiers during training, one of which is selected for each test subject based on the personalization data. We formulate classifier personalization as a selection problem and propose several algorithms to compute the set of candidate classifiers. Our experiments show that such an approach is much more efficient than adapting the classifier parameters but can still achieve comparable or better results.

4 Working Groups

4.1 Working Group Summary on *What does "Scene Understanding" mean ?*

Bob Fisher

We had quite a wide-ranging discussion from what does a Rumba understand, to would a system need to be "Conscious" to do proper scene understanding. The discussion went around a lot, but could be categorised as belonging to the following 14 topics:

- Historical
  - Scene understanding has been around for c. 40 years at least, with the outdoor scene analysis of Hanson & Riseman's Visions system, and the SRI indoor scene labeling. Plus Brook's ACRONYM system in the early 1980s.
  - How much general progress have we made? What has limited the progress?
- What does "Scene Understanding" mean?
  - Is it general? Or maybe different tasks restrict their domain of focus/labels?
  - There seem to be many different tasks that require different sorts of SU.
  - There are different levels of SU: 1) A meeting, 2) a collection of interacting people, 3) a set of people and objects in a spatial configuration.
- Are there key classes of SU?
  - Static SU:
    - What type of scene is this?
    - What are the objects in this scene?

        - What are the relationships between the objects in the scene?
        - Are the objects and relationships consistent?
        - What are the subtleties of the scene (eg. aesthetics, tidiness, etc)?
    - Dynamic SU:
        - What type of action is this?
        - What/who is doing the action?
        - What is the sequence of sub-actions?
        - Is the set of actors, actions, and sequence consistent?
        - What are the subtleties of the actions (eg. for emotion/expression analysis, lie detection)?
- What quality of sensors do we need?
    - Is it task dependent, or do we want as much as we can get? Would this get us better data, but not better understanding?
        - Faster sampling, more pixels, better depth resolution, more spectral channels, more bits to the data?
    - Could we do everything we want if we had a great low-level 2.5D+RGB video sensor? As an alternative to struggling with impoverished data, letting us focus on just the SU task?
    - Is it cheating to use specialised sensors? Eg. hyperspectral sensors for material classification? Or is it sensible to use the sensor that allow optimal SU performance?
- What is the role of temporal data?
    - Do we need it for static scene analysis?
- What representations do we need for SU?
    - Do we need all of high-level vision?
    - Should the representations be for more general-purpose SU, or should there be many different reps, each tuned for more focussed competences?
    - Does static SU require only 2D models, or 2.5/3D models?
    - Do these models need to be generative, or will discriminative do?
- Learning the content in the representations needed to do SU
    - How much "prior" or "common sense" knowledge should be innate?
    - Should we expect to learn the variety of the world from many examples, or from only a few?
    - Do we need to be able to act to learn, or can we be passive?
    - Do we need an oracle/teacher/human in the loop?
- What databases do we need to underpin SU?
    - By analogy to WordNet and ImageNet, do we need an ActivityNet and SpatialRelationNet and ObjectContextNet?
- What technologies are needed for SU?
    - Would a CNN do it all? What would be the outputs? What would be the inputs?
    - Would we need an excessive amount of hand-labeled training data (images and videos)?
    - Should we be using simulators to generate training data? At least at the initial stages? Or at just the higher levels, symbolic rather than signal level? A lot of controversy here.
- What is the Roadmap to competent SU?
    - We can do a few objects, relations and activities now.
    - There should be some serious work on a Roadmap.
- What are some of the key problems that limit better SU?
    - Imprecision of language for describing scenes

  - Many truths/many equally valid descriptions of the same scene/image.
  - Descriptions can lie at many levels.
  - Are we biased by the closed-world assumption?
  - The natural variation of the real-world. And how much of this do we need to model (which depends on the task).
  - The breadth of possible applications for SU of the same scene.
- How can we measure progress in SU?
  - Research moves around from topic to topic, but there are not good benchmarks.
  - Is the goal binary (can we do it at all?) or graded (can we do it some of the time, or be partly accurate)? How do we assess this?
  - There seems to be a large number of results that we can expect from SU. How can we tell if/what is making progress? Some results are focussed and easy to assess (How many people are in this scene?), others are open-ended (What is happening in this scene?).
  - It's a bit like a Turing Test (or Watson).
  - Maybe some assessment of whether SU is happening could be by question-oriented tests:
    * What happens next?
    * What must have happened between these 2 events?
    * Is there anything out-of-place in this scene?
    * Did anyone do something unusual?
  - Would have to avoid Eliza-like behaviour.
- Benchmarks to assess and drive progress?
  - What should they be?
  - Should they be wide ranging for general SU? Or focussed to assess specific competences?
  - The risk is developing programs to solve benchmark, not the real problem.
- Do we really need SU?
  - Yes, otherwise we might lose our jobs.
  - More seriously: Is it useless because it does not solve a task?
  - Can we hope to have a sufficiently competent general SU capable of answering any question about a scene/video?
  - Or do we instead need a set of highly competent specialised capabilities, for the 100 (?) general tasks that most people have to do.
  - Is it SU only as long as we cannot get a machine to do it, and thereafter it's simply automation?

4.2 Working Group Summary on *Dynamic Scene: Humans*

Greg Mori

Summary of topics/open questions:

- We need a big collection of data with good sampling. People are a unique object, will have lots of data but not enough labeled.
- We need to interpret groups of people from crowds as a texture down to individuals.
- We need to think about occlusion when detecting these people.
- You need to decide what you want to recognize, then you can build it?

- Can we build a semantic space of actions to rival WordNet? Bottom-up clustering will be hard, we don't have the semantic information from something like WordNet.
- Ask people to provide descriptions at multiple levels of detail when viewing an action.
- Noun and verb only? Probably not, need longer-term annotations to go to the story-level.

Data

- Granularity of representation for people
- From individuals to groups and a crowd of people
- 2d boxes
- joint angles and positions
- Annotate with everything or not
  - What can we do with lots of annotated data ?

Detection

- Rescue robotics with unusual poses, non-moving people
- Thermal cameras, multi-sensor input
- Face detection is almost solved, but not non-pedestrian people
- Dataset
  - Imagenet for humans?
  - Should we have a dataset of varied poses? Is this the main thing needed?
- What about representation?
  - Parts are good, lots of people in a crowd, just detect heads?
  - Poselets?
  - Do we need a 3d representation at all?
- Motion?
  - Range of pose in videos is larger than images (collected)
  - Should we segment the people from the videos
  - Maybe motion is not so important, we can focus on just detecting people in images?
  - If we have the video can we try prediction at any time scale
  - Subtle motion is very important in surveillance, this is hard to deal with
  - Temporal information is often just used for smoothing, not enough modeling of time
  - Motion is dependent on scale, at a far distance, it is hard to see the motion, but close-up you can segment
  - Partial occlusion of people in a dataset would be very useful
    - A few % improvement by using templates for 2 people, perhaps a little more for 3 people (Schiele et al.)
  - Label occlusion for people in a dataset
  - Interactions with objects for detection (Fua et al. ECCV14)
  - How many images do you think we need for a "good" human detector in a city scene (any photo taken from a hand-held camera)
    - Millions?
    - Stratified sampling is important
    - Rarity in the benchmark is important, need to analyze different categories
    - How do we collect such a dataset, there are lots of different cases and there is not a Wordnet equivalent to organize it
    - Maybe active learning / clustering / exemplar SVM is needed to build this
    - Could do this based on 3d body configuration representation
      - Sample possible configurations
    - PeopleNet

        - How do you label bounding boxes on these people, e.g. people on stilts
        - Large diversity is important, videos are not so good for diversity
        - Active learning approach for collecting a dataset, pay people on AMT who come up with images which are false negatives for the current algorithm
- Detecting individuals is not all there is
    - Some applications don't need person-centric or individual representations
    - Crowds of people moving
        - Is this dynamic texture?
        - Emergence of individuals from the flock, flows from the crowd
        - Phase changes
    - Do you want to find abnormal behaviour of one person in the crowd?
    - The scale of the people seems to dictate this
        - Once the people are too small we just model as a crowd
- PeopleNet could have gestures and videos
    - Go to actions of individuals

Tracking

- Evaluation of tracking for the purpose of activity analysis
    - Measuring tracklet importance using cameraman's gaze
- Hands and gesture vs. human bounding box tracking
    - Landmarks on body or a generic tracker that works on any objects

Actions

- What are the important actions to recognize
- Jump, wave hands, walk
    - Are these important?
    - Labeling crowds in hockey stadiums
    - Lists of postures and poses that are important to psychologists
    - Need help from psychologists
    - We don't have the equivalent of WordNet for actions
        - Derive it by looking at verbs in language
        - Some are not actionable
    - Given a frame, label it dangerous / not dangerous
    - Psychologists' representations should be latent variables
- These 4 problems are intertwined, but not all tracklets of all people are equally important
- Starting and ending point of an action

Group activities

- The scale of inference
    - When is motion necessary, treat the group as a texture or a whole
    - When is it necessary to reason about individuals
- A project on combining holistic crowd representations with the actions of individuals and multi-scale representations
- With images of a crowd vs. videos of a crowd, some things are not possible
    - But is this a natural task?
- Let's say you have unreliable tracking, and just get some people tracked but lots of incomplete tracklets, some false positives
    - Can you infer group activities?
    - Can we cluster or find dominant patterns?
    - Clusters of optical flow?

- How do we build systems that are robust across density and scale of people/vehicles
  - When you have rows of cars on a busy street, too hard to detect individuals, just represent optical flow
- Does looking at a group help understand occlusions / tracking for an individual?
- Should we learn these together or separately

Long-term story

- People in the event give knowledge about what is important (record gaze)
- Rich data includes video meta-data which give high-level information about the purpose of a video
  - Sports videos are probably good for this
  - Surveillance videos are bad
- Video summarization?
  - No, it's interpolate beyond what you can see
  - No, it's about in textual domain what happened
  - Yes, show me the beginning, middle, end of a story
- The longer activities, bigger pieces
  - What is happening beyond just one image or a short video
  - Describe and infer things about what is happening in a scene
  - Not everything is actually observed
  - How do you expand to a larger view
- Cooking videos, what was being prepared, exhaustively detecting all the details
- Inferring intent from few observations
- Temporal order is not fixed, events can take place in different orders
- What are the key, dominant phases in an activity
- Problems in the editing of the domain, e.g. cooking videos are edited to show only interesting clips
- How do we judge success for this?
- Summarize large video collections
  - Captions and titles are very useful
- Text is a better representation than video
  - But dogs can't cook omelettes
- Task is to ignore the mundane bits but focus on the parts of the video that people haven't seen often
- Look for 2 goals in 90 minutes of soccer
  - Should you include the boring stuff in the summary
  - E.g. include the boring parts of a game in the summary or not?
  - Storytelling constrained by the low-level detections we can do now
- How do I go from chopping onions to making soup
- How do we get datasets that contain more than just a clip-let
- Produce as much information in a summary of 4000 characters so that a cook can produce the desired output
- Use clips with (noisy) meta-data to generate a summary
- It's about how to give an instruction not just summarize a video
- Figure out what a person did over a whole day
- Build datasets with long clips that will eventually contain the events you want
  - Lots of sports videos
  - Are egocentric videos good for this?
- Analyze what people are doing over a longer time span, not just what someone is doing right now

4.3 Working Group Summary on *Recognition in static scenes (in 3D)*

Alessio Del Bue

Scene understanding has the goal of building machines that can see like humans to infer general principles and current situations from imagery. More detailed, the ultimate goal is to let a machine pass a "visual Turing challenge", e.g. to allow arbitrary questions from a human and to obtain useful answers. Sitting in a music room, there was a longer discussion about the height of a plant standing on a window ceil with a radiator below.

Starting from a single image, pointing to a single pixel, asking a straight (single sentence) question ("How tall is this?"), in a few years, the system will identify and segment the object, estimate the context and reach some conclusion that the height of the plant is xy-cm +/- a variance. Due to ambiguities, occlusions or missing information, the system should be allowed to infer to provide uncertain answers and to recheck with the human to increase the certainty. Questions can be roughly arranged into several categories, e.g. extrapolation, interpolation or straight) and answers such as category (discrete) or continuous.

- Data/Example collection: Use a robot in a room and an open internet to collect questions
- Alternative : Show random images and let people ask random questions and get people to answer them ...
- → Classify questions and learn the manifold of questions (Counting, Measuring, stupid q.)
- Parsing methods for questions (Grammars), Analysis/Rephrasing to canonical from.
- E.g. question + recognition → Text description + query expansion (cloud / big data).
  - Generative / discrimative learning, priors for regression ?
  - ...

Still problematic: Current accuracies of naïve methods are in the order of 7–12 %, the Human Baseline is 60%. Are we humans really "that good" ?

There is still the Question of metrics unsolved. For quantitative evaluation should be in relation to the quality humans can reach.

5 Schedule

Monday, February 15th, 2015

09:15–09:30	Bodo Rosenhahn: Opening
09:30–10:00	1 minute self-presentations
Chair:	Michael Yang
10:45–11:15	Anna Rohrbach: *Describing Videos with Natural Language*
11:15–11:45	Sebastiano Vascon: *Detecting conversational groups: a game-theoretic approach with sociological and biological constraints*
11:45–12:00	Discussions
Chair:	Jürgen Gall
14:00–14:30	Gabriel Brostow: *Superhuman-in-the-Loop Computer Vision*
14:30–15:00	Sanja Fidler: *Understanding Complex Scenes and People That Talk about Them*
15:00–15:30	Oisin Mac Aodha: *Image Annotation–From Machine Learning to Machine Teaching*
Chair:	Anna Rohrbach
16:00–16:30	Marco Cristani: *From groups to crowds: a social signal processing perspective*
16:30–17:00	Konrad Schindler: *Towards 3D scene understanding: how to assemble the pieces?*
17:00–17:30	Caroline Pantofaru: *Interactions Between Scene Elements*

Tuesday, February 16th, 2015

Chair:	Stefan Walk
09:00–09:30	Laura Leal-Taixé: *Learning an image-based motion context for multiple people tracking*
09:30–10:00	Michael Yang: *Sparse Optimization for Motion Segmentation*
Chair:	Jan Wegner
10:15–10:45	Angela Yao: *Gesture Recognition Portfolios for Personalization*
10:45–11:15	Roberto Henschel: *Solving Multiple People Tracking Using Minimum Cost Arborescences*
11:15–11:45	Greg Mori: *Structured Models for Recognition: Towards Sub-Category and Interaction Discovery*
11:45–12:00	Discussions
14:00–15:00	Jiří Matas, Vittorio Murino: Working Group Definition
15:30–18:00	Working Group Meeting

Wednesday, February 17th, 2015

Chair:	Roberto Henschel
09:00–09:30	Vittorio Murino: *Towards Holistic Scene Understanding: low-level cue extraction, collective activity classification, and behavior recognition*
09:30–10:00	Bob Fisher: *Fish detection, tracking, recognition and analysis with the Fish4Knowledge Dataset*
Chair:	Oisin Mac Aodha
10:15–10:45	Michele Fenzi: *Feature Regression-based Pose Estimation for Object Categories*
10:45–11:15	Jürgen Gall: *Human Pose: A Cue for Activities and Objects*
11:15–11:45	Abhinav Gupta: *Geometry, Function and Common Sense*
11:45–12:00	Discussions
14:00–18:00	Social Event: One group made a rip to Trier, and anothr group went on a 3h hike in the environment.

Thursday, February 18th, 2015

Chair:	Jörn Jochalsky
09:00–09:30	Michal Havlena: *VocMatch: Efficient Multiview Correspondence for Structure from Motion*
09:30–10:00	Jan-Michael Frahm: *Understanding Scene Dynamics from Crowd Sourced Imagery*
Chair:	Michele Fenzi
10:15–10:45	Ron Kimmel: *A Spectral Perspective on Invariant Measures of Shapes*
10:45–11:15	Matthias Reso: *Temporally Consistent Superpixels*
11:15–11:45	Jiří Matas: *History of Object Recognition: to be learned for Scene Understanding or a Scrapyard?*
11:45–12:00	Discussions
Chair:	Michal Havlena
14:00–14:30	Reinhard Klette: *Visual Odometry and 3D Roadside Reconstruction*
14:30–15:00	Jan Wegner: *Structured prediction in Remote Sensing: The key to automated cartographic mapping?*
Chair:	Matthias Reso
16:00–16:30	Min Sun: *"At-a-glance" Visualization of Highlights in Raw Personal Videos*
16:30–17:00	Alessio del Bue: *Semantic Motion Segmentation and 3D Reconstruction*
17:00–17:30	Raquel Urtasun: *Towards autonomous driving*

Friday, February 19th, 2015

Chair:	Markus Enzweiler
09:00–09:30	Esther Hobert: *Generic Object Detection in Video using Saliency and Tracking*
09:30–10:00	Stefan Walk: *Understanding scenes on mobile devices*
10:15–10:45	Shaogang Gong: *Explore Multi-source Information for Holistic Visual Search*
11:00–11:55	Working Group Meeting and Summary

Participants

- Gabriel Brostow
University College London, GB
- Marco Cristani
University of Verona, IT
- Alessio Del Bue
Italian Institute of Technology – Genova, IT
- Markus Enzweiler
Daimler AG – Böblingen, DE
- Michele Fenzi
Leibniz Univ. Hannover, DE
- Sanja Fidler
University of Toronto, CA
- Bob Fisher
University of Edinburgh, GB
- Jan-Michael Frahm
University of North Carolina – Chapel Hill, US
- Jürgen Gall
Universität Bonn, DE
- Shaogang Gong
Queen Mary University of London, GB
- Abhinav Gupta
Carnegie Mellon University – Pittsburgh, US
- Michal Havlena
ETH Zürich, CH
- Roberto Henschel
Leibniz Univ. Hannover, DE
- Esther Horbert
RWTH Aachen, DE
- Jörn Jachalsky
Technicolor – Hannover, DE
- Ron Kimmel
Technion – Haifa, IL
- Reinhard Klette
University of Auckland, NZ
- Laura Leal-Taixé
ETH Zürich, CH
- Oisin Mac Aodha
University College London, GB
- Jiri Matas
Czech Technical University, CZ
- Greg Mori
Simon Fraser University – Burnaby, CA
- Vittorio Murino
Italian Institute of Technology – Genova, IT
- Caroline Pantofaru
Google Inc. – Mountain View, US
- Matthias Reso
Leibniz Univ. Hannover, DE
- Anna Rohrbach
MPI für Informatik – Saarbrücken, DE
- Bodo Rosenhahn
Leibniz Univ. Hannover, DE
- Bernt Schiele
MPI für Informatik – Saarbrücken, DE
- Konrad Schindler
ETH Zürich, CH
- Min Sun
National Tsing Hua University – Hsinchu, TW
- Raquel Urtasun
University of Toronto, CA
- Sebastiano Vascon
Italian Institute of Technology – Genova, IT
- Stefan Walk
Qualcomm Austria Research Center GmbH, AT
- Jan Dirk Wegner
ETH Zürich, CH
- Michael Yang
Leibniz Univ. Hannover, DE
- Angela Yao
Universität Bonn, DE