Mission
The Computer Vision Lab wants to spearhead research in its area. Computer vision tries to let computers “see”, i.e., interpret the content of images. And it is going through exciting times. Increasingly, solutions let computer vision work in uncontrolled environments, i.e., “in the wild”. This created a surge in new applications, several of which consumer oriented, often via mobile apps. But people also get to experience computer vision’s recent successes via special effects in the movies, innovations in navigation, on-line shopping, modern advertising, etc., without even being aware of it. With our unit, we try to cover a wide spectrum of topics (incl. recognition, tracking, 3D reconstruction, etc.), so that we can integrate multiple state-of-the-art components into seeing systems that push the envelope of what is possible with images. The creation of successful spin-offs is the ultimate test that our methods work robustly “in the wild”.

In the courses, we aim to convey those same concepts and ideas. Our courses are designed to cover a broad range of topics, to give our students (about 120/yr) a good overview of what computer vision tries to achieve. Most of the content is explained in a written course text of over 350 pages. Recent topics are available as slide presentations, with the goal of writing a course text for those too. We aim at explaining the theoretical foundations of the different methods, while also illustrating applications. Where possible, images and videos are used to do so.
Research Activities and Achievements

As said, Prof. Luc Van Gool leads the Visual Communications unit, which focuses on:

- The recognition of specific objects and of object classes
- Tracking, and the analysis of motion patterns and gestures
- Texture analysis and synthesis, image enhancement
- (Im)painting, super-resolution, ...
- User-friendly and large-scale 3D reconstruction and modeling
- Real-world combinations and applications of the above

These topics have often been studied in isolation, but our group has worked on several systems that take a more holistic approach. For instance, when trying to build a 3D model from images, there are many pitfalls. Yet, if one would also be able to recognize what it is that one is trying to model – and thus combine 3D reconstruction with recognition – some of those can be avoided. Vice versa, when trying to recognize, it can be beneficial to not only use the images, but also the 3D layout of the visible scene.

One of our lines of research aims at traffic safety support for car drivers. We have been collaborating with Toyota for several years already. The goal is to analyze traffic scenes from mobile platforms, such as cars. In this context we develop computer techniques that detect and track people, cars, etc., relative to the driver. In order to achieve this, several modules extract complementary information. The overall scene interpretation is the result of a joint (Bayesian) optimization. It results from multiple interactions between the modules. The mobile platform uses a stereo setup. Its 3D data allow for the estimation of a ground plane. This reduces the search space for people (pedestrians) in the images, as they are supposed to walk on this plane. On the other hand, once people have been detected, the estimated positions of their feet influence the estimation of the ground plane. There are many such (mutual) interactions that are taken account of.

Trackers follow the people in order to predict their positions in subsequent frames and thereby facilitate their detection, and these trackers benefit from the results of pedestrian detectors. The tracks are not defined relative to the platform, but the 3D world. And so is the motion of the platform. When extracting the 3D model of the scene and the motion of the mobile platform therein, it is important to only use point correspondences from the static parts of the environment, and no points on moving agents, such as pedestrians, which therefore have to be detected to be eliminated from the point correspondence search.

Another project that combines multiple cues is the ERC Advanced Grant ‘VarCity’ (Variation & the City). The goal is to build 4D city models. Both the static and dynamic aspects are to be captured: buildings and vegetation are to be modeled in 3D, but also typical traffic flows need to be added so as not to end up with virtual ghost towns. VarCity combines 3D analysis with tracking and recognition at different levels. For instance, when modeling the facade of a building, the model ought to be highly structured, along concepts such as floors, windows, balconies, etc. Building such semantic models calls for the ability to recognize those building components. We have also demonstrated that the views that one should best select for 3D reconstruction from images depend on the type of facade component and therefore vary across the facades, thereby again combining 3D reconstruction and object class recognition. On the other hand, once extracted, these semantically enriched models can be made compact, clean, high-resolution, and realistically looking in comparison to the usual 3D models derived from all the images in the traditional bottom-up fashion. As input for the modeling of the static part, we use image-based mobile mapping data (images and 3D point cloud). The project will use the semantic or ‘procedural’ models also for refinements such as inpainting and Level-of-Detail rendering. It should be emphasized that the very way these 3D city models are constructed prevents people, cars (with their license plates), interiors behind windows, etc. to show up in the final model. Rather than dealing with anonymity in a separate post-processing step [e.g. blurring faces], anonymity and privacy are guaranteed through the very way in which the models are being produced.

In VarCity, the modeling of the traffic flows starts from intermittent video data. The idea is that, increasingly, people take videos continuously while moving around. Such data will therefore be available from many places at many times, but this spatio-temporal distribution is not under our control. Bursts of information need to be integrated into overall, largely interpolated traffic flow patterns. Thus, at some points and some times, observations of the actual traffic flows are available, without us being able to pick those places and times. Of course, cities typically also have a number of fixed traffic surveillance cameras installed, which are predictable. In any case, the traffic flows that are inferred from the combination of these video data should reflect the actual traffic densities and compositions (i.e. containing cars, bikes, buses) at all places and times as well as possible. This requires state-of-the-art trackers and detectors for traffic agents and novel approaches for the inference of the traffic flows where observations are missing. As for the representation or visualization of those flows, these are consciously designed to only simulate the actual statistics in terms of speed, composition, and density. Therefore, the trajectories of individual pedestrians or cars are not replicated, for obvious reasons of privacy protection. In those visualizations, the virtual traffic agents also do not take on the appearance of the real ones.

There are many other projects and research activities that run in parallel at our lab. For instance, the group generates body models and analyses facial expressions for tele-medicine. A physician can examine a remote patient via a touch-sensitive robot. In other projects we reconstruct dynamic scenes in 3D from only uncalibrated images. This involves solving for the unknown relative scales between the moving objects. We also extract and label the layout of scenes in real-time.

Currently, when developing object class detectors, usually a separate detector is trained for each class. If one would want to deal with additional classes, ever more detectors need to be applied and computation times go up quickly. Our lab has succeeded in building up a framework that can integrate the detection of additional classes more efficiently. We also recognize food and dishes for dietary purposes, analyze human poses and clothing, automatically select interesting images and summarize consumer videos, complete 3D shape models based on object type, find faces and their feature points “in the wild”, etc.

Pedestrian tracking

Human pose extraction

For this work, the Visual Communications unit has received several prizes, e.g.:

2013: Best Paper award at the Workshop on Application of Computer Vision (WACV)
2012: Best Paper award at Workshop on Computer Vision in Vehicle Technology
2012: Best Scientific Paper award at the International Conference on Pattern Recognition (ICIP)
2012: International Society PhotoGrammetry & Remote Sensing Helava Prize (best paper in 4 years)
2011: Best Paper at the British Machine Vision Conference (BMVC)
2011: Best Impact Paper at the British Machine Vision Conference (BMVC)
2010: Best Student Paper at the International Conference on Image Processing (ICIP)
2009: Best Student Paper at SCA
2009: Innovation Prize Electrosuisse
2008: Most cited paper Journal Computer Vision and Image Understanding (CVIU)
2007: Best Paper award at Asian Conference on Computer Vision (ACCV)
2007: Best Paper award at IEEE Conference on Computer Vision and Pattern Recognition
1998: Da Vinci Prize at the International Conference on Computer Vision (ICCV)