Review

of the doctoral dissertation of **Mr Maciej Kurc**,

entitled **'Hybrid techniques of depth map estimation and their application in three-dimensional video systems'**

by **Atanas Gotchev**,

Professor of Signal Processing at the Faculty of Information Technology and Communication Sciences, Tampere University, Finland

The dissertation addresses the challenging problem of sensing 3D visual scenes by multiple sensors and fusing the heterogeneous data into a video-like 3D representation, where the scene geometry is represented along with the colour video modality. The problem arises from the ever-lasting demand for understanding and recreating 3D visual scenes. Human users want to experience the visual world in its genuine entirety: not only in true colour, high spatial resolution, and high dynamic range but also in all three dimensions. 3D allows for more realistic perception, feeling of 'being there', understanding the spatial structure and location of objects, and accomplishing tasks more accurately. Along with human users, machines also need an accurate 3D scene representation. Most notably, these are (semi-) autonomous vehicles and other robots, which have to localize their positions with respect to the environment and to map the latter.

In order to analyse and recreate 3D visual scenes, one has to accurately sense and reconstruct the 3D structure. The classical and most widely spread approach is to use two or more cameras, seeing the scene from different perspectives. Object points are projected on different cameras' pixels. By finding (stereo) correspondences, and knowing the camera parameters, one can triangulate and find the corresponding 3D world coordinates. These can be organized in a map, referred to as depth map, usually being aligned with the camera views. Multi-camera depth estimation is a powerful approach to find the scene geometry however, in many cases it is not sufficient. Most notably, this is the case when the scene is textureless and corresponding points cannot be matched easily. Furthermore, finding depth from stereo correspondences requires computing, which takes time. Therefore, researchers have attempted constructing specific active depth sensors, which can be complementary to the colour sensors. They should sense the depth directly and thus augment the colour views with no further computations. Various depth sensors have different operational principles and the presented thesis studies the class of sensors, which operate by emitting a continuous harmonic wave in the near infrared spectrum and then measuring the phase difference with the reflected-by-scene-objects wave, which is a measure of the objects' distances from the camera. This is the so-called Time-of-flight depth sensing principle. While a promising technology, ToF comes with its own limitations. Sensing reflected IR wave requires wider pixels in order to catch more light, which in turn reduces the spatial resolution. ToF sensors have different field of view compared to conventional video cameras. This imposes the clear problem: how to make ToF sensors working together with colour cameras – knowing their relative positioning, and fusing the multi-modal data into an aligned view+depth 3D representation.

The author of the thesis has correctly identified these problems and has formulated the primary goal of the thesis as **fusing video and ToF depth data**, where the fusion process means increasing the depth spatial resolution and aligning it on the plane of the colour camera. Subsequently, this main goal imposes a sequence of more specific goals, as follows:

1. Synchronize multiple heterogeneous sensors (colour and ToF) to get the imagery streams at the same moments of time.
2. Estimate camera parameters, both intrinsic and extrinsic, in order to correctly interpret trans-camera correspondences
3. Rectify multi-camera settings, in order to get all sensors on the same place and facilitate the correspondence search.
4. Pre-process the ToF to enhance the corresponding imagery. That is, correct possible depth measurement errors and suppress any other noise components.
5. Improve the multi-view multi-depth depth composition and ensure temporal consistency of the depth map sequences.

The specified goals are ambitious as they address challenging problems. Despite decades of research, there are still open issues in multi-camera synchronization, calibration, and multi-sensor data fusion. They are clearly difficult and significant enough and require new research on post-graduate level.

The dissertation is structured in seven chapters, dealing with various aspects of the above-mentioned problems. In addition, there are four annexes presenting designs, data and experimental results in their entirety. The dissertation is completed with a list of concepts, symbols and abbreviations, and bibliography.

Section 2 is devoted to the estimation of camera parameters in multi-camera systems and the subsequent multi-image rectification.

The author has considered a linear arrangement of several colour camera, where the depth cameras can be placed arbitrarily, while maintaining a FoV overlap. Given that hybrid setup, the first goal is to get the colour cameras jointly rectified and the second goal is to make a parametric estimation of two or more depth cameras.

The chapter offers a detailed overview of camera basics, including projective geometry, pinhole camera model, and calibration and rectification methods. Based on the presented state of the art, the author proposes some modifications aimed at mutually calibrating and rectifying multiple colour cameras. The proposed modifications are based on imposed constrains on and assumed approximations of the corresponding transformations. Cameras are assumed to be equidistantly located on a line and their individual positions are restricted to affine transformations. The proposed techniques for rotation and translation compensation are technically correct and rather straightforward. The experiments for validating the performance are quite limited. While they are illustrative about the performance, more quantitative experiments with different scenes and against state of the art would be beneficial.

The problem of estimating camera extrinsics in a multi-sensor setup is considered next in the chapter. The state of the art is adequately presented. The author's proposal is based on the use of two depth cameras bringing directly the scene geometry. Estimation of the rotation parameters is performed by a singular value decomposition. Furthermore, the author proposes a specific 3D calibration marker for a single calibration point and quantifies its performance against the conventional checker board 2D pattern. The experiments demonstrate the advantages the new 3D pattern offers.

Section 3 is devoted to the ToF depth measurement correction. More specifically, three corrections are considered: 1) geometrical corrections; 2) calibration of ToF depth measurements, and 3) denoising of ToF data.

Geometrical corrections are needed in order to convert measured distances to world depth coordinates. Calibration is need for correcting sensor systematic errors. Denoising is subsequently needed to correct any remaining erroneous measurements caused by the sensor or the environment and jointly characterized as noise. Both noise components in a single frame and in subsequent frames are considered.

The need for a good geometrical correction is well motivated. The fact is that ToF sensor manufacturers do not provide access to sensor parameters and a custom and transparent procedure is needed. The author has illustrated the implemented geometry adjustment by a single experiment.

To achieve a metric calibration of a depth sensor, one needs to relate the ToF measurements with the measurements by another means. In the thesis, these reference measurements are accomplished by assuming the pinhole camera model and using calibration patterns. The author's aim has been to design a simple setting and a corresponding practical technique. With this aim, he has proposed a linear model, which accounts for the transformation from phase measurements to metric distances. The proposed setup includes a calibration pattern with circles moved a few times (e.g. 3) at different distances while also being titled. The approach is notably simple and clearly relates the pinhole camera model with the ToF measurements.

The third group of corrections is related with the presence of noise in any sensor measurements. In the case of ToF sensors, the noise is caused by objects with low reflectivity, reflecting a weak signal, which is susceptible to other effects, such as trapped light, multi-path reflection and sensor noise. A good denoising method should smooth the depth planes while preserving the edges delineating objects at different depth. Denoising of a single spatial frame would take into account the spatial scene structure only. Temporal averaging of successive frames would be very effective (due to the central limit theorem) however, temporal frames acquire also object motions and a simple averaging would blur the scene. These factors motivate the author to seek a solution, which combines edge-preserving spatial filtering with motion-adaptive temporal filtering. The classical bilateral filter is selected as a spatial filter, and first order regression filter is selected to operate over successive frames. Motion adaptation is implemented through blocking the temporal filtering for areas where motion is detected. Motion detection is implemented over the sequence of intensity images. As they might also be noisy, a good amount of discussion in the thesis is spent on estimating the noise variance in the intensity images as this variance sets the threshold, which distinguish motion.

I have a few questions related to this part of the work.

1. The author uses the terms *spatial* and *temporal* noise. This nomenclature needs clarification in terms of factors causing the noise and the corresponding statistical noise model(s).
2. How the performance of the presented method would change with respect to different noise levels and objects with different reflectance?

The author has mentioned several times that his approach cannot be generalized. In fact, I think, it can and it should, as this is the purpose of any scientific research. While the proposed denoising method seems doable and efficient, it would benefit from more extensive benchmarking against varying sensing conditions and with respect to the state of the art.

Section 4 is devoted to the synchronization of video and depth cameras. It emphasizes the importance of time synchronization of multiple sensing cameras for the correct disparity estimation and contains a comparison of

different synchronization techniques. Factors such as jitter are analysed for their influence on the generation of a trigger signal from a video Genlock signal.

Based on the analysis, the author has designed a new device, which gets a Genlock signal as an input and generates a trigger signal as an output. The design is thoroughly documented in an Appendix.

Section 5 comes as the central chapter in the thesis as it discusses the fusion of video and depth data. The proposed approach aims at projecting the depth data on the colour camera plane, performing aggregation of depth maps coming from multiple depth cameras and eventually obtaining a high quality, high-resolution depth map aligned with the colour map.

The author discusses the deficiencies of other approaches and comes up with a proposal to convert the multiview depth data into meshes, which then can be projected on the colour camera plane and interpolated at the positions of the colour pixels. The author has motivated the choice to use quadrilaterals instead of triangles, when constructing the mesh out of depth measurements. Depth aggregation is accomplished by determining the max depth value, after converting all meshes to the colour camera plane.

The author has paid special attention to the problem of dis-occlusions. He has clearly motivated the use of multiple depth cameras for handling occlusions. Then a technique for breaking the mesh so to make the dis-occlusion regions discontinuous is proposed.

At this stage, I have the following question to the doctoral candidate: Is there any issue of *interference* between emitted/reflected signals while simultaneously using two or more active depth cameras?

Further, in Chapter 5, the author addresses the problem of fusing two different depth modalities: one coming as a result of the multi-camera depth and another coming from the active depth sensors. The proposed approach is to modify the cost function serving the depth estimation with a term driven by the ToF depth data. This modification is motived by analysing the existing cost function terms, namely data and smoothness terms. The aim is to incorporate a term responsible for image edges. Furthermore, different cost functions are discussed in order to identify the best convex combination of the candidate terms.

Having the modified cost function in place, the author proposes a number of multi-view multi-depth capture system in terms of number of cameras and camera topologies. Even though his attempt is to eventually process video data, he has specifically focused on the case of still image frames.

The chapter contains also a discussion on how to assess the quality of the reconstructed depth maps. The author argues on the use (or no use) of virtual views, synthesised using the obtained depth maps and how much these virtual views are representative for the quality of the depth maps. It might be instructive to have part of this discussion during the public defence, as some details need to be further clarified. More specifically, the author raises the opinion that a ToF-sensed depth map cannot be assessed through virtual view synthesis. This somehow contradicts the idea of fusing multiple depth maps obtained by different means and I would like to discuss this with the author during the defence. Another highly interesting point to be addressed during the defence is how to interpret the results in Tables 5.6.7 and 5.6.8. An inspection of these tables reveal that there is no improvement in the case of fused depth maps compared to the case of estimated depth maps. A provocative question: Is there any use of ToF data then?

At the same time, comparisons of different depth maps in terms of bad pixels show clear improvement in the case of fused maps. Perhaps the problem is not in the fusion and quality of ToF data but in the view synthesis

algorithm (which is not within the thesis scope but simply used off-shelf). What is the candidate's opinion on this?

In overall, I find the proposed modifications meaningful and worth considering in multi-view muti-depth systems. The lack of quantitative improvement in the case of comparisons of differently synthesized virtual views needs to be analysed more thoroughly.

Section 6 addresses the problem of depth map refinement via inter-view consistency improvement.

It starts with an overview of the multi-view video compression problems and corresponding compression standards. Then, the problem of depth map inter-view consistency is analysed and existing solutions are overviewed. Based on this, the author formulates a proposal, with the core idea to exchange depth information between multiple-view depth maps and to utilise a purposely-designed inconsistency measure to guide that exchange.

I find this chapter most significant in terms of scientific results. The conducted experiments clearly demonstrate that the proposed method overcomes the state of the art and therefore is of practical importance for the multi-view multi-depth compression standards.

In summary, this is a solid work on doctoral level. Its main feature is that it addresses the scientific problem of depth sensing and fusion in its entirety: starting with location and calibration of cameras, through pre-processing, enhancement, data fusion and ensuring multi-view consistency for the needs of effective compression. The proposed solutions extend the current knowledge on the topic. The material is abundant and demonstrates the candidate's personal and deep involvement in all stages of designing, experimenting and analysing the 3D video capture systems and algorithms. There are places, where the analysis would benefit from a more profound comparison with the state of the art and playing with varying sensing conditions. It seems that in his attempt to cover everything, the author opted to skip such experiments for the sake of manageable dissertation length. While I believe that such experiments would further strengthen the contributions, I can confidently declare that the proposed techniques are original and highly meaningful and they demonstrate the author's profound understanding of the multi-sensor data and corresponding processes. The dissertation meets the Polish national requirements for granting a doctoral degree. I am looking forward to having an expert discussion with the candidate during the public defence.

10.05.2019, Tampere

Atanas Gotchev, Professor of Signal Processing
Head, 3D Media Group
Director, Centre for Immersive Visual Technologies /CIVIT/
Phone: +358 40 8490733; email: Email: atanas.gotchev@tuni.fi

Faculty of Information Technology and Communication Sciences
Tampere University
www.tuni.fi/en
Human Potential Unlimited.