VarCity – the Video: the Struggles and Triumphs of Leveraging Fundamental Research Results in a Graphics Video Production

Kenneth Vanhoey, Carlos Eduardo Porto de Oliveira, Hayko Riemenschneider,
András Bódis-Szomorú, Santiago Manén, Danda Pani Paudel, Michael Gygli, Nikolay Kobyshev,
Till Kroeger, Dengxin Dai, Luc Van Gool
Computer Vision Laboratory, ETH Zürich

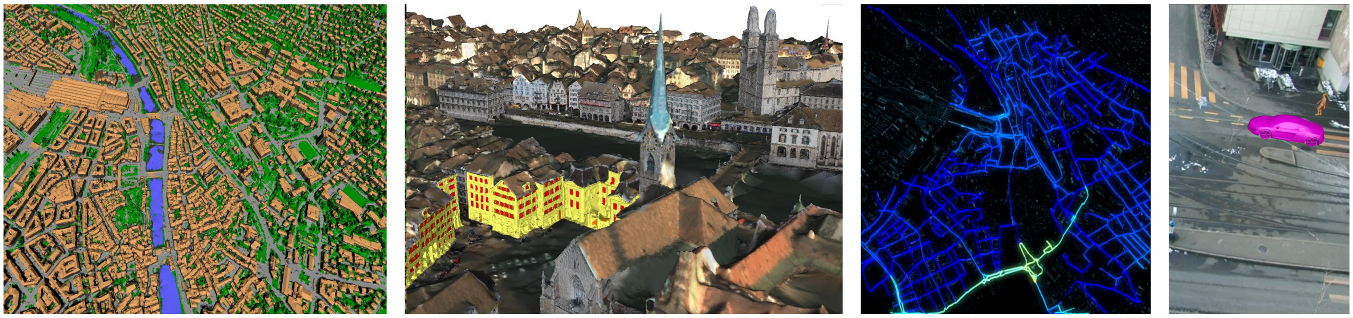

Figure 1: Data used in our video. First: distant view of automatic city-scale 3D reconstruction with semantic segmentation of its composition into buildings, roads, vegetation, and water. Second: closer view of the textured 3D model with overlaid building-level semantic segmentation into facades, windows, balconies, and roofs. Third: distant view of a traffic model estimated for a specific hour in a weekday, based on public webcam footage. Fourth: closer view of the city's liveliness model where avatars simulate the actual data, preserving people's privacy.

ABSTRACT

VarCity – the Video is a short documentary-style CGI movie explaining the main outcomes of the 5-year Computer Vision research project VarCity. Besides a coarse overview of the research, we present the challenges that were faced in its production. These were mainly induced by two factors: i) imperfect raw research data produced by automatic algorithms that had to be included in the movie, and ii) human factors, like federating researchers (and later a CG artist) around a similar goal many had a different conception of, while no one had a detailed overview of all the content. Successive achievement was driven by some ad-hoc technical developments but more importantly of detailed and abundant communication and agreement on common best practices.

ACM Reference format:
Kenneth Vanhoey, Carlos Eduardo Porto de Oliveira, Hayko Riemenschneider, András Bódis-Szomorú, Santiago Manén, Danda Pani Paudel, Michael Gygli, Nikolay Kobyshev, Till Kroeger, Dengxin Dai, Luc Van Gool. 2017. *VarCity – the Video*: the Struggles and Triumphs of Leveraging Fundamental Research Results in a Graphics Video Production. In *Proceedings of SIGGRAPH '17 Talks, Los Angeles, CA, USA, July 30 - August 03, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3084363.3085085

1 INTRODUCTION

VarCity is a 5-year computer vision research project that arrives to an end this year. It involved over a dozen researchers at ETH Zürich in Switzerland and several industrial partners. Its goal is to model the city automatically, based on images that are either available on social media or are captured by dedicated acquisition devices like cars or drones. Besides producing compact and scalable 3D reconstructions of a whole city, a particularity of the project is to augment the data with additional knowledge using machine learning techniques. This results in an understanding of the city's static composition (see semantic segmentation at city or building scale in figure 1), dynamic movement (see, *e.g.*, the traffic and pedestrian models in figure 1) and virtual online life on social media. In turn, this knowledge allows for instant querying of data like, *e.g.*, live monitoring of available parking spots in the streets, automatic measurements of building sizes or wall surfaces, neighborhood visualization for architectural planning or urban design, and so on.

To communicate our work encompassing over 70 scientific papers [1] to the outer public and to our funding agency – the European Research Council – we created a CGI short movie leveraging automatically produced data. In this we faced several challenges induced by the nature of our production. A single versatile 3D artist joined our forces, composed of fundamental researchers only, to produce this animated movie. This includes leveraging our 3D reconstruction and static and dynamic understanding of the full city of Zürich into a documentary-style video. Instead of the usual modeling of assets and animations, we use raw data generated by automated algorithms, hence it is heterogeneous and imperfect:

SIGGRAPH '17 Talks, Los Angeles, CA, USA
© 2017 Copyright held by the owner/author(s). 978-1-4503-5008-2/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3084363.3085085

noisy meshes having millions of polygons, millions of images, estimated camera models, displacement trajectories, traffic models, etc. 3D artists and their tools are not necessarily used to dealing with this data. Conversely, the people producing the data (our researchers) are not experts in non-technical communication, or communication to artists, which lead to organizational and human challenges. We present a brief overview of the underlying research in section 2, and attempt to convey the story of the challenging and original production of *VarCity – the Video* in section 3.

2 COMPUTER VISION RESEARCH

During the 5 years of the project, over 70 scientific computer vision papers were produced [1]. The main concrete outcome is a semantic and dynamic understanding of the city, in the form of automatically produced data after analyzing images acquired by a fly-over, a drive-through, social media or public camera footage. It includes

- city-scale geometric reconstructions from aerial footage, optionally enhanced with street-side footage [3],
- a semantic understanding of the city composition at city-scale (city partitioning) and building scale (building segmentation) [6],
- an understanding of traffic flows from density analysis on public camera footage, both at city-scale (global flows) and street-level (car trajectories and pedestrian densities) [5],
- analysis of large amounts of online resources on social media (*e.g.*, images on Flickr or videos on Youtube) allowing for detection of interesting elements and automatic composition of summary video's of events or landmarks [4].

These elements are useful for many applications [2], like urban planning (roadwork, architectural or emergency planning), real estate valuation, video games and so on. Note that all these algorithms operate *automatically* or semi-automatically by mathematical or artificial intelligence tools (*e.g.*, deep learning) that process input images, and respect people's *privacy by design*. While all our examples are shown in the city of Zürich, these properties allow for repeated and large-scale processing for many cities.

3 THE VIDEO PRODUCTION

3.1 Leveraging unusual data

3D artists are accustomed to using 3D data (like models, textures or cameras) that were either self-designed or designed by other 3D artists using data production software. The main challenges we faced with our data is their inhomogeneous, sometimes imperfect nature and their potential large scale. For example, our city-scale mesh has around 20 million faces, we leverage colored point clouds which are not supported as such in most commercial 3D editing software, and we have to translate conceptual traffic flow models into 3D visualizations.

For most problems, we came up with solutions to interpret the data in commercial 3D rendering software. For example, we registered all our 2D and 3D data in the same unique georegistered reference frame, which facilitates compositing. Conversely, it was not possible to load the point clouds in the rendering software and display them satisfyingly, *i.e.*, without heavy aliasing issues. We resorted to dedicated engineering to produce the animated visualization of 3D point clouds. As a final example, data flows like traffic models and car and pedestrian densities and trajectories were incorporated by intermediate data representations, like spline curves overlaid in 3D space. These were then taken as an input for producing a fancy visualization of that information.

3.2 Human relations

One of the most challenging aspects was handling human relations in the team. Unlike usual video productions, top-down decisions coming from a director are unthinkable in our academic research environment. We traditionally rely on people's will to collaborate and there are rarely top-down decisions being made. Hence editorial decisions had to be made horizontally. Moreover, researchers that produced the data all had their own opinion on how their results should be shown, and how technical the related explanations should be, possibly too technical for a wider audience.

To deal with these issues, clear editorial rules were agreed on early-on. Among this was the mandatory writing by the researchers of a detailed script draft composed of three tracks: i) what is shown: the data and viewpoints shown at a given point in time, ii) what non-technical elements are to be explained with the above, and iii) what technical elements could be highlighted by keywords or bibliographical references. Most importantly, the separation of two levels of technicality was critical. In the following production step, the non-technical elements were incorporated first, which constructed the storyline and often conveyed that more technical elements are maybe not necessary.

4 CONCLUSION

Looking back at this video production, what have we researchers learned about how to produce a video conveying our research to the general public? First, that the storytelling should be determined by people understanding the research on a high level only, but not involved in it directly: the more the research is understood, the more technical the end result is going to be, even subconsciously. Second, we found a good practice to associate everyone to contribute into a common framework that was fixed a priori. Producing consistent data (*e.g.*, georegistered even if this is often seen as useless for the research) or script writing in two levels of technicality are an example of this. Finally, although this is often not possible in publicly-funded institutions, working with a professional CGI artist was crucial in upgrading this movie to a quality level suitable for a wider audience, as he has both the creative mind and the experience to highlight key messages in a subtle way.

REFERENCES

[1] 2017. The VarCity project. (2017). Over 70 research papers and full video available on https://varcity.ethz.ch/.

[2] F. Biljecki, J. Stoter, H. Ledoux, S. Zlatanova, and A. Çöltekin. 2015. Applications of 3D city models: State of the art review. *ISPRS* 4, 4 (2015), 2842–2889.

[3] A. Bódis-Szomorú, H. Riemenschneider, and L. Van Gool. 2016. Efficient Volumetric Fusion of Airborne and Street-side Data for Urban Reconstruction. In *ICPR '16*.

[4] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. 2014. Creating Summaries from User Videos. In *ECCV '14*.

[5] S. Manen, M. Gygli, D. Dai, and L. Van Gool. 2017. PathTrack: Fast Trajectory Annotation with Path Supervision. *ArXiv e-prints* (2017). arXiv:cs.CV/1703.02437

[6] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool. 2014. Learning Where To Classify In Multi-View Semantic Segmentation. In *ECCV '14*.