How to Give a Good Scientific Talk

Dr. Vladislav Golyanik, MPI for Informatics

Computer Vision and Machine Learning
for Computer Graphics

Seminar – Summer Term 2023

max planck institut
informatik


242:2 • Elgharib, M. and Mendiratta, M. et al.
Recently, reenactment algorithms are combining model-based and learning-based scene reconstruction and rendering. They showed the ability to control and modify facial expressions and poses in video filmed from a frontal perspective in a highly photo-realistic manner [Kim et al. 2018; Nagano et al. 2018; Suwajanakorn et al. 2017; Thies et al. 2016]. They have also been successfully employed in various applications, such as avatar generation [Thies et al. 2019], visual dubbing [Kim et al. 2019], video post-production [Fried et al. 2019], and virtual assistant generation [Thies et al. 2020]. Similar methods were also used to improve virtual reality teleconferencing with head-mounted displays (HMDs). Here, it was shown that reenactment techniques, in combination with additional sensors attached to the HMD [Burgos-Artizzu et al. 2015] can remove the occlusion the display from a user's face, and thus portray an unobstructed view of the user or her avatar to every participant of the VR teleconference [Lombardi et al. 2018, 2019; Olszewski et al. 2016; Thies et al. 2018a]. However, wearing such non-see-through VR HMDs while walking around is quite inconvenient, to impractical.
We, therefore, present a new approach to enable people to start egocentric video conferencing, i.e., calls where they are mobile and in indoor and outdoor environments. We want our approach to be similarly convenient and non-encumbering as audio calls with a headset, while additionally transmitting a frontal video image of the person. To this end, we propose an egocentrically worn hardware setup and a new algorithm to achieve this goal. Our approach uses a single commodity RGB fish-eye camera mounted to the side of an ordinary VR headset. Since the camera observes the user's face from a side perspective, we argue that mere production of future smart and augmented reality glasses will make it easy to integrate stately nonmastered cameras in this way. The related designs of Google Glasses or the Snapchat Spectacles have shown this. While a camera placed in this way minimizes obstruction to the user's field of view, it records a highly distorted and incomplete facial view that is not directly useful for applications in a videoconferencing application (see Fig. 1, input). We therefore present a new conditional generative translation network that learns the mapping from the highly distorted egocentric view as input. An adversarial loss that operates on a sequence of estimated frontalized images ensures temporal consistency while a perceptual loss is employed to produce high fidelity results.
Our frontalization algorithm purposefully refrains from estimating an intermediate representation of the full face performance on the basis of a 3D face model, as it is done in the majority of the many previous face reenactment methods [Kim et al. 2018; Thies et al. 2019]. 3DMMs usually lack the variable dimensions to represent all fine-grained nuances of eye gaze, eye blinks, facial micro-expressions, or eyebrow gestures and tongue motion, which are important for conveying realism. Moreover, 3DMMs are computationally even if these dimensions were parametrically modeled, estimating them from a starkly oblique and distorted view is non-trivial. Therefore, our frontalization method only uses weak conditioning with a neutral 3D face model without face expressions, and transfers the fine-grained expression details from the egocentric view to the frontal view by means of learned direct video-to-video mapping.
Our lateral Sidney-to-frontal transfer method solves a much harder problem than established frontalization settings [Cao et al. 2019; Peng et al. 2017; Yin et al. 2017; Zhang et al. 2019], and utilize learned frontalization algorithms [Hassner et al. 2015; Huang et al. 2017], the frontalization network also learns to disentangle the identity from the head pose. Yin et al. [2017] use a GAN conditioned on the the frontalized synthetic rendering to produce photo-realistic results. The output is also constrained to maintain the identity of the examined image. The approach of Cao et al. [Cao et al. 2018, 2019] uses a 3D representation to handle pose variations in the input image or map. A discriminator loss learns to differentiate between frontal and non-frontal views.
Image translation-based techniques [Isagunar et al. 2015; Wiles et al. 2018; Zhang et al. 2019] learn to frontalize faces without a parametric 3D model. This bypasses some limitations imposed by the restricted expressiveness of learned parametric expression and shape models [Cao et al. 2013; Li et al. 2017]. Wiles et al. [2018] frontal faces have the minimum rank across all poses. They formulate face frontalization as an optimization problem with nuclear norm minimization. A statistical prior is learned from frontal images. Zhang et al. [Zhang et al. 2019] propose a flow-based approach for face frontalization. A flow field is initialized by DFF-flow [Liu et al. 2008] and refined through a convolutional neural network. Most recently, Hassner et al. [2015] and Huang et al. [2017] employ a deep encryption technique and suit use producing photo-realistic and temporally coherent video outputs [Cao et al. 2018, 2019; Sagonas et al. 2015; Yin et al. 2017; Zhang et al. 2019]. In contrast, our method produces photo-realistic frontalizations that are temporally coherent. It does so by bounding between a strongly distorted fish-eye camera and a more regular frontal camera (see Fig. 1).
Face reenactment. Recent reenactment methods [Kim et al. 2017; Wiles et al. 2018] can also be used for frontalization. X2face [Wiles et al. 2018] trains a network to extract a face embedding based on a single image. The extracted latent code is used to synthesize a new image of the target face with new expressions. Isola et al.'s pix2pix [Isola et al. 2017] uses a Conditional Generative Adversarial Network (CGAN) to translate an input image from one domain to another one. Recently, this approach was used for many face-to-face translation tasks [Kim et al. 2018; Thies et al. 2020, 2016]. Unlike face frontalization, reenactment focuses more on photo-realistic editing in the same camera perspective.
2.1 Face Frontalization
Face frontalization techniques are commonly designed to transform large or profile face poses in a camera view, where larger parts of the face are occluded, into complete and frontal views of the face. Existing approaches can be divided into face model based [Cao et al. 2018, 2019; Hassner et al. 2015; Peng et al. 2017; Yin et al. 2017; Zhang et al. 2019] and image translation based [Isola et al. 2017; Sagonas et al. 2015; Wiles et al. 2018; Zhang et al. 2019]. Model based techniques use a parametric 3D Morphable Model (3DMM) to represent faces [Garrido et al. 2016]. Such model provides a parametric control over the head pose and hence allows frontalizing for face as observed from a front-looking camera. Zhu et al. [Zhu et al. 2016] and Peng et al. [Peng et al. 2017] learn the parameters of the face model from the input face using a deep neural network. The network is trained using pairs of profile and frontal view faces. Such data is obtained by synthesizing the profile images from the frontal images. In Peng et al. [Peng et al. 2017], the frontalization network also learns to disentangle the identity from the head pose. Yin et al. [2017] use a GAN conditioned on the the frontalized synthetic rendering to produce photo-realistic results. The output is also constrained to maintain the identity of the examined image. The approach of Cao et al. [Cao et al. 2018, 2019] uses a 3D representation to handle pose variations in the input image or map. A discriminator loss learns to differentiate between frontal and non-frontal views.
To summarize, we make the following contributions:
• A light-weight capturing setup that enables hands-free video-conferencing and is easy to be integrated into smart glasses.
• A real-time video-to-video translation technique that uses a new conditional neural network adversarially trained with a temporal discriminator to transfer even subtle face expressions and natural eye motion from an egocentric fisheye to a frontal view.
• A thorough analysis showing that our approach reconstructs frontal video 5x5 more accurately than established image-to-image translation methods (pix2pix [Isola et al. 2017]), visually outperforms 3DMM-regression based solutions, and runs in real time at 29.4 ms per frame.
2 RELATED WORK
We survey computer graphics and vision techniques that can potentially enable video-conferencing in our scenario of a an egocentric camera. We also discuss the most related face frontalization, model-based and reenactment-based approaches. Frontalization techniques produce a frontal complete view of the face from an incomplete side view, while reenactment methods transfer facial motions onto prerecorded video [Kim et al. 2018; Thies et al. 2020, 2016]. Unlike face frontalization, reenactment focuses more on photo-realistic editing in the same camera perspective.
2.1 Face Frontalization
Face frontalization techniques are commonly designed to transform large or profile face poses in a camera view, where larger parts of the face are occluded, into complete and frontal views of the face. Existing approaches can be divided into face model based [Cao et al. 2018, 2019; Hassner et al. 2015; Peng et al. 2017; Yin et al. 2017; Zhang et al. 2019] and image translation based [Isola et al. 2017; Sagonas et al. 2015; Wiles et al. 2018; Zhang et al. 2019]. Model based techniques use a parametric 3D Morphable Model (3DMM) to represent faces [Garrido et al. 2016]. Such model provides a parametric control over the head pose and hence allows frontalizing for face as observed from a front-looking camera. Zhu et al. [Zhu et al. 2016] and Peng et al. [Peng et al. 2017] learn the parameters of the face model from the input face using a deep neural network. The network is trained using pairs of profile and
Egocentric Videoconferencing • 242:3
inspired by pix2pix [Isola et al. 2017] and uses a U-Net architecture as a generator that gets synthetic renderings of the underlying 3DMM as input. Kim et al. [Kim et al. 2019] presented a reenactment technique which maintains the speaking style of the target identity. The work shows the importance of this feature during visual dubbing. pix2GAN [Nagano et al. 2018] generates a person aligned avatar from a single image. A translation network trained on synthetic renderings is used. The work also shows the importance of rendering and its corresponding photo-real version. Deferred Neural Rendering [Thies et al. 2019] introduced neural textures which store high-dimensional neural descriptors. Such textures are interpreted by a neural network to produce a photo-realistic output. Both, neural textures and the interpreting network are trained in conjunction based on a short target video sequence, where the 3DMM face model is reconstructed based on tracked facial landmarks [Thies et al. 2016]. While the approach is general and can be applied to novel view point synthesis and scene editing, they also demonstrate high quality facial re-rendering and reenactment. In contrast to the model-based approaches, there are also techniques that are not relying on a 3DMM face model prior. Based on a sophisticated multi-camera setup, Lombardi et al. [Lombardi et al. 2018] learn a deep appearance model. They learn a latent code from a view-dependent texture descriptor that is interpreted by a decoder network. As follow-up work they show the ability of driving a highly photo-realistic avatar through a VR headset [Wei et al. 2019].
Recently audio and text driven reenactment approaches have been proposed. Here facial expressions are not extracted from a source video, but rather from a source audio or a text script. Fried et al. [Fried et al. 2019] presented a system that edits a video according to the expression parameters of the 3DMM. This allows text-based reenactment in photo-realistic and temporally smooth manner. Several approaches for audio-driven reenactment are available [Chung et al. 2017; Olszewski et al. 2016; Suwajanakorn et al. 2017]. Thies et al. 2020], Thies et al. [Thies et al. 2020] presented Neural Voice Puppetry, an approach for estimating facial expressions from the audio signal based on a latent expression space of a deep audio-to-Speech [Hannun et al. 2014] to produce character-agnostic frame-by-input audio. A network then translates the login into the parameters of a blendshape expression model. A synthetic rendering of a face model is produced, followed by neural rendering for photo-realistic results. Suwajanakorn et al. [Suwajanakorn et al. 2017] define the mouth shape with a number of keypoints, for which they regress their positions from only the audio signal. With proper image compositing of the predicted mouth shapes, they show impressive high quality lip-synchronization for the supervised target actor. Karras et al. [Karras et al. 2017] and Taylor et al. [Taylor et al. 2016] used the audio signal to assist performance capture from a VR-headset. Here, audio-based alignment techniques are used to map same utterance from different subjects into the same animation parameters. We show that purely audio driven solutions do not suffer in our egocentric videoconferencing setting since important non-verbal expression only appear on video.
3 DATA RECORDING
We propose a recording setup for obtaining the training data of our solution (see Fig. 2-a). The training data consists of paired egocentric and frontal views of the same person recorded in sync. The egocentric video are temporally synchronized with the frontal video at calibration stage. We have two data recording setups: one for a dynamic scenario and another for a sitting scenario (see Fig. 2-a, top row). They both use the same egocentric camera but differ in the setup of the frontal camera.
Egocentric Camera: We use a low-cost RGB fish-eye camera to capture the facial expression (ELP-usb8mp02m-l180). The camera is attached to the frame of an eye-glass such that it minimizes obstruction to the field of view. The camera records a video stream (see Fig. 1-b). It has a diagonal field of view of 180 degrees, and records images with resolutions up to 1280 × 1960 at 30fps. Fig. 2-c shows data samples captured by the egocentric camera. During test time, these views are the input to our method. Specifically, our algorithm learns to estimate the corresponding full face frontalization. This learning is supervised by a frontal camera.
Frontal Camera: We capture the frontal RGB frontal camera placed in front of the user to capture the face from the target perspective. All our experiments in the dynamic scenario are shot using a commodity mobile phone camera. A commodity HD camera is used for recording our experiments in the sitting scenario. The dynamic scenario resembles situations where the user is moving around in an environment with changing illuminations and background. In this case, the frontal camera is mounted on a tripod. The frontal camera is a way to allow good face coverage (see Fig. 2-a, top-left), capturing the face from a frontal view at a fixed location with respect to the face. In the sitting scenario we place the supervising camera on a tripod, and the user sits in front of it (see Fig. 2-a top-right).
In many sequences the user reads a collection of 11 english phonetically balanced sentences [Harvard 2019] displayed on the laptop screen. Each paragraph is a sentence containing all the 16 Latin letters. This captures a wide variety of visemes commonly used in everyday speeches. In other sequences, subjects were asked to talk freely, imitating a phone call, discussing a popular topic and so on. In the dynamic scenario the user walks, in either outdoor or indoor environment. In the sitting scenario the user sits on a chair in front of a desk in an office room. Fig. 2-a, bottom row, shows egocentric and frontal camera recordings, we use a transient event, i.e., a mobile phone screen observed from both cameras plays a video of sassily black frames, but with a single white frame every 10 seconds (see Fig. 2-a, bottom, zoom on region). We start recording from both cameras and wait until the white frame is observed. We use this white frame to temporally synchronize the egocentric and frontal recordings. The calibration is only done once at the start of the data recording.
We recorded 27 sequences with 13 identities, and extracted them at 24 frames per second. Sequences are on average 14000 frames long. The original resolution of the egocentric view is 1280 × 1960 while the original resolution for the frontal view is usually 1920 × 1080. We manually take a tight crop around the face for both videos and resize the resolution to 256 × 256 while maintaining the aspect ratio. In all experiments and comparisons we use 7500 frames for training.

...

After you have read and understood the papers...

...

After you have read and understood the papers...

...you will have to present them.

max planck institut informatik

Outline

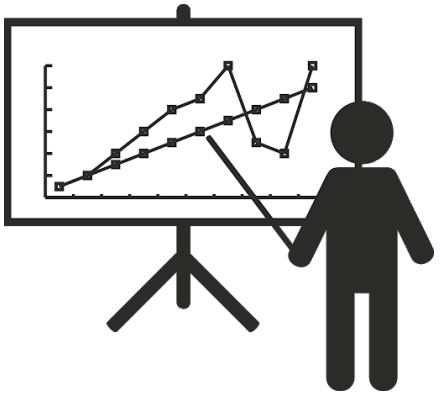
* Structuring your story
* Preparing your data and information
* Preparing and giving the presentation
* Concluding your presentation
* Questions and answers

Presentation Structure :: Basic Rule

* Say what you are going to say (introduction)
* Say it (give the core talk)
* Say what you said (summarise and conclude)

Images: http://blogs.nature.com/naturejobs/2017/01/11/scientific-presentations-a-cheat-sheet/
https://www.barnesandnoble.com/blog/7-books-you-wont-believe-are-nonfiction/

Presentation Structure :: Basic Rule



* Say what you are going to say (introduction)
* Say it (give the core talk)
* Say what you said (summarise and conclude)

vs

This is about scientific findings and implications:
Do not try building suspense and then unveiling a surprise ending.

Images: http://blogs.nature.com/naturejobs/2017/01/11/scientific-presentations-a-cheat-sheet/
https://www.barnesandnoble.com/blog/7-books-you-wont-believe-are-nonfiction/

The Story

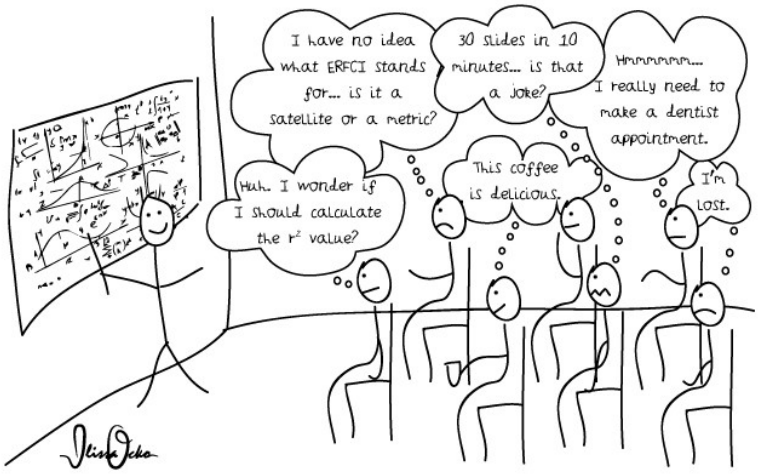
* Structure and tell the story logically
* Exemplary structure of the presentation:
  + Title page (title, date, authors, venue, acknowledgements)
  + Seminar specifics: recap of the previous topic
  + Introduction / Motivation (including an overview and related works)
  + Approach (technical details of the method, maths)
  + Experimental Results (including evaluation methodology, interpretation of the results and discussion)
  + Conclusion (summary and core implications)

The Story

* Structure and tell the story logically
* Exemplary structure of the presentation:
  + Title page (title, date, authors, venue, acknowledgements)
  + Seminar specifics: recap of the previous topic
  + Introduction / Motivation (including an overview and related works)
  + Approach (technical details of the method, maths)
  + Experimental Results (including evaluation methodology, interpretation of the results and discussion)
  + Conclusion (summary and core implications)

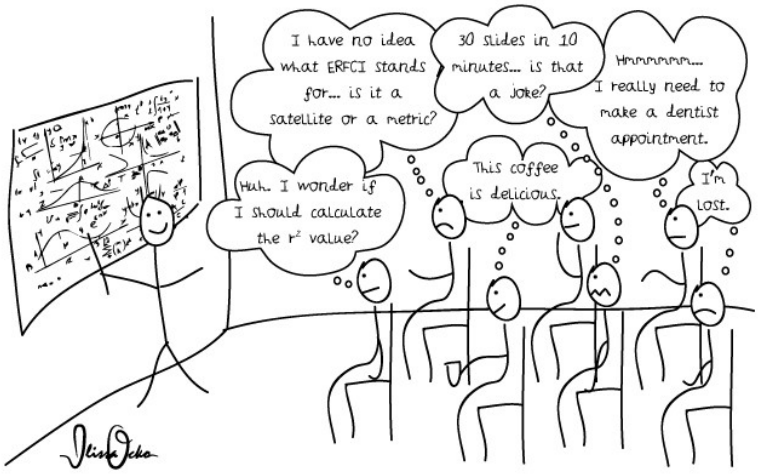
Audience

* Why are you giving this presentation?
* To whom are you giving this presentation?
* What are your expectations from that talk?
* What are the expectations of the audience?
* Is the presentation live or online?
* How much time do I have?



Audience

* Why are you giving this presentation?
* To whom are you giving this presentation?
* What are your expectations from that talk?
* What are the expectations of the audience?
* Is the presentation live or online?
* How much time do I have?

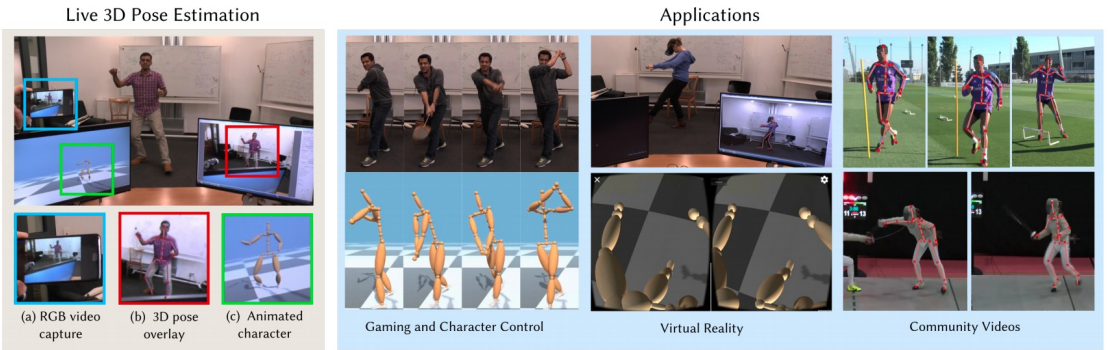


+ Keep that in mind while preparing the talk
+ Edit / adjust the slides

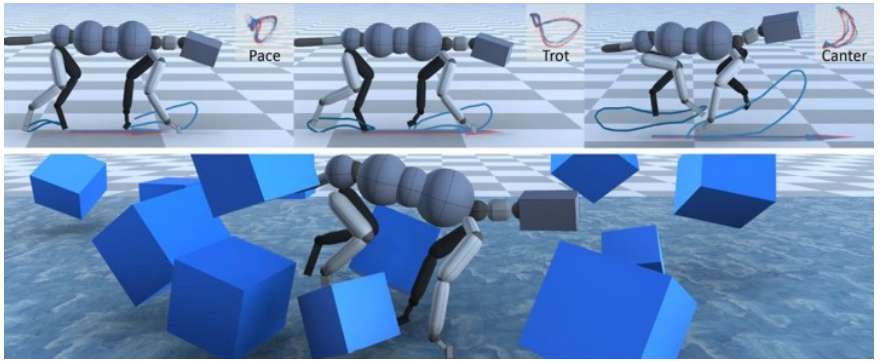
Audience :: University Seminar

* Audience with broad technical background
* Many topics: Provide an overview of state of the art
* Message:
    + Why the problem is important
    + Why the proposed solution is novel and impactful?
    + What are the main ideas and insights?
    + “Being a graduate student”: discussion, ideas for improvement
    + To include a slide or not: How important is it for the story? Will the audience understand and value the point?

Preparing the Talk :: Overview Figures



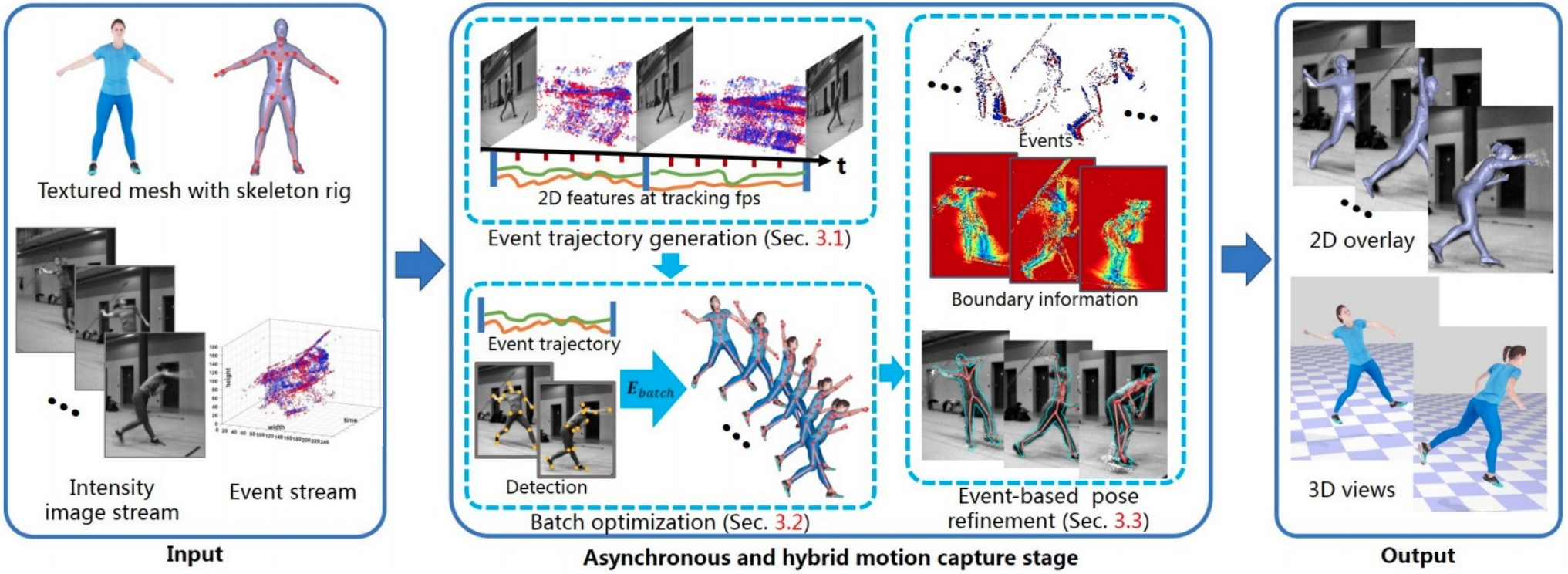
Mehta *et al.*, SIGGRAPH 2017.



Luo *et al.*, SIGGRAPH 2020.

* A figure with a summary of findings, overview of the method, problem or a core concept
* Helps to motivate why the problem is important
* If you use web sources, reference the source

Preparing the Talk :: Overview Figures



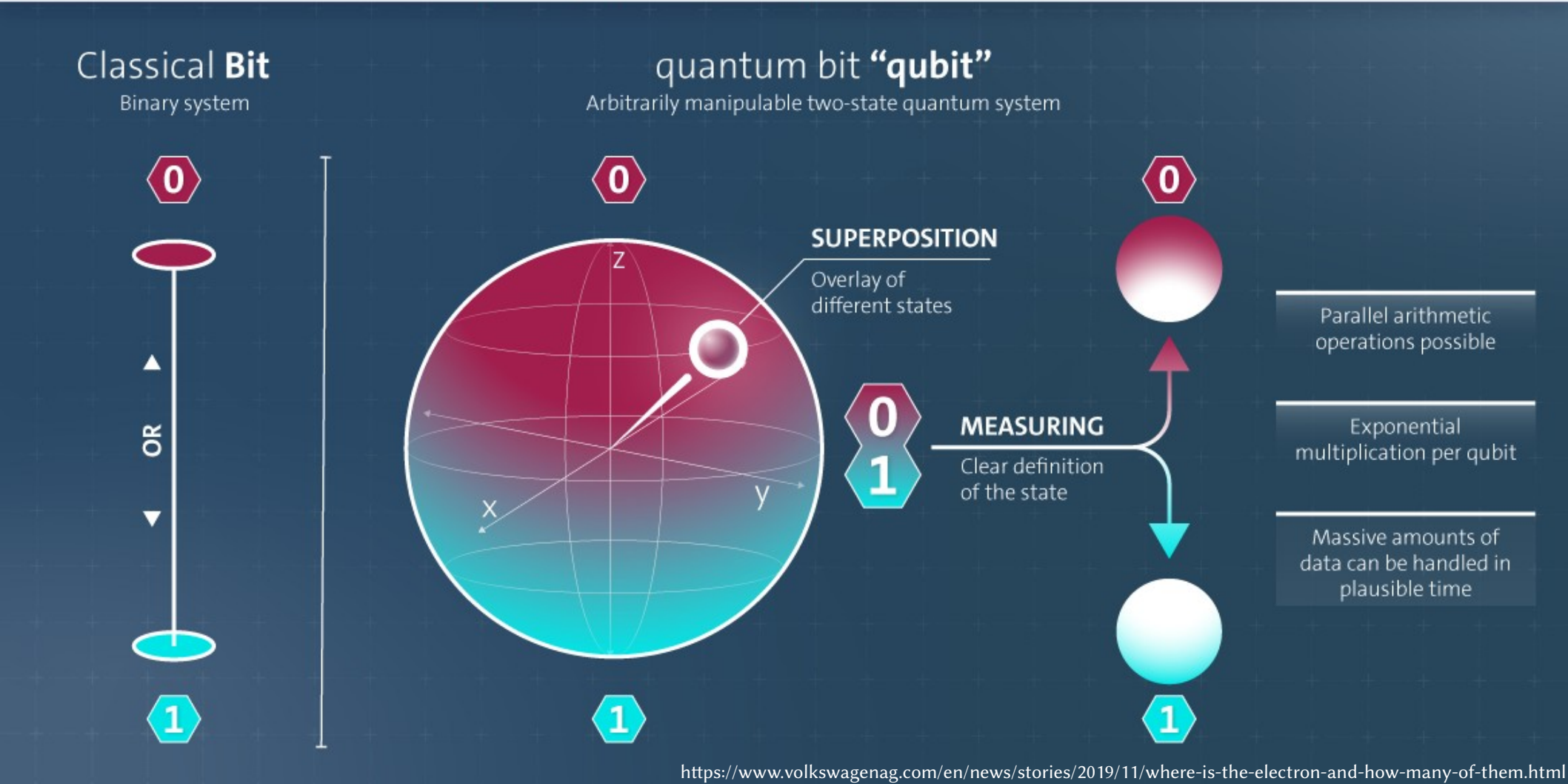
Xu *et al.*, CVPR 2020.

Example :: What is a Qubit?

Qubit. Quantum computing encompasses tasks which can be performed on quantum-mechanical systems [53]. Quantum *superposition* and *entanglement* are two forms of parallelism evidenced in quantum computers. A *qubit* is a quantum-mechanical equivalent of a classical bit. A qubit $|\phi\rangle$ — written in the *Dirac* notation — can be in the state $|0\rangle$, $|1\rangle$ or an arbitrary *superposition of both states* denoted by $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$, where α and β are the (generally, complex) probability amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$. In quantum computing, the state $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$ denoted by $|+\rangle$ is often used for initialisation of a qubit register. The state of a qubit remains hidden during the entire computation and reveals when measured. If qubits are *entangled*, measuring one of them influences the measurement outcome of the other one [59]. During the measurement, the qubit's state irreversibly collapses to one of the basis states $|0\rangle$ or $|1\rangle$. Efficient physical realisation of a qubit demand very low temperatures. Otherwise, thermal fluctuations will destroy it and lead to arbitrary changes of the measured qubit state.

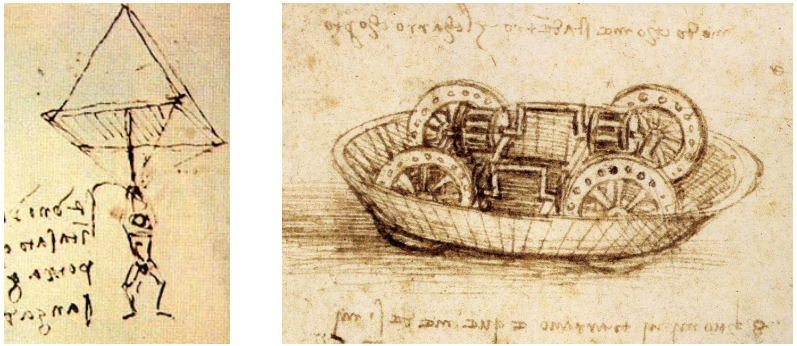
HOW A QUANTUM COMPUTER WORKS

Principle of superposition allows parallelism in the calculations

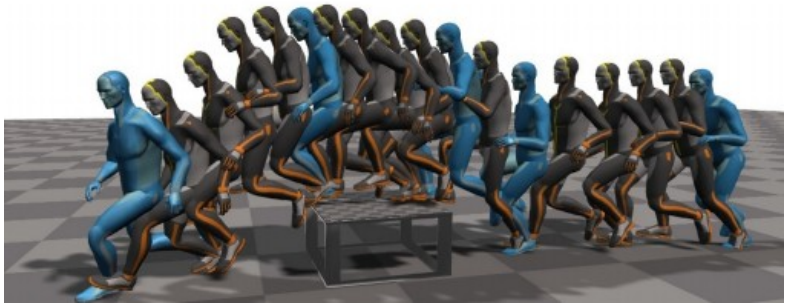


https://www.volkswagenag.com/en/news/stories/2019/11/where-is-the-electron-and-how-many-of-them.html

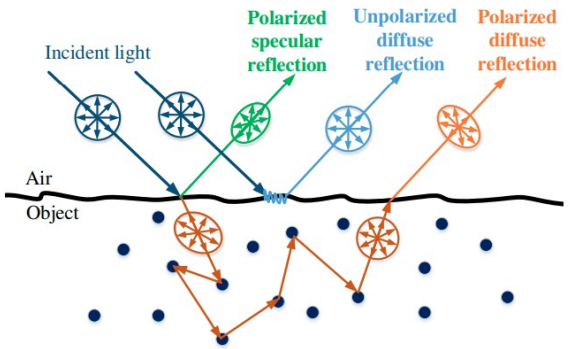
Example :: Overview Figures



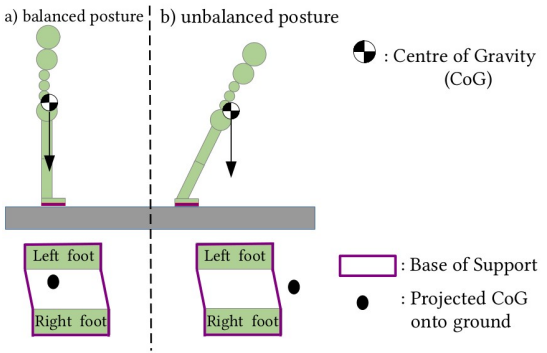
Technical Drawings of da Vinci.



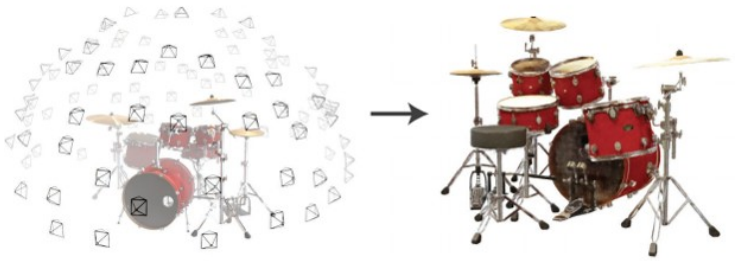
Harvey *et al.*, SIGGRAPH 2020.



Cui *et al.*, CVPR 2017.



Shimada *et al.*, SIG'ASIA, 2020.



Mildenhall *et al.*, ECCV, 2020.

Using Tables

date	discharge (cf/s)	precipitation (in/day)
1-Nov	631	0
2-Nov	808	0
3-Nov	794	0.08
4-Nov	826	0
5-Nov	1060	1.09
6-Nov	1080	0.48
7-Nov	1040	0.28
8-Nov	779	0
9-Nov	686	0
10-Nov	670	0
11-Nov	696	0.53
12-Nov	831	0.23
13-Nov	985	0.45
14-Nov	1080	0.14
15-Nov	1350	0.65
16-Nov	1430	0
17-Nov	2440	1.6
18-Nov	2280	0
19-Nov	2040	0
20-Nov	1830	0.55
21-Nov	1650	0
22-Nov	1560	0
23-Nov	1520	0.39
24-Nov	1410	0
25-Nov	1320	0
26-Nov	1310	0.11
27-Nov	1450	0.78
28-Nov	1560	0.22
29-Nov	1550	0.45
30-Nov	1480	0

date	discharge (cf/s)	precipitation (in/day)
1-Dec	1480	0.07
2-Dec	2920	0.96
3-Dec	2380	0
4-Dec	1990	0
5-Dec	1770	0
6-Dec	1620	0.1
7-Dec	1500	0
8-Dec	1420	0
9-Dec	1350	0
10-Dec	1290	0
11-Dec	1280	0.1
12-Dec	1330	0.47
13-Dec	1280	0
14-Dec	1250	0.57
15-Dec	1190	0.04
16-Dec	1180	0
17-Dec	1160	0.17
18-Dec	1120	0.01
19-Dec	1080	0
20-Dec	1070	0
21-Dec	1080	0
22-Dec	1060	0
23-Dec	1060	0.18
24-Dec	1050	0
25-Dec	1050	0.5
26-Dec	986	0
27-Dec	1010	0
28-Dec	1010	0.07
29-Dec	977	0
30-Dec	972	0
31-Dec	957	0

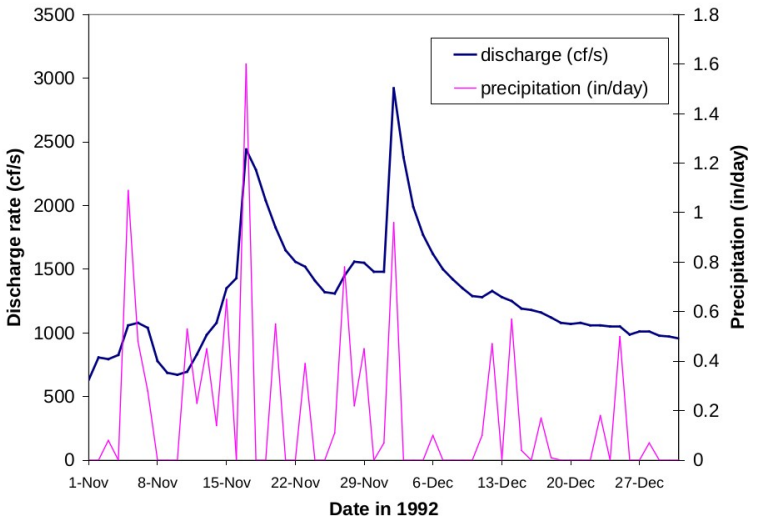
Discharge of the Esopus Creek (Coldbrook, NY) and precipitation at Slide Mountain, NY (source: USGS/NCDC)

Using Tables

date	discharge (cf/s)	precipitation (in/day)
1-Nov	631	0
2-Nov	808	0
3-Nov	794	0.08
4-Nov	826	0
5-Nov	1060	1.09
6-Nov	1080	0.48
7-Nov	1040	0.28
8-Nov	779	0
9-Nov	686	0
10-Nov	670	0
11-Nov	696	0.53
12-Nov	831	0.23
13-Nov	985	0.45
14-Nov	1080	0.14
15-Nov	1350	0.65
16-Nov	1430	0
17-Nov	2440	1.6
18-Nov	2280	0
19-Nov	2040	0
20-Nov	1830	0.55
21-Nov	1650	0
22-Nov	1560	0
23-Nov	1520	0.39
24-Nov	1410	0
25-Nov	1320	0
26-Nov	1310	0.11
27-Nov	1450	0.78
28-Nov	1560	0.22
29-Nov	1550	0.45
30-Nov	1480	0

date	discharge (cf/s)	precipitation (in/day)
1-Dec	1480	0.07
2-Dec	2920	0.96
3-Dec	2380	0
4-Dec	1990	0
5-Dec	1770	0
6-Dec	1620	0.1
7-Dec	1500	0
8-Dec	1420	0
9-Dec	1350	0
10-Dec	1290	0
11-Dec	1280	0.1
12-Dec	1330	0.47
13-Dec	1280	0
14-Dec	1250	0.57
15-Dec	1190	0.04
16-Dec	1180	0
17-Dec	1160	0.17
18-Dec	1120	0.01
19-Dec	1080	0
20-Dec	1070	0
21-Dec	1080	0
22-Dec	1060	0
23-Dec	1060	0.18
24-Dec	1050	0
25-Dec	1050	0.5
26-Dec	986	0
27-Dec	1010	0
28-Dec	1010	0.07
29-Dec	977	0
30-Dec	972	0
31-Dec	957	0

vs



Discharge of the Esopus Creek (Coldbrook, NY) and precipitation at Slide Mountain, NY (source: USGS/NCDC)

Using Maths

$$\mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^{|Z|}, \mathbf{w}) = \sum_{\zeta \in Z} \alpha_\zeta \, \mathfrak{E}_{\text{data}}(\mathbf{T}^\zeta) +$$
$$+ \sum_{\zeta \in Z} \beta_\zeta \, \mathfrak{E}_{\text{pICP}}(\mathbf{T}^\zeta) \; + \gamma_\zeta \sum_{\zeta \in Z} \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^\zeta, \mathbf{w}) +$$
$$+ \eta \, \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|Z|} \lambda_\zeta \, \mathfrak{E}_{\text{c.}}(\mathbf{T}^\zeta).$$

Using Maths

$$\mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^{|Z|}, \mathbf{w}) = \sum_{\zeta \in Z} \alpha_\zeta \, \mathfrak{E}_{\text{data}}(\mathbf{T}^\zeta) +$$
$$+ \sum_{\zeta \in Z} \beta_\zeta \, \mathfrak{E}_{\text{pICP}}(\mathbf{T}^\zeta) \; + \gamma_\zeta \sum_{\zeta \in Z} \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^\zeta, \mathbf{w}) +$$
$$+ \eta \, \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|Z|} \lambda_\zeta \, \mathfrak{E}_{\text{c.}}(\mathbf{T}^\zeta).$$

vector of frame-to-frame segment transformations

segment-to-segment connectivity weights

brightness constancy

$$\mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^{|Z|}, \mathbf{w}) = \sum_{\zeta \in Z} \alpha_\zeta \, \mathfrak{E}_{\text{data}}(\mathbf{T}^\zeta) +$$

projective ICP

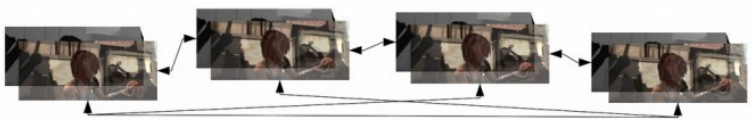
$$+ \sum_{\zeta \in Z} \beta_\zeta \, \mathfrak{E}_{\text{pICP}}(\mathbf{T}^\zeta) \; + \sum_{\zeta \in Z} \gamma_\zeta \, \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^\zeta, \mathbf{w}) +$$

lifted segment pose regularizer

$$+ \eta \, \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|Z|} \lambda_\zeta \, \mathfrak{E}_{\text{c.}}(\mathbf{T}^\zeta).$$

robust weight optimizer

multiframe pose concatenation



optimization over multiple frames

Using Maths

$$\mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^{|Z|}, \mathbf{w}) = \sum_{\zeta \in Z} \alpha_\zeta \, \mathfrak{E}_{\text{data}}(\mathbf{T}^\zeta) + \\ + \sum_{\zeta \in Z} \beta_\zeta \, \mathfrak{E}_{\text{pICP}}(\mathbf{T}^\zeta) \; + \gamma_\zeta \sum_{\zeta \in Z} \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^\zeta, \mathbf{w}) + \\ + \eta \, \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|Z|} \lambda_\zeta \, \mathfrak{E}_{\text{c.}}(\mathbf{T}^\zeta).$$

vector of frame-to-frame segment transformations

segment-to-segment connectivity weights

brightness constancy

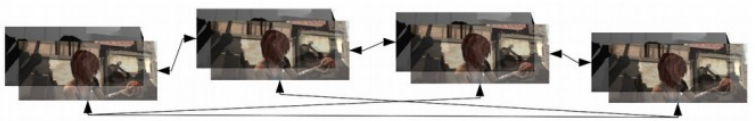
$$\mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^{|Z|}, \mathbf{w}) = \sum_{\zeta \in Z} \alpha_\zeta \, \mathfrak{E}_{\text{data}}(\mathbf{T}^\zeta) + \\ + \sum_{\zeta \in Z} \beta_\zeta \, \mathfrak{E}_{\text{pICP}}(\mathbf{T}^\zeta) \; + \sum_{\zeta \in Z} \gamma_\zeta \, \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^\zeta, \mathbf{w}) + \\ + \eta \, \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|Z|} \lambda_\zeta \, \mathfrak{E}_{\text{c.}}(\mathbf{T}^\zeta).$$

projective ICP

lifted segment pose regularizer

robust weight optimizer

multiframe pose concatenation

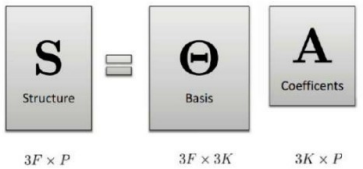


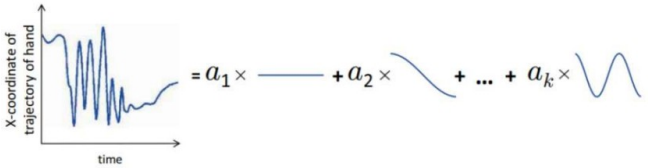
optimization over multiple frames

* use equations at little as possible and as much as needed

Using Maths

$$\mathbf{E}_{\text{traj}}(\boldsymbol{\theta}, \mathbf{z}) = \left\| (\mathbf{1}_T \otimes \bar{\mathbf{S}}) + f_{\boldsymbol{\theta}}(\mathbf{z}) - (\boldsymbol{\Phi} \otimes \mathbf{I}_3)\mathbf{A} \right\|_\epsilon, \quad \boldsymbol{\Phi} = \begin{pmatrix} \phi_{1,1} & \cdots & \phi_{1,K} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \cdots & \phi_{T,K} \end{pmatrix}$$

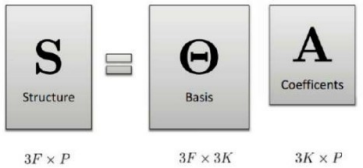


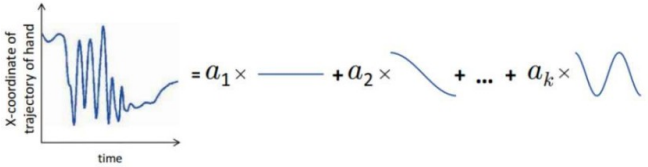


Sidhu *et al.*, ECCV, 2020.

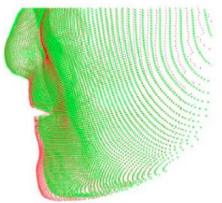
Using Maths

$$\mathbf{E}_{\text{traj}}(\boldsymbol{\theta}, \mathbf{z}) = \left\| (\mathbf{1}_T \otimes \bar{\mathbf{S}}) + f_{\boldsymbol{\theta}}(\mathbf{z}) - (\boldsymbol{\Phi} \otimes \mathbf{I}_3)\mathbf{A} \right\|_{\epsilon}, \quad \boldsymbol{\Phi} = \begin{pmatrix} \phi_{1,1} & \cdots & \phi_{1,K} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \cdots & \phi_{T,K} \end{pmatrix}$$





$$\mathbf{E}_{\text{spat}}(\boldsymbol{\theta}, \mathbf{z}) = \underbrace{\sum_{t=0}^{T-1} \sum_{\mathbf{p} \in \mathbf{S}_t} \left\| \mathbf{p} - \frac{1}{|\mathcal{N}(\mathbf{p})|} \sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} \mathbf{q} \right\|_1}_{\text{Laplacian smoothing}} - \lambda \underbrace{\sum_{t=1}^{T} \|\mathcal{P}_z(\mathbf{G}_t \mathbf{S}_t)\|_2}_{\text{depth control}}$$



Sidhu *et al.*, ECCV, 2020.

General Rule :: Presenting Methodology

* A scientific talk is always about

HOW and WHY

* Explain what you do
* What is new and innovative
* **AND** motivate why this is the way to go

General Rule :: Presenting Methodology

* A scientific talk is always about

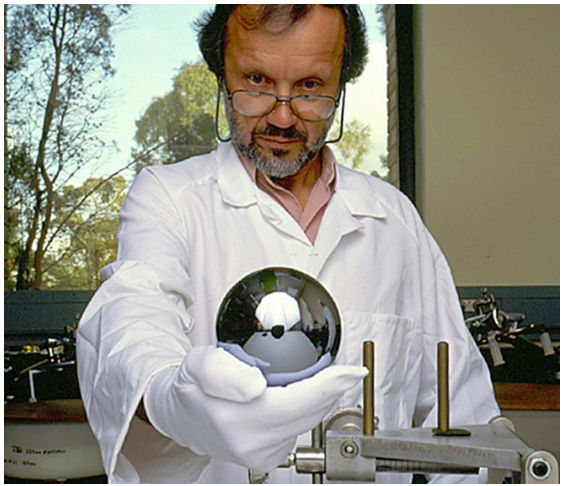
HOW and WHY

* Explain what you do
* What is new and innovative
* **AND** motivate why this is the way to go

THIS INFLUENCES THE STORY

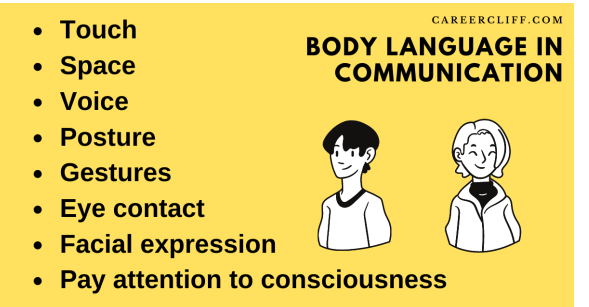
Preparing and Polishing Presentation

* Use 3-7 bullets per page
  + avoid writing out complete sentences
* No more than one minute per slide on average
* Check the slide appearance consistency
* No sound unless it is part of results
* Videos are often results in visual computing
* Spelling and writing style
  + Use the same font (or a few fonts)
  + Check the text for typos; check the grammar
  + Decide between British and American English, and use the chosen language consistently



https://www.wired.com/2010/10/platinum-silicon-kilogram-standard/

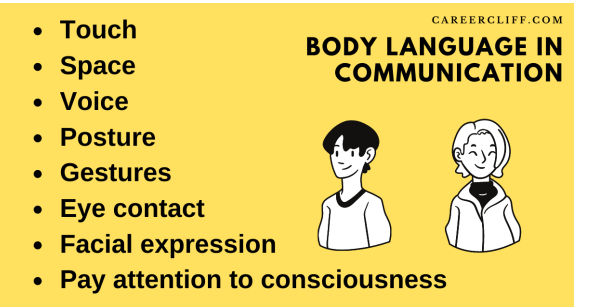
Preparing Yourself



* The way how you present yourself is as important as your slides
* Immerse yourself in what you are going to say
* Make sure that you are familiar with the equipment, check your equipment (microphone, *etc.*)

Image: https://medium.com/@indiayuvagi/body-language-in-daily-life-ac75ff6f2453

Preparing Yourself



* The way how you present yourself is as important as your slides
* Immerse yourself in what you are going to say
* Make sure that you are familiar with the equipment, check your equipment (microphone, *etc.*)
* Perception of gestures and body language
  + Eye contact with the audience
  + Use intonation in combination with the visual tools (*e.g.*, colours)
  + Rehearsing is very important! Be on time, know what you want to say, prepare transitions between the slides/papers

Image: https://medium.com/@indiayuvagi/body-language-in-daily-life-ac75ff6f2453

Rehearsing

* Practice – actually stand up and say the words out loud
    - discover what you do not understand and develop a natural flow
* Do not memorise the talk, do not over-rehearse
* Stay within the time limit
* The Feynman Technique: *a mental model and a breakdown of the thought process to convey information using concise thoughts and simple language [1].*

[1] https://medium.com/taking-note/learning-from-the-feynman-technique-5373014ad230

Rehearsing

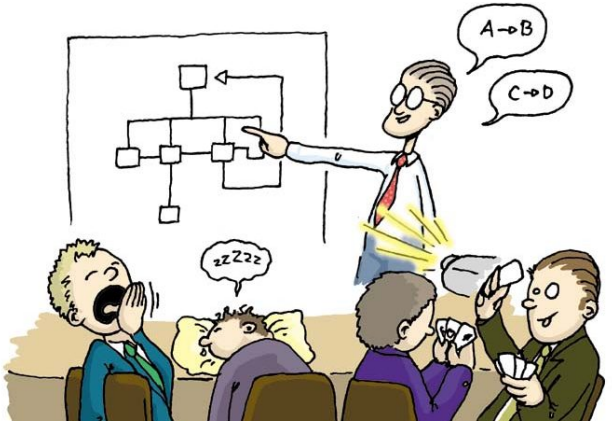
* Practice – actually stand up and say the words out loud
  - discover what you do not understand and develop a natural flow
* Do not memorise the talk, do not over-rehearse
* Stay within the time limit
* The Feynman Technique: *a mental model and a breakdown of the thought process to convey information using concise thoughts and simple language [1].*

If you can't explain it simply, you don't understand it well enough.

A. Einstein.

[1] https://medium.com/taking-note/learning-from-the-feynman-technique-5373014ad230

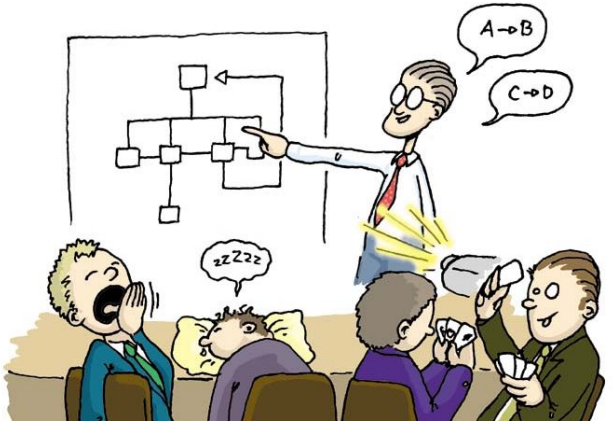
Presenting



* Make yourself comfortable, speak freely, be enthusiastic but do not rush
* Ensure that people can hear you well and see your shared screen

https://mindthegraph.com/blog/scientific-presentation/
https://www.business2community.com/communications/z-mindful-presenting-high-impact-public-speaking-presentation-skills-b-brevity-01949085

Presenting



* Make yourself comfortable, speak freely, be enthusiastic but do not rush
* Ensure that people can hear you well and see your shared screen
* Starting is the most difficult part
  + memorise the first lines
* Nervousness is normal, don't worry about stopping to think

https://mindthegraph.com/blog/scientific-presentation/
https://www.business2community.com/communications/z-mindful-presenting-high-impact-public-speaking-presentation-skills-b-brevity-01949085

Concluding the Presentation

* Announce the ending so that people are prepared
* Have only a few concluding statements (the core points)
* Come back to the big picture and summarize the significance of your work in that context
* Open up new perspective (could be another slide)
    + describe future work
    + raise questions and potential implications
* Think carefully about the final words (which people tend to memorise)

Concluding the Presentation

* Announce the ending so that people are prepared
* Have only a few concluding statements (the core points)
* Come back to the big picture and summarize the significance of your work in that context
* Open up new perspective (could be another slide)
    + describe future work
    + raise questions and potential implications
* Think carefully about the final words (which people tend to memorise)
* <u>Seminar specifics:</u>
    + compare two papers
    + common conclusion for both papers
    + present own ideas

Questions and Answers

* Difficult questions can help improving your skills, writing and research
  + Identifies parts the audience did not understand
  + Focuses and adds an additional dimension to your analysis
* You can repeat the question using your own words
  + This gives you time to think
  + Helps in understanding the question by more people
  + Presents an opportunity for clarification
* Be concise in your answers, do not drift away
* Anticipate questions, prepare backup slides if required
* Do not say that the question is bad or it has been already addressed
* Never demean the question or questioner

Moderating the Discussion

* You will be assigned as a moderator and get a set of questions one day before the appointment
* Most probably, some questions will be already addressed; all questions cannot be addressed due to time limits
    + 2-4 questions to each paper, up to 2 questions to both papers
    + you decide which questions are the most relevant and engaging
* Prepare a set of points to discuss
    + Weaknesses / Limitations of the methods
    + Comparisons between the papers
    + Ask other participants about their ideas
    + Build bridges to other talks in the seminar
    + Points you were unclear about while reading the papers

Conclusions

* Structure your content in a way that is comfortable for you and your audience
* Filter out core aspects and build convincing story
* Use figures, videos and maths appropriately
* Rehearse and present within the time limit
* Be prepared for questions

12 Rules for a Bad Talk

* Cram as much onto each slide as you can
* Use tables with lots of data
* Make your plots really complex
* Use as many slides as possible
* Embrace obfuscation
* Over-run your time

…

Best Presentation-Ever Bingo

Didn't pre-load the presentation	Over-ran time	Used as many bullet points as humanly possible	
	Apologized for unreadable slides	Acted as if had never used PowerPoint	Embraced Obfuscation
Used incredibly complex plots		Used as many slides as humanly possible	Crammed as much as possible onto each slide
Included a video fail	Didn't check the presentation worked beforehand		Used tables with more data than any sane person could read

Source: https://medium.com/edge-of-innovation/how-to-give-the-best-scientific-presentation-ever-c87e202718cf

Materials Used

This talk is a revised version of
How to Give a Good Scientific Talk by C. Theobalt, 2017.
Some ideas are from
How to Give a Good Talk by S. Pfirman (Cornell University) and
How to give Scientific Presentations by T. Williams (Texas A&M University).

Thank You!