The oral exam will last 30 minutes and will consist of one application question followed by two theoretical questions. Please find below a non exhaustive list of possible application questions. The list of theoretical question is instead exhaustive, i.e., it contains all the topics that you should learn about the course.

Application questions

1. Summarize the building blocks of a visual odometry (VO) or visual SLAM (VSLAM) algorithm.
2. Augmented reality (AR) is a view of a physical scene augmented by computer-generated sensory inputs, such as data or graphics. Suppose you want to design an augmented reality system that super-imposes text labels to the image of real physical objects. Summarize the building blocks of an AR algorithm.
3. Suppose that your task is to reconstruct an object from different views. How do you proceed?
4. Building a panorama stitching application. Summarize the building blocks.
5. How would you design a mobile tourist app? The user points the phone in the direction of a landmark and the app displays tag with the name of it. How would you implement it?
6. Assume that we have several images downloaded from flicker showing the two towers of Grossmünster. Since such images were uploaded by different persons they will have different camera parameters (intrinsic and extrinsic), different lighting, different resolutions and so on. If you were supposed to create a 3D model of Grossmünster, what kind of approach would you use? Can you get a dense 3D model or it will be a sparse one? Please explain the pipeline that you propose for this scenario.
7. Assume that you move around a statue with a camera and take pictures in a way that the statue is not far from the camera and always completely visible in the image. If you were supposed to find out where the pictures were taken, what would you do with the images? What kind of approach would you use? Since the camera motion is around the statue, the images contain different parts of the statue. How do you deal with this problem?
8. Suppose that you have two robots exploring an environment, explain how the robots should localize themselves and each other with respect to the environment? What are the alternative solutions?

Theoretical questions

01 – Introduction

Definition of VO. VO vs VSLAM vs SFM. Assumptions. Working principle and building blocks. Dense vs semi-dense vs sparse. Bundle Adjustment vs Pose graph optimization (formulas and explanation).

02-03 – Image Formation

1. Blur circle.
2. Pin-hole approximation (proof).
3. Definition of vanishing points and lines. Prove that the parallel lines intersect at vanishing points and show how to compute it mathematically.
4. How do you build an Ames room (the floor and the walls); try to sketch a concept.

5. Relation between field of view and focal length.
6. Perspective projection equations including lens distortion and world to camera projection (derivation of perspective equations in matrix form using homogeneous coordinates). Normalized image coordinates and geometric explanation.
7. Definition of general PnP problem (what's the minimum number of points and what are the degenerate configurations). Working principle of P3P algorithm (non-linear algorithm for calibrated cameras: what are the algebraic trigonometric equations that it attempts to solve?). How do we solve PnP using a linear algorithm (derive DLT equations for 3D object or planar grids) and what is the minimum number of point correspondences it requires, why?
8. Omnidirectional cameras (only definition of central and non central cameras): what type of mirror ensure central projection? Spherical model: illustrate equivalence between perspective and omnidirectional model. What do we mean by normalized image coordinates on the unit sphere?
9. Given an image and the associated camera pose, how would you superimpose a virtual object on the image (for example, a virtual cube). Describe the steps involved.

04 – Filtering and edge detection

1. Working principle of filters: convolution vs correlation. Box filter vs Gaussian filter (what are the pros and cons of either one?). Median filter (when do we need a median filter?) Gaussian filters: why should we increase the size of the kernel if sigma is large (i.e., sigma close to the size of the kernel)? Boundary issues.
2. Edge detection: working principle with 1D signal; noise effects; differential property of convolution; how do we compute the first derivative along x and y? Laplacian of Gaussian operator: why should we use it and what effect does it have on the image? Properties of smoothing and derivative filters (what is the sum of the coefficients of a smoothing filter, and of a derivative filter?). Illustrate Canny edge detection. What is non-maxima suppression and how is it implemented?

05-06 – Point feature detection, descriptor, and matching

1. What is template matching and how is it implemented? (Mathematical expression). What are the limitations of template matching? Can I use it to recognize any car?
2. Similarity metrics: SSD, SAD, NCC, Census transform. What is the intuitive explanation behind SSD and NCC (hint: represent images as vectors)?
3. Feature extraction: what are good feature to track: definition of corners and blobs and their pros and cons.
4. Harris corner detector: intuitive illustration using Moravec definition of corner, flat region, and edge; show how to get to the second moment matrix from the definition of SSD and first order approximation (show that this is a quadratic expression and what is the intuitive interpretation of the second moment matrix using ellipse (what does the ellipse represent?). What is the M matrix for an edge, for a flat region, for an axis-aligned 90-degree corner, and for a non-axis-aligned 90-degree corner? What do the eigenvalues of M reveal? Harris detection vs Shi-Tomasi detection. Is Harris rotation, illumination and scale invariant? Why? What is the repeatability of the Harris detector after a recalling of 2?
5. Scale-invariant detection: how does automatic scale selection work? What are the good and the bad properties that a function for automatic scale selection should have or not have? How

can we implement scale invariant detection efficiently? (show that we can do this by resampling the image vs rescaling the kernel). What is the Harris Laplacian and what is its repeatability after a rescaling of 2?

6. What is a feature descriptor? (patch of intensity value vs histogram of oriented gradients). How do we match descriptors?
7. SIFT detection and descriptor: how is the keypoint detection done in SIFT and how does this differ from Harris Laplacian? How does SIFT achieve orientation invariance? What is SIFT descriptor built? What is the repeatability of the SIFT detector after a rescaling of 2? And for a 50 degree viewpoint change? Illustrate the 1^{st} to 2^{nd} closest ratio of SIFT detection: what's the intuitive reasoning behind it? Where does the 0.8 threshold come from?
8. Brief overview of FAST, SURF, BRIEF, ORB and BRISK. Pros and cons of Harris, SIFT, SURF and FAST and BRIEF, ORB and BRISK in terms of localization accuracy, relocalization, and efficiency (see recap table in the slides).
9. Describe two different ways of tracking features between frames (hint: exercises and VO project).

07 – Multiple View Geometry

1. SFM vs 3D reconstruction: definition.
2. Stereo vision: definition of disparity. Simplified case and general case; mathematical expression of depth as a function of baseline, disparity and focal length? Apply error propagation to derive expression of depth uncertainty. How can we improve the uncertainty? Large baseline vs small baseline. What's the closest depth a stereo camera can measure? Can you derive it mathematically? Stereo vision general case: show mathematically how we can compute the intersection of two lines, both linearly and non linearly. What is the geometric interpretation of the linear and non linear approaches and what error term do they minimize? (write it mathematically). Correspondence problem: epipolar geometry; definition of epipole, epipolar line, and epipolar plane. Draw epipolar lines for two converging cameras, for a forward moving camera, for a side-moving camera.
3. Stereo rectification and mathematical derivation of rectifying homographies.
4. Disparity map. How is it computed?
5. How to establish stereo correspondences with subpixel accuracy?
6. Describe one or more simple ways to reject outliers in stereo correspondences.
7. Is stereovision the only way of estimating depth information? If not, list alternative options.

08-09– Multiple view geometry 2 and 3

1. What's the minimum number of correspondences required for calibrated SFM and why? Derive the epipolar constraint. Definition of Essential matrix. The 8-point algorithm (derivation). How many rotation and translation combinations can the essential be decomposed in? Geometric interpretation of the epipolar constraint. Relation between Essential and Fundamental matrix. Normalized 8-point algorithm. Quality metrics for Fundamental matrix estimation (directional error, epipolar line, and reprojection error).
2. RANSAC. Why do we need RANSAC? What is the theoretical maximum number of combinations to explore? After how many iterations can RANSAC be stopped to guarantee a given success probability? What is the trend of RANSAC iterations vs the fraction of outliers and vs the minimum number of points to estimate the model? How do we apply RANSAC to

the 8-point algorithm? And to the DLT? How can we reduce the number of RANSAC iterations for the SFM problem (1- and 2-point RANSAC).

3. Definition of Bundle Adjustment (mathematical expression and illustration). Hierarchical SFM. Sequential SFM for monocular VO. What are keyframes? Why do we need them and how can we select them? General definition of VO and comparison with respect VSLAM and SFM. Definition of loop closure detection (why do we need loops?). List the most popular open source VO and VSLAM algorithms. Differences between feature-based and direct methods.
4. RANSAC: In practice, can you fully rely on the formula that predicts the optimal number of iterations? (hint: especially when the inliers themselves are noisy, RANSAC exercise).
5. Why is it important to normalize the point coordinates in the 8-point algorithm? Describe one or more possible ways to achieve this normalization.

10 – Multi-view Stereo

1. Working principle (aggregated photometric error). What are the differences in the behavior of the aggregated photometric error for corners, flat regions, and edges? What is the Disparity Space Image (DSI) and how is it built in practice? How do we extract the depth from the DSI? How do we enforce smoothness (regularization) and how do we incorporate depth discontinuities (mathematical expressions)? What happens if we increase lambda (the regularization term)? What if lambda is 0? And if lambda is too big? What is the optimal baseline for multi-view stereo? What are the advantages of GPUs?

11 – Tracking

1. Illustrate tracking with block matching.
2. Tracking with differential methods: describe the underlying assumptions, derive the mathematical expression, and meaning of the M matrix. When Is this matrix invertible and when not? What is the aperture problem and how can we overcome it? What is optical flow?
3. Recap of pros and cons of block-based vs differential methods for tracking. Lucas-Kanade algorithm: working principle and derivation of the underlying mathematical expression (only first two slides titled “derivation of the Lucas-Kanade algorithm”, slide pp. 55-56). What is the Hessian matrix and for which warping function does it coincide to that used for point tracking? Discussion on Lucas-Kanade: failure cases and how to overcome them. How do we get the initial guess? Illustration of coarse-to-fine Lucas-Kanade implementation. Illustrate alternative tracking using point features.
4. List one or more possible ways of discarding wrong feature tracks in practice.

12 – Place Recognition

1. Bag of words: inverted file index; what is a visual word? Why do we need hierarchical clustering? How does K-means clustering work? Explain and illustrate image retrieval with Bag of Words. Discussion on place recognition: what are the open challenges and what solutions have been proposed?

13 – Visual inertial fusion

1. Why an IMU for VO? How does a MEMS IMU work? What’s the drift of an industrial IMU?
2. What is the IMU measurement model (formula)? What causes the bias in an IMU? How do we model the bias? How do we integrate the acceleration to get the position (formula)? Loosely

coupled vs tightly coupled visual inertial fusion: definition. How can we use non-linear optimization-based approaches to solve for visual inertial fusion (mathematical expression and graphical illustration of the pose graph)?

14 – Event-based Vision

1. What's a DVS and how does it work? What are its pros and cons vs standard cameras and vs high speed cameras? Can we apply standard camera calibration techniques? How can we compute optical flow with a DVS? Intuitive explanation of why we can reconstruct the intensity. What is the generative model of a DVS? What is a DAVIS sensor? Can you write the equation of the event generation model and its proof?

15 – Visual (inertial) Odometry

1. List one or more possible ways to decide when to triangulate new landmarks based on feature tracks.
2. What are keyframes in Visual Odometry? Why are they needed? Are they strictly necessary? How to decide whether a given frame should be a keyframe?
3. Do you know any popular Visual (Inertial) Odometry algorithm? (one or more). Can you explain briefly how they work?
4. Suppose you have implemented a monocular Visual Odometry pipeline. How would change it to work as well with a stereo camera?