

Midterm Exam

CS223B

Stanford CS223B: Introduction to Computer Vision, Winter 2007

Full Name: _____

Email: _____

Welcome to the CS223B Midterm Exam!

- The exam is 5 pages long. Make sure your exam is not missing any sheets. The exam has a maximum score of 80 points. You have 60 minutes.
- The exam is open book, open notes, but closed cell phones and online devices.
- Write your answers in the space provided. If you need extra space, use the back of the preceding sheet.
- Write clearly and be concise.
- All points will be manually counted before certification.
- SCPD students: If you are taking this exam off campus, you have to fax it to (650) 725-1449 exactly 60 minutes after receipt. Alternatively, you can Email your answers to cs223b@gmail.com.

Question	Points
1 (20 max)	
2 (20 max)	
3 (20 max)	
4 (20 max)	
total	

1 Stereo Calibration

20pts

Suppose you are calibrating a stereo rig with two cameras, whose optical axes are known to be parallel (and pointing in the same direction). The intrinsics of both cameras are known. *Hint: there may be more unknowns that you might think!*

1. What is the rank of the fundamental matrix?

Answer: Rank is 2 as in the general fundamental matrix; the lack of a rotation doesn't have an effect here.

2. Suppose we have k features in the scene which can be seen by both cameras (with known correspondence). What is the number of unknown parameters in the stereo rig? How many constraints can we derive from the k features? Explain.

Answer: Unknowns overall are 4 (3 for translation, and 1 for rotation around the optical axis). Unknowns in the scene: $3k$, which correspond to the x - y - z -coordinates of each point feature. Thus, the total is $4 + 3k$. From the k features we can derive $4k$ constraints, since each feature is seen in each camera image, and each image provides 2 constraints.

3. How many features k are required (at a minimum) recovering the fundamental matrix? Carefully derive your answer (we'll be checking the math and subtract points even if it's one off).

Answer: A key insight here is that we cannot recover scale. In fact, the fundamental matrix does not even contain a scale parameter. Hence the recoverable unknowns are $3k + 3$. With that we get

$$4k \geq 3k + 3 \implies k \geq 3 \quad (1)$$

4. How would you apply the RANSAC algorithm to this problem if the feature correspondences are unknown? Provide your answer as pseudo-code.

Answer: Many answers possible, but here is one that is very simple. Correspond 3 feature pairs at random and compute from that the fundamental matrix. Then check if we can find pairs of points for which the reconstruction error is small using the fundamental matrix. If sufficiently many such points exists, reestimate the fundamental matrix with these points, and report success. Otherwise repeat. Alternatively, RANSAC could be run for a fixed number of iterations.

2 Structure From Motion

20pts

Consider a structure from motion (SFM) problem with a calibrated camera, where the camera motion is constrained to shift along the camera's optical axis (no rotation). Denote the number of images by m , and the number of features by k .

1. Suppose we have $k = 5$ point features. How many camera images do we need to recover both the motion and the structure? (Assume known correspondence.) When deriving this expression, clearly state [A] the number and type of unknowns [B] the unknowns that can be recovered, and [C] and the number of constraints we get from the images.

Answer: [A] Camera motion is in 1D. Hence we have m unknowns for the camera motion. On top of this we have $3k = 15$ unknowns for the structure, for a total of $m + 3k = m + 15$ unknowns.

[B] We cannot recover the origin of the camera motion (1 parameter), and the global scale (1 parameter). The number of recoverable parameters is hence $m + 3k - 2 = m + 13$.

[C] We get $2mk = 10m$ constraints from the images.

This gives us

$$10m \geq m + 13 \implies m \geq \lceil 13/9 \rceil = 2 \quad (2)$$

2. Suppose we know the camera motion. What conditions must be fulfilled for a point feature, so that the SFM process can recover its 3-D coordinates? Name 2 such conditions.

Answer: It must be visible in at least two images, and it may not lie on the optical axis.

3. Now let's look at correspondence. Suppose we see a point feature in one camera image, and then seek to identify the same feature in a different camera image. Where shall we look for this feature? Be specific!

Answer: Because of the limited nature of the camera motion, it suffices to search along a half-line in the image (or a half-curve if the lens distorts the image). The line must go through the image center (principal point) and the location of the feature in the previous image. Such a half-line is similar to the epipolar line of a stereo rig.

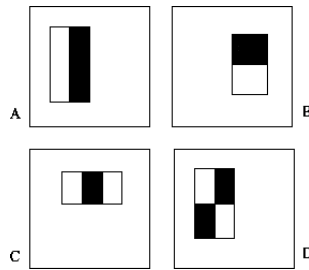
4. Now suppose one of the feature moves such that the motion is identical to the camera motion, except for the fact that the feature is located somewhere in the visual field of the camera. Where would SFM localize this feature?

Answer: SFM would believe that the feature is infinitely far away. Note that this is a special case: for the general motion case, SFM might not be able to find a good solution.

3 Viola/Jones

20pts

In class, we discussed the paper "Robust Real-Time Object Detection" by Viola and Jones. This paper defines the following Haar features:



Here white corresponds to -1 and black to $+1$.

1. Are these features computable with a linear filter? Explain why/why not.

Answer: Of course they are, they are easily written as convolutions.

2. Which of these features are separable? Explain why/why not.

Answer: All four are, since in each case the convolution in x and y can be carried out separately. The fourth is perhaps the one that is least obvious, but the decomposed filter convolves first with a kernel of the type

$$\begin{pmatrix} -1 & -1 & -1 & +1 & +1 & +1 \end{pmatrix} \quad (3)$$

and then with the transpose of this kernel.

3. Explain how these features can be computed using FFT (fast Fourier transform), and discuss some of the strength and limitations of the FFT method.

Answer: Compute the FFT of the image and the feature, then take the dot product, and compute the inverse FFT. FFTs might be more efficient than the simple-minded implementation of convolution involving large sums, yet it is less efficient than Viola/Jones computation. FFTs are also cyclic, hence may give erroneous result at the image border.

4. Viola and Jones do *not* use the FFTs. Why?

Answer: Because of efficiency. The Viola Jones algorithm computes all necessary statistics in time linear in the number of pixels, whereas FFTs require long-linear time. However, their trick does not generalize to most other linear filters.

4 True or false

20pts

Correct answer is 2 point per question; a false answer results in minus 1 point.

TRUE FALSE In the SFM problem, assume the camera motion is constrained to shift in the x - y -direction (no z -shift and no rotation). This implies that the orthographic assumption in Tomasi/Kanade's SFM technique is *exact*.

Answer: FALSE, the orthographic assumption is the limit case where the scene is infinitely far away from the cameras.

TRUE FALSE A perspective projection is also affine.

Answer: FALSE, the division in a perspective projection cannot be expressed by an affine projection matrix.

TRUE FALSE A weak perspective projection is an orthographic projection with a scale factor.

Answer: TRUE.

TRUE FALSE SIFT features are invariant to occlusion.

Answer: FALSE, occlusion implies that a feature is not even visible.

TRUE FALSE The Viola/Jones algorithm discussed in class is particularly suitable for object recognition from a single training image.

Answer: FALSE, the opposite is true: it typically requires thousands of training images. SIFT features are much better suited for that.

TRUE FALSE An image pyramid provides (approximate) invariance to scale.

Answer: TRUE, this is the exact purpose of a pyramid.

TRUE FALSE Orthographic projections preserve straight lines and right angles.

Answer: FALSE, just imagine an orthographic projection of a cube—it's easy to tilt any of the side surfaces in the projection.

TRUE FALSE Canny edges are the preferred features for optical flow because of the aperture effect.

Answer: FALSE, edges tend to yield aperture effects, corners are therefore much better.

TRUE FALSE Correspondence: When calibrating a stereo rig, it suffices to search corresponding features on epipolar lines.

Answer: FALSE, during calibration we cannot tell where the epipolar line is.

TRUE FALSE Vignetting predominately occurs at the border of an image, and it is usually corrected in the calibration process by adjusting the calibration parameters k_i (for $i = 1, 2, \dots$).

Answer: FALSE, the parameters k_i address lense distortion, not vi-