Problem set 1, 6476 Computer Vision, spring 2021
Name: Haoran Wang

Unless otherwise stated, there is a single best answer for each multiple choice question.

[p1] *Human Vision*. The human eye is analogous to cameras in many ways. Which of the following statements about the human are true? *Choose **all** that are correct*
[✓ (a)] Rods have little to do with color perception in bright scenes.
(b) Cones are active only during low light situations.
(c) Rods are highest density in the Fovea (center of our visual field).
(d) Colors are perceived in bright scenes based on the relative color sensitivities of Rods versus Cones.
[✓ (e)] People typically have 3 types of cones sensitive to 3 different ranges of wavelengths (color blind persons can be an exception).

[p2] *Cameras*. A typical camera sensor (as you'd find in a cell phone) measures which attributes of the photons coming from a scene? *Choose **all** that are correct*
[✓ (a)] The number of "bounces" a photon took off various scene surfaces
[✓ (b)] The position that a photon hit the sensor (equivalent to the angle that a photon passed through the "pinhole")
[✓ (c)] The amount of time that a photon existed
(d) The rate of photons hitting the sensor (flux)
(e) The distance a photon traveled
(f) The wavelength of a photon hitting the sensor

[p3] *Image Formation*. Which of the following properties of 3d lines are preserved during image formation under perspective projection with the pinhole camera model?
(a) relative lengths of lines
(b) angles between lines
[✓ (c)] parallelism between lines
(d) straightness of lines

Image formation with a pinhole camera model is described mathematically as
$p_{image} = K\,[\,R \mid T]\,p_{world}$ where p_{image} is an image coordinate and p_{world} is a world coordinate.

[p4] *Image Formation*. What is the matrix K?
[✓ (a)] This "intrinsic" matrix converts 3d camera coordinates to image coordinates.
(b) This matrix integrates the exposure over time accounting for world and camera motion.
(c) This matrix accounts for distortions caused by the camera lens.
(d) This "extrinsic" matrix rotates and translates the world coordinate system to align with the

camera coordinate system.

[p5] *Image Formation*. What is the matrix [R | T]?
(a) This “intrinsic” matrix converts 3d camera coordinates to image coordinates.
(b) This matrix integrates the exposure over time accounting for world and camera motion.
(c) This matrix accounts for distortions caused by the camera lens.
(d) This “extrinsic” matrix rotates and translates the 3D world coordinate system to align with the 3D camera coordinate system.

[p6] *Image Formation*. How many degrees of freedom does the matrix [R | T] have?

6 degree of freedom

[p7] *Image Formation*. In lecture we mentioned that K can be modeled with different numbers of free parameters -- in fact the book mentions 8, 7, 5, 3, and 1 parameter versions. The 1 parameter version, which models only the *most important* parameter, is of this form:

?	0	0
0	?	0
0	0	1

What is this “?” parameter and what is its physical meaning?

f_x and f_y, the focal length for the sensor x and y dimension.

[p8] *Frequency Domain*. Convolution in the spatial domain is equivalent to element-wise multiplication in the Fourier domain. Why might this be useful?
(a) Since element-wise multiplication is very fast, it may be faster to convert the image and filter to the frequency domain, multiply them, and convert back to the image domain rather than performing the convolution in the spatial domain as we did in project 1.
(b) Many common filters only exist in the Fourier domain and have no spatial domain representation, so we have to work in the Fourier domain for those filters.
(c) When filtering in the spatial domain you can have boundary artifacts where the filter doesn’t fit into the image near the edges of the image. In the Fourier domain there are no boundary artifacts because it magically hallucinates content outside the image boundaries.
(d) Working in the Fourier domain is more numerically accurate.

[p9] *Filtering*. If we convolve an image A with filter B to produce result C (A * B = C), can we invert the convolution (i.e. recover A from B and C)?
(a) No, never. Convolution always removes information (e.g. blurs away high frequencies) and there's no way to recover that information even approximately.
(b) Kind of. Filters typically attenuate but do not completely remove particular frequencies from A (equivalently, the fourier transform of a filter is mostly non-zero amplitudes). These attenuated frequencies can be approximately recovered by division of C by B in the Fourier domain. As particular frequencies of B approach zero amplitude this approximation becomes worse.
(c) Yes, sometimes. For some trivial cases (e.g. B is identity or a shift filter) then A is easy to recover. For all other filters you cannot even approximate A given B and C.
(d) Yes, always. Convolution is multiplication in the Fourier domain, and multiplication is inverted by division. Thus any convolution is invertible.

[p10] *Features*. Why do the majority of local features, such as the SIFT feature you implemented, characterize the spatial *gradients* of intensities in an image patch rather than the raw pixel intensities?
(a) gradients are invariant to image rotation.
(b) gradients are invariant to image scaling.
(c) gradients are invariant to constant brightness shifts.
(d) gradients don’t have noise but pixel intensities do.

[p11] *Model Fitting and Outlier Rejection*. We want to use a Hough transform to find arbitrary triangles in an image. Triangles can be parametrized by the location of the 3 corners for a total of 6 degrees of freedom. How much memory is needed to run this Hough transform?
(a) 6 ints to store the image coordinates of the best model currently discovered.
(b) K ^ 6 ints, where K is the number of discrete bins for each spatial dimension.
(c) 6 ^ K ints, where K is the number of discrete bins for each spatial dimension.
(d) 6 * K ints, where K is the number of discrete bins for each spatial dimension.

[p12] *Model Fitting and Outlier Rejection*. RANSAC is an algorithm to robustly find the best parameters for a model. To find the best line parameters given a large set of points, briefly outline the steps of the RANSAC algorithm

To find the consensus set of inliers, repeat the below two steps until there is enough inliers:

1. In the first step, a sample subset containing minimal data items is randomly selected from the input dataset. A fitting model and the corresponding model parameters are computed using only the elements of this sample subset. The

cardinality of the sample subset is the smallest sufficient to determine the model parameters.

2. In the second step, the algorithm checks which elements of the entire dataset are consistent with the model instantiated by the estimated model parameters obtained from the first step. A data element will be considered as an outlier if it does not fit the fitting model instantiated by the set of estimated model parameters within some error threshold that defines the maximum deviation attributable to the effect of noise.

To find the best line parameters, RANSAC achieves its goal by repeating the following steps for a fixed iteration:

1. Select a random subset of the original data. Call this subset the hypothetical inliers.
2. A model is fitted to the set of hypothetical inliers.
3. All other data are then tested against the fitted model. Those points that fit the estimated model well, according to some model-specific loss function, are considered as part of the consensus set.
4. The estimated model is reasonably good if sufficiently many points have been classified as part of the consensus set.
5. Afterwards, the model may be improved by re-estimating it using all members of the consensus set.

This procedure is repeated a fixed number of times, each time producing either a bad line parameter which is rejected because too few points are part of the consensus set, or a good line parameter together with a corresponding consensus set size. In the latter case, we update the best line with the corresponding consensus set if its consensus set is larger than the previously saved best line parameters.

[p13] *Model Fitting and Outlier Rejection*. Which of the following is a potential advantage of RANSAC over the Hough transform?
(a) With RANSAC it is easier to simultaneously fit multiple independent models (e.g. 3 different affine transformations) which fit different parts of the data well.
(b) With RANSAC there is no need to discretize the model's parameter space and store an accumulator array over this parameter space.
(c) The stopping conditions of RANSAC are more clearly defined than the Hough transform.

[p14] *Stereo*. Which of the following statements correctly describes the relationship between disparity and depth?
(a) Depth and disparity are independent.
(b) Depth is proportional to disparity.
(c) Depth is proportional to the square root of disparity.
(d) Depth is inversely proportional to disparity.

[p15] *Stereo.* Which of the following is a property of the Fundamental Matrix F? x and x' are homogeneous pixel coordinates, from cameras at location o and o' respectively, represented as

3 x 1 vectors.
(a) $x^T * F * x' = 1$ only when points x and x' correspond to the same 3d point.
(b) $x^T * F * x' = 0$ implies that points x and x' could be projections of the same 3d point.
(c) $x^T * F = x'$ for points x and x' at the same depth.
(d) $F = x' * x^T$ if x' and x are the projections of the other cameras (o and o').

[p16] *Tracking and Optical flow*. In the context of tracking and optical flow, explain the "aperture problem" and how it is mitigated?
(a) The "aperture problem" is an illusion specific to human perception of motion. The illusion does not need to be mitigated by tracking algorithms.
(b) Tracking and optical flow is ambiguous with simple intensity features. Using color and texture descriptors mitigates the aperture problem.
(c) Without knowing the parameters of the camera aperture, tracking and optical flow are fundamentally ambiguous. Camera calibration eliminates the ambiguity.
(d) In small image regions, true motions can be ambiguous if the image content is flat or one-dimensional. Reasoning about larger windows or only tracking corners reduces the ambiguity.

[p17] *Tracking and Optical flow*. Which of the following is **NOT** an assumption of the classical Lucas-Kanade optical flow discussed in class:
(a) Parallel motion: points only move parallel to the imaging plane in 3d
(b) Brightness constancy: projection of the same point looks the same in every frame
(c) Spatial coherence: points move like their neighbors
(d) Small motion: points do not move very far