# 2 - Image Formation

**How does the size of the pinhole affect the sharpness of the image?**

A bigger pinhole delivers a brighter image, but the sharpness will be low.

A smaller pinhole delivers a dimmer and sharper image (wenn zu klein, dann nimmt das Licht die Welleneigenschaft an - wird unschärfer)

The image cannot be infinitely sharp, due to the light passing through as a pencil of rays.

Correction proposal: However, decreasing the pinhole size can lead to diffraction (Beugung) which also leads to a blurry image. Therefore we need a lense to reliably create a sharp image.

**What is a vanishing point when projecting from 3D to 2D?**

Parallel lines in the real world (3D) converge at a vanishing point in the image (2D) (if they are not parallel to the imaging plane)

**Why is perspective projection a non-linear transformation?**

(x, y, z) → (f’ x/z , f’ y/z) is a non-linear transformation because the division by z is non-linear.

**How can one convert from homogenous vector to non-homogenous vector?**

By multiplying the homogenous vector by the perspective projection [ **I** | 0 ] (maybe only for matrix?)

→ for vector: divide each entry by the third coordinate (2D) (or fourth (3D), depending on the dimension, basically divided by last component in vector, which is the weight)

**What are extrinsic and intrinsic camera transformations?**

Objects live in “world coordinates”.

Camera has a “camera coordinate” system.

Image is defined in “image coordinates”.

* Extrinsic camera transformation takes world coordinates into camera coordinates. The needed parameters are rotation and translation of the camera. Here we have six degrees of freedom.
* Intrinsic camera transformation describes the image formation process. A needed parameter is the camera calibration matrix **K**, which captures the so-called intrinsics, which are the principal point coordinates (p\_x, p\_y) which is basically the offset of the principal point from the boundary of the image/sensor, the focal length (f), and the pixel magnification factors (m\_x, m\_y), that translate units of measurements (m, mm, …) into unitless pixels. An additional and optional parameter is the skew to model non-rectangular pixels (not used in the lecture 2)

**What are the features/parts that govern a pinhole?**

* Lightproof barrier with a pinhole.
* Film/sensor where the image plane lies. The image there is upside down.
* The virtual image plane lies in front of the pinhole. The image there would be right sid

**What is radial distortion?**

This type of distortion usually occurs due to unequal bending of light. The rays bend more near the edges of the lens than the rays near the center of the lens. Due to radial distortion, straight lines in the real world appear to be curved in the image. The light ray gets displaced radially inward or outward from its ideal location before hitting the image sensor. Pinhole cameras never have radial die up.

stortion, it can only happen with lenses.

**How would you formulate the camera parameters estimation with homogeneous least squares?**

Find the right nullspace of A using singular value decomposition (SVD)

* Decompose equation system matrix **A = USVT**
  + **ATA = (USVT)T (USVT) = VSTUTUSVT = VS2VT**
  + The eigenvectors of **ATA** are the right singular vectors of A (i.e. columns of **V**)
  + The eigenvalues of **ATA** are the square of the singular values of **A**
  + To find the eigenvector of **ATA** with the smallest eigenvalue, we compute the last right-singular vector of **A**
* Assume that singular values are stored, i.e. **S** = diag(s1, …, s12), si+1 ≤ si
* Take last right singular vector **p** = **v**12

# 3 - Cameras

**Why do we need multiple planes for camera calibration?**

If all points lie on a plane, we will get degenerate solutions. Using multiple planes avoids that.

**Why does the image become blurry when the size of the aperture is very small?**

If the pinhole is too small, the light spreads out due to its wave characteristics. This is called the diffraction effect (“Beugungsunschärfe”).

**Which points are in focus when using a camera with lens?**

All parallel rays converge to one point on a plane located at the focal length f. It's called focal point. Other points (that are close to or farther away) project to a “circle of confusion” in the image, because the bundle of light rays will not converge to a single point.

Based on the thin lens formula, any point satisfying

is in focus, where D is the distance between the lens and the object, D’ is the distance between the lens and the image sensor and f is the focal length of the lens.

**What is vignetting?**

It's the effect when the image gets darker towards the corners. Lenses have a black tube around the assembly of optical elements. Parts of the light that goes through the first lens element near the outer circles hits the black plastic at some point and gets lost

The property of real-world lenses is vignetting, which is the tendency for the brightness

of the image to fall off towards the edge of the image

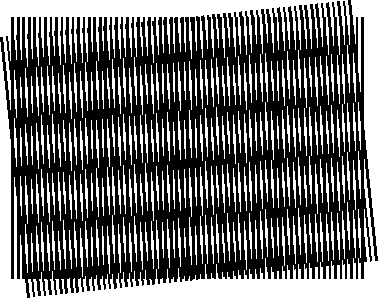
**What is the role of an image sensor?**

To convert photons into a digital signal.

**What is moiré effect?**

Parts of the image get e seen when looking through ordinary window screens at another screen or background. It can also be generated by a photographic or eleccolored where there isn't any color due to the (erroneous) interpolation of the RGB values (fine black and white detail in image misinterpreted as color information). This effect can get avoided by using an artificial low-pass filter (artificial blur)

Moiré effect is a visual perception that occurs when viewing a set of lines or dots that is superimposed on another set of lines or dots, where the sets differ in relative size, angle, or spacing. The moiré effect can abtronic reproduction, either deliberately or accidentally.



# 4 - Image Filtering, Edges & Image Pyramids

**Is convolution a linear operator?**

Yes, convolutions are a special case of a matrix-vector-product.

**What is the shift invariance property of convolution operation?**

First shifting the image by a certain spatial offset and then filtering it delivers the same result as first filtering and then shifting the output.输入移动，输出也移动。所以不便

**What is the difference between convolution and correlation?**

Convolution requires mirroring the kernel in both x- and y-direction around the center (rotate kernel 180 degrees), correlation does not.镜像滤镜

Convolution operator: ∗

Correlation operator: ⊗（相关，相似度）

**Why do the coefficients sum to one in the average filter?**

To conserve the brightness level of the input image.

<1太暗，>1太亮

**How is Gaussian filter separable?**

P15 行列分别卷积

**Why do Gaussian smoothing (besides downsampling) when creating a pyramid?**

Downsampling: taking every other pixel (literally, there is no averaging).

Downsampling without applying Gaussian smoothing conserves all parts of the spatial frequency spectrum. The lower resolution images cannot correctly represent the high frequency signals, instead these high frequencies get misinterpreted using other (and erroneous) spatial frequencies (aliasing).

Applying Gaussian smoothing removes the parts of the spatial frequency spectrum that can’t be represented by a lower resolution image, and thus avoiding the aliasing effect (i.e. introducing artificial information).

高频不能被低分辨率显示，所以用高斯平滑下采样。

**What are the goals of edge detection?**

* good detection: filter responds to edge, not to noise
* good localization: detected edge near true edge
* single response: once per edge

**How to compute edge strength?**

Gradient magnitude, which is the norm of the gradient vector.

# 

# 5 - View-Based Recognition & PCA

**What is the detection/localization tradeoff in edge detection?**

The goals *good detection* and *good localization* contradict each other. Smoothing the image to remove the noise results in a blurrier image. This reduces the localization ability of the representation because of the blurry edges. However, this ensures that noise is not falsely detected as an edge

边缘检测中的检测/定位权衡是什么？

良好检测和良好定位的目标相互矛盾。 平滑图像以去除噪声会导致图像更模糊。 由于边缘模糊，这降低了表示的定位能力。 但是，这确保了噪声不会被错误地检测为边缘

**Under what assumptions is the Canny edge detector optimal?**

* assume we can only do linear filtering
* additive independent and identically distributed (i.i.d.) Gaussian noise

**What is normalized correlation in the context of template matching?**

The dot product of the template with the image patch divided by the norm of a and b computes the cosine of the angle between the two vectors.

It is the distance in the angular space. This representation is contrast invariant.

**What is meant by principal directions in PCA?**

The direction of the maximum variation of the data.

**？How do we project the image onto basis vectors when performing PCA?**

We first subtract the mean from the image. Then we multiply the result with the transposed of a given basis vector to get the length of the projection onto the basis vector.

**What is the relation between eigenvalues and direction of maximum variance?**

Largest eigenvalue is maximum variance. The corresponding eigenvector is the direction of the maximum variance.

**What is the advantage of using SVD for PCA?**

Using SVD eliminates the need to explicitly build and store the covariance matrix. SVD can be appied to any matrix, numerically stable, computationally efficient.

Svd also allows for an economic computation of the first biggest eigenvalues and eigenvectors without having to calculate the whole decomposition

# 6 - Interest Points

**What do the basis vectors of PCA look like for natural images?**

(Besides the basis vectors, we save the mean, which is a gray image). The first two basis vectors are the derivatives of the gaussian filter (if x or y direction comes first is random and depends on the images and noise). The next basis vectors look like the second derivatives of the gaussian, then third derivatives etc.

**How can we decide on the number of eigenvalues to choose when approximating data vectors with PCA?**

We can either choose the minimal number that works for the use case, i.e. we choose a given number, try if it works well for us and if not, increase the number by hand.

Or we choose as many eigenvectors so that the total variance they cover exceeds a certain percentage, e.g. the total variance covered should be 0.9 of the total variance. The total variance covered can be calculated by the sum of the eigenvalues of the eigenvectors used.

P7 1.2+公式

**Why are the limitations of PCA for appearance-based instance recognition?**

Challenges of appearance-based instance recognition:

* Viewpoint changes (translation, image-plane rotation, scale change, out-of-plane rotation)
* Illumination
* Clutter
* Occlusion
* Noise

The assumption that the subspace is linear is not reasonable. An object cannot be represented in a linear subspace.

PCA will only work if the images are well aligned in a sense that the orientation is canonical.

**What is the local measure of interest point uniqueness?**

Spawn a window at a certain location in the image and shift the window around. If the content in the window changes, the content is an interesting point in the image. If the content doesn't change, the content is not that interesting.

**shifting the window in any direction causes a big change**

**What do the smallest and largest eigenvalues of a structure tensor represent?**

Eigenvectors (x): Encode in which direction the content of the window changes the most & the least.

Eigenvalues (λ): Encode how much the content of the window changes.

* x+: direction of the largest increase in E
* λ+: amount of increase in x+
* x-: direction of the smallest increase in E
* λ-: amount of increase in x-

**What are the advantages of the Harris detector?**

No need to compute a square root (which is a computationally expensive operation). It is a good approximation and faster to compute than the smaller of the two Eigenvalues λ-.

**Harris detector is rotation invariant. computationally faster.**

**How are the interest points obtained from the Harris detector different from Hessian interest points?**

Harris detector is a corner detector, Hessian detector is a blob detector. Interest points obtained from Hessian detector do not lie exactly on the corner, but are slightly displaced.

**How can one compare the interest points of image patches at different scale?**

Interest point in original image detected. Extract a patch from both images around the interest point/ where the interest point is supposed to be.

Inefficient approach: Iteratively increase patch size in the scaled image while matching similarity until correct patch size is found. This approach is computationally inefficient.

Better approach: Apply Gaussian smoothing on the original image patch, which simulates zooming in. The advantage of smoothing instead of zooming in, is that the interest points remain in the same location.

create a scale space and do non-maximum suppression over the scales by finding the scale that maximizes the laplacian for a given keypoint -> canonical scale

detecting interest points->extracting patches->computing feature descriptors->comparing feature descriptors

# 7 - Matching Interest Points & Single-View Geometry

**Which interest points detection algorithm have a better repeatability rate?**

Standard Harris is bad. Adapted (Harris-Laplace) is better. More modern ones like SIFT & SURF are slightly better than Harris-Laplace, but not by much.

**repeatability rate 公式**

**How do you compute the canonical orientation of image patches ?**

TODO

**Why does the SIFT descriptor divide the 16x16 window into a grid of 4x4 cells?**

To achieve rotation and illumination dependencies from the distinctive and invariant interest point, and to further reduce the dimensionality of the descriptor

**Why does log-polar coordinate system work better than the Euclidean coordinate system for shape context descriptors?**

If we have different instantiations of the same shape, they are sometimes related by radially distorted versions. Using a log-polar coordinate system & histogram allows representing the shape as viewed locally from an interest point in a way that is invariant to certain shape transformation, so that it is sensitive to nearby changes in geometry and less sensitive to changes far away from the center, so it can deal with certain shape distortions.

**\* nearby geometry change sensitive**

**\* better for shape distortion**

**How do we evaluate the performance of different feature matchers?**

With the ROC curve (“Receiver Operator Characteristic”)

It is generated by counting the amount of correct/ incorrect matches for different thresholds. The goal is to maximize the area under the curve.

ROC plots True positives on the y axis and False positives the x axis

**How can we obtain the homography matrix H from the projection matrix P?**

x = P X

x = H x\_pi

remove the third column of P since we assume Z=0 for the plane?

TODO

**What are some applications of a homography?**

* Panorama stitching
* Image registration
* stereo rectification

**Why are 4 point correspondences required for estimation of a homography?**

**x’ ∝ H x**

Because we are dealing with homogeneous coordinates, we don’t know the actual depth of the light ray, we only know that the direction is the same. That means that we only know that **x’** is proportional to **H x**. We don’t know whether they have the same length, we only know that it has the same direction.

If we know that a homogeneous 2D vector is proportional to another homogeneous 2D vector, it gives us two constraints. Because if the direction is the same in homogeneous 2D space, this ties two d.o.f. from the total eight d.o.f. of the Homography. Thus, we need at least four point correspondences to actually estimate the Homography. The more point correspondences we have, the better we can determine the Matrix.

-> short version: H has 8 degrees of freedom and we get 2 equations per point correspondence (x and y component). That means we need at least 4 correspondences

**?How do we scale the equation system matrix so that all values lie between -1 and**

**+1?**

1. Calculate s as the distance of the point that is the furthest away from the origin divided by 2 (for one image)
2. Compute t as the average location of all points (for one image)
3. Calculate T matrix and T’ matrix
4. Scale the points x and x’ using T: u = Tx and u’ = T’x’
5. Use u and u’ instead of x and x’ in the equation system matrix

(Why do we do this?)

In a 1000x1000 pixel image, the x and y pixel values lie in the range [0, 999]. Some entries of the matrix can be as big as a million or as small as one. These large discrepancies between values can lead to numerical errors when using finite precision floating point arithmetic. Conditioning the matrix so that all the coefficients are in the same order of magnitude circumvents this problems. So: Numerical Stability

# 8 - Single & Two-View Geometry

**How do we get the Homography matrix H from the Projection matrix P ?**

We remove the third column of P. This follows by definition that the Z value in our world coordinate system is set to 0 (this makes calculation easier). Therefore, the third column in our projection matrix is then always multiplied with 0 and can be left out. The new matrix (without the third column) is then our homography matrix H.

same as above

**What condition do you use for estimating the Homography matrix H ?**

Scale and shift points to be in [-1, 1]. Apply coordinate normalization transformation → estimate the Homography H → undo normalization to get the desired coordinate system.

**How many number of trials are required for estimating matrix H with 30% inliers and z=99%?**

w = 0.3, z = 0.99, d = 4

**Show that cross product is equivalent to the matrix vector product (slide 47) ?**

TODO

**What constraints are used for deriving the Triangulation method ?**

Assumption:

* Two different views of the same scene
  + Either scene is static and camera was moved
  + Or scene is dynamic and two pictures were taken at the exact same time

TODO

**What is the disadvantage of Triangulation compared to Epipolar Geometry ?**

start of recording l9: triangulation requires knowledge of camera pose and calibration, ie.e. intrinsics and extrinsics.

Triangulation via the non-linear approach is the most accurate method, but more complex than the geometric midpoint approach or the linear approach. For two cameras, we have to find roots of a 6th degree polynomial. For more than two cameras, we have to initialize with a linear estimate and then optimize with iterative methods.

**Define epipoles and epipolar lines.**

Epipole: Image location of the optical center of the other camera (Can intersect the virtual image plane outside of the visible area). Example: Epipole e2 is the location of the camera center of Camera 1 in the coordinate system of Camera 2.

Baseline: Line connecting both optical centers. Both epipoles lie on the baseline.

Epipolar Plane: Plane through both camera centers and world point. The baseline falls into that plane.

Epipolar Lines: Intersection of the epipolar plane with the image plane. The line constrains the location where a particular feature from one view can be found in the other (It tells us the potential location of the corresponding points). There is one epipolar line per world point. All epipolar lines intersect at the epipoles. The epipolar lines are in general not parallel. If the image plane of both cameras are parallel, then the epipolar lines are parallel as well, the epipoles would be at infinity.

**How do Epipolar lines help with feature matching ?**

In order to match interest points from the first image to the second image, we have to search for the interest point only on the epipolar line in the second image. The point cannot lie anywhere else in the second image but on the epipolar line.

**What is the epipolar constraint ?**

The points p1 in the first image plane and p2 in the second image plane are related by that constraint. This constrain constrains one degree of freedom. Knowledge of p1 with camera translation **t** and camera rotation **R** gives us the epipolar line of p2.

Two views of the same 3D point must satisfy **p1**T [**t** x (**R p2**)] = 0

*Popular exam question: How many d.o.f. does this constraint constrain?*

It constraints 1 degree of freedom, and we still have one degree of freedom left. Mathematically, we have an equation that yields a scalar value on the right-hand-side, so this has one degree of freedom. Intuitively, we map a point in one image to a line in the second image, but we still have one degree of freedom, where on the line the point can be.

**Prove that the rank of the essential matrix is 2.**

* Essential matrix **E** is a 3x3 matrix, so it has 3 eigenvalues.
* **E** is singular, it has rank 2 (third eigenvalue is zero).
* The two remaining eigenvalues are equal.
* 5 d.o.f. (translation + rotation have 6, but scale is arbitrary).

# 9 - Two-View Geometry & Vision

**What are epipolar lines?**

Intersection of the epipolar plane with the image plane

**How is epipolar constraint expressed in terms of the essential matrix?**

Epipolar constraint: **p1**T [**t** x (**Rp2**)] = 0

Express cross-product as matrix: t x v = [**t**]x **v**

Essential Matrix: **E** = [**t**]x **R**

Epipolar constraint with essential matrix: 0 = **p1**T **E p2**

**What is the relation between epipoles and the essential matrix?**

Epipoles are the (left/ right) null-space of the essential matrix:

**e1**T **E** = **E**T **e1** = 0

**E e2** = 0

—

(p.25)

**x**1 = **K**1 **p**1 **x**1 = **K**1 **p**1

**p**1 = **K**1-1 **x**1 **p**1 = **K**1-1 **x**1

with **p**1, **p**2 calibrated points, **x**1, **x**2 uncalibrated points

—

**What is the relation between fundamental matrix and essential matrix?**

**F** = **K**1-T **E** **K**2-1

Combination of the Essential matrix **E** and the unknown calibration matrix **K**.

The epipolar constraint still holds.

Fundamental matrix is singular, has rank 2. It no longer holds that the two non-zero eigenvalues are the same, that holds only for calibrated cameras.

**How can numerical instability be resolved in the eight-point algorithm?**

Products in the matrix **F** can have coefficients that have a wide range in terms of magnitude. We can condition the matrix to get numerical stability by scaling and shifting points to be in range [-1, 1].

Important: undo the normalization at the end.

**How to obtain the fundamental matrix in absence of point correspondences between two images?**

knowing the extrinsics and intrinsics of both views, as well as the camera positions, the fundamental matrix can be computed as follows:

**What is the relation between binocular disparity and depth?**

From known geometry of the cameras and estimated disparity, recover depth in the scene.

* d: disparity
* f: focal length
* B: baseline
* Z: depth

**What is the importance of stereo rectification?**

If the two images are not in the required setup, use stereo rectification to rewarp them such that they are.

It’s almost impossible that both cameras point in the exact same direction (only translation in x direction). The camera directions will be slightly different, and thus the epipolar lines will not be parallel but slightly tilted. To make the epipolar lines parallel, we can apply the rectification setup, where we map both images to a common plane that is parallel to the baseline vector.

The downside of this is that the output images are no longer rectangular, but this is not a problem.

all epipolar lines are parallel to the horizontal axis of the image

**What is the drawback of window-based matching for depth estimation?**

No window size fits every condition. Adaptive window size delivers better results.

# 10 - Dense Geometry & Dense Motion Estimation

**How can the matching criterion in stereo be made robust?**

Non-local constraints:

* Uniqueness
  + For any point in one image, there should be at most one matching point in the other image
* Ordering
  + Corresponding points should be in the same order in both views
* Smoothness
  + We expect disparity values to change slowly (for the most part)

**What is the shortest path algorithm for scalene stereo?**

Put a cost on the edges of the graph. Cost at left and right node is higher than the diagonal node, so diagonal direction is preferred.

**Can we obtain spatially coherent scanlines using the dynamic programming approach?**

Yes, we get spatially coherent scanlines. Within a scanline we get a lot of geometric details. The geometry is a lot more accurate than with the window based matching approach.

We estimate the disparity in each scanline independently of all the other scanlines. That leads to the streaking artifacts. We cannot extend dynamic programming to 2D, so we can’t use dynamic programming to find specially coherent disparities/ correspondences on a 2D grid.

**How can the motion field be different from optical flow?**

Motion field: 2D motion field, representing the projection of the 3D motion of points in the scene onto the image plane. (Can be the result of camera motion, object motion, or both)

Optical flow: 2D velocity field describing the apparent motion in the images.

Motion field describes real motion, either of the object in the scene or the camera. Optical flow can occur even if both the objects in the scene and the camera are static, e.g. when a moving light source illuminates the scene. The specular highlight moves over the surface causing optical flow, but no motion field.

see also barber pole: motion field: sideways (true motion), optical flow: upwards

**What assumptions are made when computing optical flow?**

* Brightness constancy assumption:
  + Image measurements (e.g. brightness) in a small region remain the same, although their location may change
* Spacial coherence assumption:
  + Neighboring points in the scene typically belong to the same surface and hence typically have similar 3D motions
  + Since they also project to nearby points in the image, we expect spatial coherence in the image flow.

**What are the limitations of simple flow estimation algorithm?**

* Very inefficient
* The motions are discrete, which is usually not true in practice

**What makes the optical flow constraint equation a convex problem?**

It holds under the following assumptions:

* Assume small motion
* Assume brightness varies smoothly

approximated SSD using first order taylor series approximation of the intensity at time t+1 and location (x+dx, y+dy)

**How can multiple constraints be combined to estimate the velocity?**

We cannot detect neither the optical flow, nor the motion field. If a single constraint line from a single pixel is not enough, then we need a constraint line from a different pixel. Then we can determine the intersection. With it, we can not only estimate the normal component, but the whole motion b.c. we localized the motion in 2D space.

**What makes the structure tensor ill-conditioned for motion estimation in homogenous area?**

* Textured area:
  + A lot of texture, structure tensor can be inverted → We can get motion
* Single edge:
  + Structure tensor is not well-behaved. In an interest point, both eigenvalues are large. At an edge, one eigenvalue is nearly zero. Thus, the structure tensor is not well invertible
* Homogenous area:
  + No texture, structure tensor is nearly zero, so both eigenvalues are nearly zero. Therefore, the structure tensor is poorly conditioned.
* Surface boundary:
  + The surface of a closer object moves much faster through the window than the surface of a distant object. Window based approach assumes that the motion in the entire window is the same. Window based optical flow technique is not applicable at surface boundaries.

# 11 - Object Recognition

*Some of these questions are from the 11th lecture, but they relate to image motion.*

**What are the advantages of inverse warping over forward warping?**

Forward warping: Sending the pixel to the new location in the second image. When the pixel lands between pixels, we attribute the source pixel’s value to all four target pixels with the respective weight. Therefore, for every target pixel, we would need to save the color value of the source pixels and their weight in the target pixel.

When zooming in, if you think of the pixels as infinitesimal, you get a gap between the pixels. This area is then black, b.c. there is no source pixel that contributed to that target pixel. When treating the pixel as a square, you can warp it to a quadrilateral, and avoid the black areas. This approach is computationally much more expensive and is therefore not done.

Inverse warping: Iterate over all pixels in the second image and look up where they should come from in the first image. In the first image, then do an image lookup. When the pixel lands between two pixels, it is less of a problem than in forward warping, b.c. you can just interpolate the value, which is typically easy to do. We can use nearest neighbor, bilinear or bicubic interpolation.

**forward is computationally more expensive**

**inverse can interpolate the value , which is easy to do.**

**What is the limitation of nearest neighbor interpolation?**

Nearest neighbor interpolation leads to jagged artifacts, especially for rotation and projective distortion.

**What is the limitation of LK method?**

The LK method only works for small motions. (But this can be compensated by using iterative estimation and coarse-to-fine estimation.)

LK assumes that the flow is constant in a small region (spatial smoothness), so at the motion boundaries, discontinuities in the flow are smoothed over, although they exist (window too big). In areas with little texture, there is not enough image information to estimate the image flow (window too small).

LK is a local optical flow method.

**What are the disadvantages of view-based approaches in recognition?**

* They are severely challenged by these common variations. To make them work, we would need an unmanageable amount of examples (training data)
  + Viewpoint changes (translation, image-plane rotation, scale changes, out-of-plane rotation)
  + Illumination
  + Clutter
  + Occlusion
  + Noise
* They do not generalize well. Almost any variation that hasn’t been captured in the training data will not be handled gracefully.
* Training data is expensive: Humans have to gather and label it.

**What is the color constancy problem?**

The pixel colors change with the illumination. We need to get the white-balance in the camera correct if we want to have the accurate colors in different lighting settings. Also the overall brightness could be different.

**What is the receptive field?**

/\*Any local descriptor (e.g. filter, filter combination) can be used to build a histogram. Examples:

* Gradient magnitude
* Gradient direction
* Laplacian

\*/

A receptive field is the region in the input which affects the output after applying e.g. a filter.

Each feature response (feature vector) is therefore not a value of a single pixel, but takes into account the neighborhood of the pixel.

Analog to biology: In biology the receptive field is the area of the retina (here image) that influences the firing rate of a neuron (here node activation)

# 12 - Bag of Words for Object Classification

**When do you build the codeword representation (Train/Test)?**

The codeword dictionary gets built during the learning phase (train).

It gets used during the recognition phase (test).

**What invariances does using visual BoW representation provide ?**

Bag of Words model provides invariance to spatial localization of certain features.

*If you use too many visual words, they will be very distinctive, particularly of the training images. If you now have a test image that shows the same object with small changes (e.g. a different brand of car), it might look a bit different and that visual word will therefore not be detected, b.c. the model did not see that particular car before.*

*If the dictionary is too small, then the visual words are not distinctive enough.*

*The good size of a dictionary is between a hundred and a few thousand visual words. Then the visual words are sufficiently specific that they can distinguish between categories, but not overly specific so that they still can generalize.*

**How would you account for spatial information while using vBoW feature**

**representation ?**

* part-bases representation: object as set of parts

=> model relative locations between parts and appearance of part

* pyramid math kernel: combine visual words with spatial information

=> use SVM and histogram kernels

**Can you devise a weighting scene for code words used in code book i.e only using**

**most discriminative words in the code books ?**

TODO

I think something like tf.idf weighting is meant here, i.e. weighting visual words more if they appear less frequent in the set of images that we use to build the dictionary.

k-means?

**Explain the Bayesian decision rule and its drawbacks ?**

Using a given x as input, we decide for category C1 if

We do not need normalization.

A classifier obeying this rule is called a Bayes optimal classifier.

The problem is that it assumes that we know what the likelihood and prior is and it is very difficult to know that.

**Can Linear Discriminant Classifier solve the XOR problem i.e. X/Data = [(-1, -1), (-1, +1), (+1, -1), (+1, +1)], Y/Classes=[-1, +1, +1, -1] ?**

No, an XOR problem cannot be solved with a linear discriminant classifier, because the data is not separable with a linear decision boundary.

**What are various ways of extending SVM for multi-class classification ?**

Question is not exam relevant (see Q&A session)

**Derive the gradient of the softmax non-linearity with respect to it's input.**

TODO

# 13 & 14 - Convolutional Neural Networks & 15 - Object Detection

**What is the limitation of softmax non-linearity for multi-class classification?**

TODO (was skipped in the lecture)

**In which scenario is gradient descent computationally expensive?**

The gradient is computed using all training data points. This step gets computationally expensive when there are a large amount of data points.

**Under what condition is a two layer neural network a universal approximator?**

Only on a closed interval where enough data is given

**What is the main drawback of the bag-of-words model?**

A bag-of-words model only considers the visual words and the frequency, but disregards spatial information of where the visual words are located in an image.

A bag-of-words model cannot distinguish two images if both have the same visual words in the same frequency, even if one of them has the visual words arranged in a for humans logical way.

**What is the role of a pooling layer in CNNs?**

It reduces the number of parameters in the classification layer/ the size of the feature maps. and increases the receptive field

**What is a receptive field in the context of CNNs?**

A receptive field is the part of the input that is responsible for the activation of a certain neuron. Each neuron has therefore only a small receptive field and thus is affected only a few dozen pixels of the input image. By using pooling, the feature resolution gets reduced. Therefore, a single neuron has a much bigger receptive field, b.c. it has more pixels in the input that influence it.

the region in the input space that a particular CNN's feature is affected by. A receptive field of a feature can be described by its center location and its size.

# Klausurfragen (Gedächtnisprotokoll)

TODO: Alles ohne Gewähr, kann sein das Antworten falsch sind, Verbesserungen gewünscht

**Wie funktioniert Camera Calibration? Erklären Sie die Hauptschritte davon. Welche Inputwerte sind gegeben?**

Via collinearity we get that xi x PXi = 0, to solve Ap = 0 we need 6 correspondences xi and Xi to create matrix A  
1. Build A from the first 2 rows of xi x Xi

2. Get p as right nullspace of A via SVD(A) and by taking the last right singular vector

3. To get K, *R* and c, P = K\*[*R*|*-R*c] = [M | m], by decomposing M via RQ-decomposition, here R = K and Q=*R*, c can be retrieved as the nullspace of P via SVD

**Was bedeutet Vignetting?**

Vignetting means that the borders of the image are black. This can be caused by lens flaws, if light rays get blocked.

**Was passiert, wenn das Loch des Pinhole Camera Modells zu klein oder zu groß ist?**

We get blurred images, if it is too large, the pencil of rays is large.

If the aperture is very small, blurring is caused by diffraction effects (Beugungsunschärfe). Also the picture gets darker when the aperture is smaller and might not be bright enough to work with.

**Was ist der Nachteil einer Linearen Repräsentation, wie z.B. PCA?**

The assumption that the subspace is linear is not reasonable. An object cannot be represented in a linear subspace.

PCA will only work if the images are well aligned in a sense that the orientation is canonical.

**Wieso werden im PCA-Verfahren Eigenvektoren mit den größten Eigenvalues betrachtet?**

The corresponding eigenvector is the direction of the maximum variance. The eigenvalue represents the magnitude of the direction of largest variance that decorrelates the data (under gaussian assumption).

Projecting onto them minimizes the reprojection error.

**Erklären Sie, auf Berechnung bezogen, die Beziehung zwischen Gaussian und Laplacian Pyramid. Wieso wird bei Laplacian Pyramid die Frequenzdekomposition durchgeführt?**

LoG can be approximated by a DoG

We can get Li = Gi - expand(Gi+1)

Li represents the frequencies that are lst by downsampling in the gaussian pyramid and can be recovered as Gi = Li + expand(Gi+1). We can thus recover the original image G0.

Li can also be used for sharpening.

What is meant with Frequenzkomposition? -> frequency decomposition into subbands

**Wie funktioniert der RANSAC Algorithmus?**

RANSAC-Algo:

1. Sample a minimal set of points (e.g. d=4 for homography)
2. construct the model (e.g. homography)
3. measure the support

repeat for k iterations that are retrieved via k = log(1-z)/log(1-w^d) to satisfy a desired z.

At the end, take the model with the most inliers, take all the inliers with this model and build a big equation system to compute the final model.

**Erklären Sie, wie Bag of Words Model funktioniert.**

Learning:

on a labeled dataset perform:

1. feature detection and representation, e.g. find keypoints and represent via SIFT
2. extract codewords via vector quantization (e.g. extract them via k-means clustering)
3. represent each training image via a BoW vector
4. train a classifier

Recognition:

1. find keypoints of query image
2. represent as BoW vector
3. predict

**Geben Sie die Definition des Harris Point Detektors an. Erklären Sie die Parameter von Harris Point detektor.**

*TODO:* nicht ganz sicher was mit Parametern gemeint ist.

Harris detector is an alternative for lambda\_, f = det(H) - alpha \* trace(H)^2 where H is the structure tensor, alpa is estimated empirically.

**Erklären Sie, wie die Kantendetektion bei der ersten und zweiten Ableitung funktioniert**

*TODO:* Evtl. ist hier auch der canny algorithmus gemeint bzw. dass man kernel benutzt?

1st order: edges = local optima

2nd order: edges = zero-crossings

**Wie wird die Boundary Behandlung bei der Kantendetektion durchgeführt?**

To ensure equal size regarding the original image the boundaries can be padded.

Different strategies are: constant(0-padding), wrap, clamp, mirror

**Welche drei Annahmen werden getroffen, um den Kanade Algorithmus durchführen zu können?**

TODO: kp was alle 3 sind, habe nur die 2 die direkt auf der Folie bei Taylor approximation stehen

1. small motion
2. Brightness constancy: image measurements (e.g. brightness) in a small region remain the same although their location may change
3. Neighboring points in the scene typically belong to the same surface and hence typically have similar 3D motions
   1. since they also project to nearby points in the image, we expect partial coherence in the image flow

**Welches Aperture-Problem tritt bei Image Motion auf?**

The linearized OFCE u\*Ix +v\*Iy + It = 0 is a line equation, but we have two unknowns u and v.

We can only find the normal velocity and can’t estimate the true motion. We therefore need multiple constraints i.e. use the spatial coherence assumption and estimate via a window.

**Angenommen die Kamera schießt ein Bild, rotiert um 30°, und schießt ein weiteres Bild. Kann der Stereovision Algorithmus durchgeführt werden?**

*TODO:* Nicht sicher. Auch nicht sicher ob Stereo Vision nur binocular stereo meint oder auch epipolar geometry

No, since stereo vision assumes a baseline > 0 to estimate depth.

**Was ist der Unterschied zwischen Fundamental und Essential Matrix?**

**F** = **K**1-T **E** **K**2-1

Combination of the Essential matrix **E** and the unknown calibration matrix **K**.

The epipolar constraint still holds.

Fundamental matrix is singular, has rank 2. It no longer holds that the two non-zero eigenvalues are the same, that holds only for calibrated cameras. E has 5 degrees of freedom, F has 7.

**Erklären Sie das Problem der Numerischen Stabilität beim Homography.**

In a 100x100 pixel image, the x and y pixel values lie in the range [0, 999]. Some entries of the matrix can be as big as a million (e.g. calculating xx’, with x=1000 and x’=1000) or as small as one. These large discrepancies between values can lead to numerical errors when using finite precision floating point arithmetic. Conditioning the matrix so that all the coefficients are in the same order of magnitude circumvents this problem.

**Welche Probleme können beim Window Based Matching in Stereovision auftreten?**

No window size fits every condition. Adaptive window size delivers better results.

small window: problems with homogeneous areas

large window: problems with textured areas

Also, the similarity constraint can be challenged by: repetition, textureless surfaces, specularities

**Angenommen das Bild wird geglättet(= Smoothing) und danach Ableitung berechnet? Wie kann diese Prozedur vereinfacht werden?**

We can save one convolution by first combining the gaussian filter and the derivative filter into a derivative of gaussian and then applying this to the image.

**Erklären Sie den Ansatz, wie die Anzahl der Hauptkomponenten beim PCA-Verfahren bestimmt werden.**

We can define n=0.9 s the amount of variance that should be captured by the projection.

We can then take the eigenvectors that correspond to the first D eigenvalues that satisfy:

Sum\_i=0\_D lambda\_i >= n\*Sum\_i=0\_M lambda\_i where M = number of eigenvalues

**Was ist der Unterschied zwischen Korrelation und ...(Keine Ahnung^^)?**

TODO: evtl. correlation und convolution?

Convolution requires mirroring the kernel in both x- and y-direction around the center (180 degree rotation of kernel), correlation does not.

Convolution operator: ∗

Correlation operator: ⊗

# <100 Words Definitions

**Computer vision**: Developing computational models and algorithms to interpret digital images and understand the visual world we live in.

**Application examples:** Face Recognition, Human pose estimation, Drive assistance

**Artistic cues:** Size, Light, vanishing point, objects further away get blurrier

**Human visual cues:** Stereo parallax, motion parallax, Shadow, Convergence

**Pinhole camera:** Object reflection lights go through an aperture of a barrier and hit a film/sensor, where the object is reproduced as an upside down image.

**Perspective projection:** uncalibrated projection; Projection is a matrix multiplication in homogeneous coordinates

**Orthographic projection:** Also called parallel projection. The distance from the image plane to the center of projection is infinite. → Camera calibration

**Coordinate transformations:** Extrinsic transformation takes world coordinates into camera coordinates. Intrinsic transformation takes camera coordinates and transforms them into pixel “coordinates”, which describes the image.

**Spatial sampling:** continuous signal \* sampling operator = sampled image

**Vanishing points:** Parallel lines in the real word converge in the image into a vanishing point.

**Perspective distortions:** There is a distortion in perspective, when exterior objects appear bigger while having the same size.

**Homogeneous coordinates:** Adding another coordinate (1).

**Projection matrix:** P = [ M | m ] = [ K R | K R c ̃ ]

**Calibration matrix / camera intrinsics:** K = (m\_x 0 0, 0 m\_y 0, 0 0 1)\*(f 0 p\_x, 0 f p\_y, 0 0 1) = (alpha\_x 0 beta\_x, 0 alpha\_y beta\_y, 0 0 1)

**Extrinsics:** Extrinsic transformation takes world coordinates into camera coordinates

**Linear camera calibration:**

**Homogeneous least squares:**

**Estimating projection matrix:**

**Aperture:**

**Thin lens formula:**

**Depth of field:**

**Field of view:**

**Camera artifact:**

**Color Cameras:**

**Bayer pattern:**

**Color sensors:**

**Linear filtering:**

**Convolution kernels:**

**Image smoothing:**

**Separable filters:**

**Boundary handling:**

Median filter:

Morphology:

Gaussian pyramid:

Aliasing:

Template-based recognition:

Edge detection:

Canny edge detector:

Non-maximum suppression:

Laplacian Laplacian pyramid:

Template-based matching:

SSD:

Subspaces:

Dimensionality reduction:

Correlation PCA:

SVD:

Eigenrepresentations:

Eigenfaces:

Appearance manifolds:

Interest points:

Structure tensor:

Harris points:

Hessian points:

Invariant features:

Scale space:

Scale selection:

Harris-Laplace points:

Performance evaluation:

Local descriptors:

SIFT features:

Shape context:

Homographies:

Coordinate normalization:

RANSAC:

Panorama stitching:

Stereo:

Triangulation:

Epipolar geometry:

Epipolar constraint:

Essential matrix:

Fundamental matrix:

Eight-point algorithm:

Binocular stereo:

Disparity:

Rectification:

Baseline:

Window-based matching:

Normalized correlation:

Motion field:

Optical flow:

Image interpolation:

Brightness constancy:

OFCE:

Aperture problem:

Lucas-Kanade:

Image registration:

Image warping:

Coarse-to-fine estimation:

View-based recognition:

Bag-of-Words model:

Color histograms:

Histogram distances:

Receptive field histogram:

Vector quantization:

BoW “features”:

Bayesian decision theory:

Discriminative & generative approaches:

Histogram & Pyramid match kernels:

Neural Networks:

Back propagation:

Stochastic gradient descent:

Convolutional Neural Networks:

Non-linearities:

Spatial pooling:

Network Architectures:

Convolutional vs. fully connected layers:

Receptive fields:

Role of depth in CNNs:

CNN training:

Adversarial examples:

Sliding window detector:

HOG:

Bootstrapping:

False positives etc.:

R-CNN and variants: