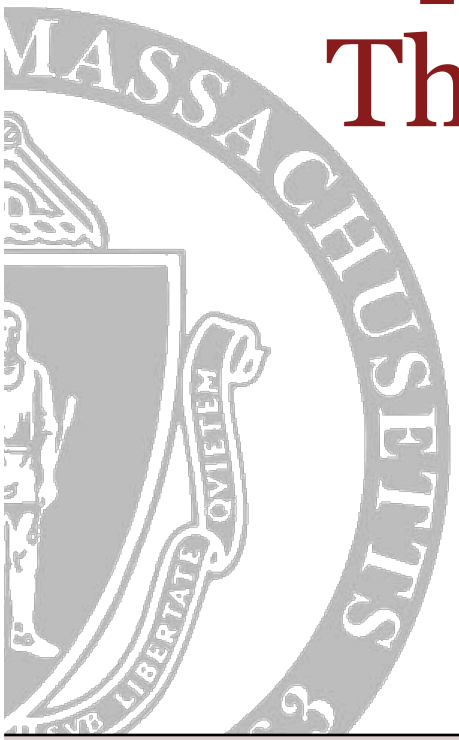
UMassAmherst

Lecture 3: Adaptive Image Similarity: The Sharpening Match



Erik Learned-Miller
with Laura Sevilla Lara, Manju Narayana,
Ben Mears, Evan Shelhamer

Computer Science Department

Lecture 3 in 1 slide

- In *comparing images*, people use "binning" to introduce spatial invariance.
- Big bins don't allow us to do fine discrimination.
- Small bins don't give us enough invariance.
- What should we do?
  Answer: Adapt the bin size specifically for the current images being compared.
  - *The sharpening match – a dynamic procedure for comparing images.*

Outline

- Similarity measures in vision—general remarks
- Piece 1: Sensitivity to position in image comparison
  - What's the right histogram bin size?
- Piece 2: Image matching with gradient descent
  - Overcoming problems with traditional blurring approaches using *distribution fields.*
- Putting the pieces together:
  the sharpening match.
- Some related results
  - Basin of attraction studies
  - Tracking experiments

Similarity measures in vision

- Right similarity measure depends on goal.
- The way humans evaluate similarity strongly depends upon what they are comparing.

How Similar are These Images?

How Similar are These Images?

Design a Similarity Function F such that…

F( , ) > F( , )

There is No Universal Similarity Function

There is No Universal Similarity Function

Totallylookslike.com

There is No Universal Similarity Function


"Lower Level"


"Higher Level"

Take Home Messages

1. The useful notion of similarity depends upon the goal.

Take Home Messages

1. The useful notion of similarity depends upon the goal.
2. Human similarity judgments are related to the **strength of our models.**

Strength of Models Example: Human Face Rec.

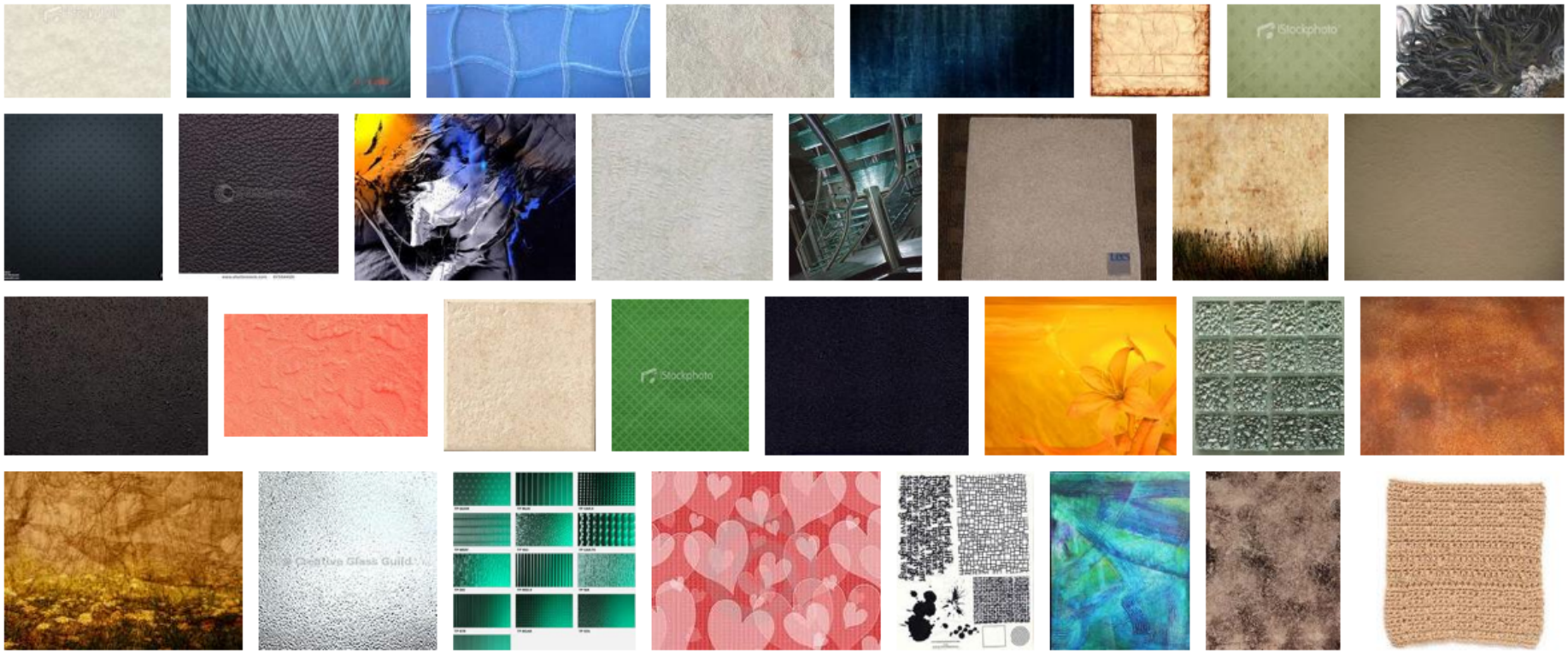
- Human models for upright faces
  - Very strong
  - Can distinguish among large number of faces
- Human models for upside-down faces
  - Less strong
  - Can’t distinguish among as many upside-down faces

Similarity as a function of model strength

Schwaninger et al., 2003

Low level similarity

- Try to establish similarity for very general images.



Outline

- Similarity measures in vision—general remarks
- Piece 1: Sensitivity to position in image comparison
  - What's the right histogram bin size?
- Piece 2: Image matching with gradient descent
  - Overcoming problems with traditional blurring approaches using *distribution fields.*
- Putting the pieces together:
  the sharpening match.
- Some results
  - Basin of attraction studies
  - Tracking experiments

Some "Low Level" Vision Problems

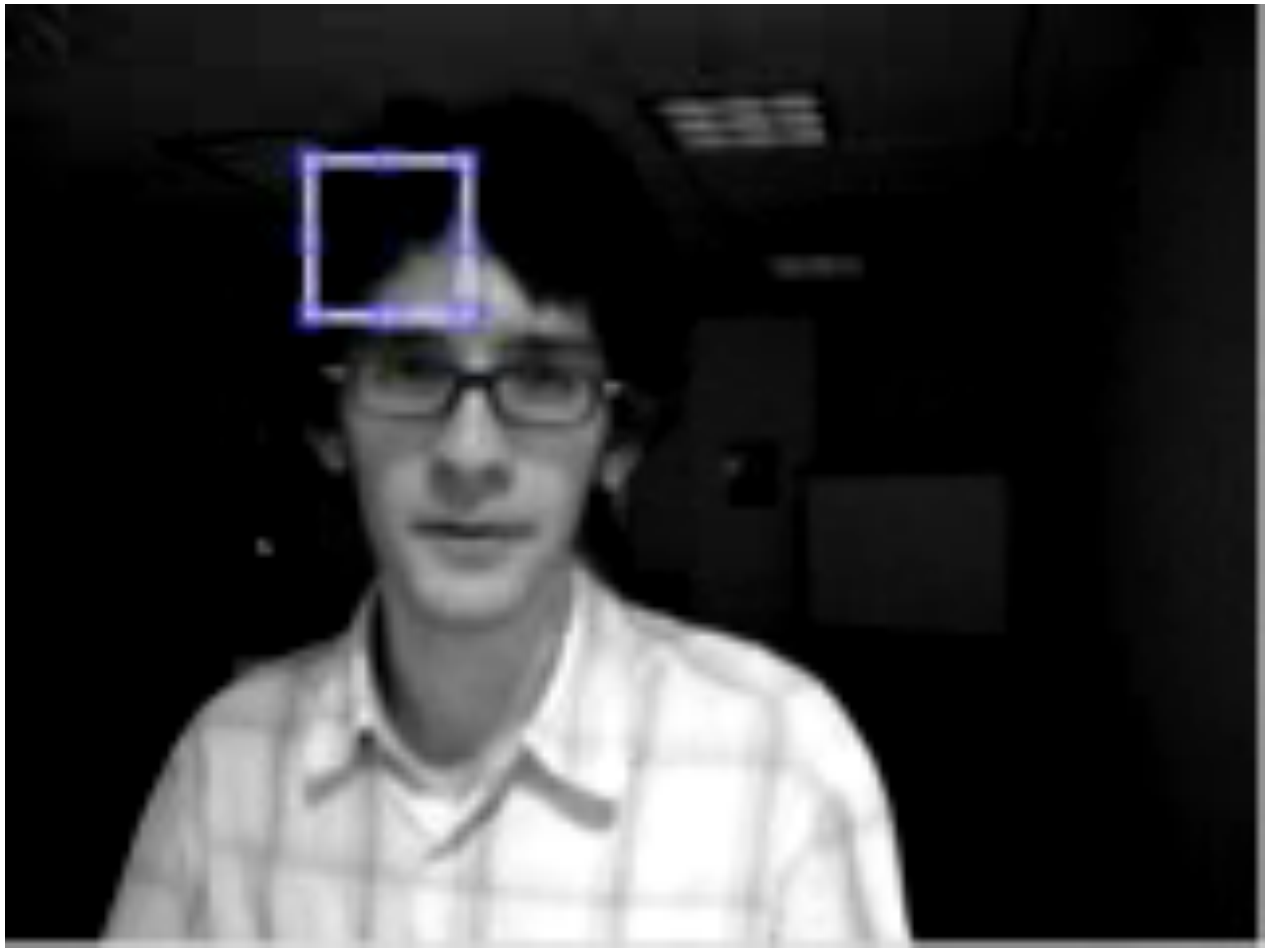
- Tracking
- Backgrounding
- Optical Flow
- Stereo
- Affine Invariant Matching
- Medical image registration
- Image stitching

Some "Low Level" Vision Problems

- Tracking
- Backgrounding
- Optical Flow
- Stereo
- Affine Invariant Matching
- Medical image registration
- Image stitching

- What makes these "low level"?
  - Weak models of appearance

A Sample Application: Tracking of General Objects



Basics of Tracking

Frame T

Frame T+d

Basics of Tracking

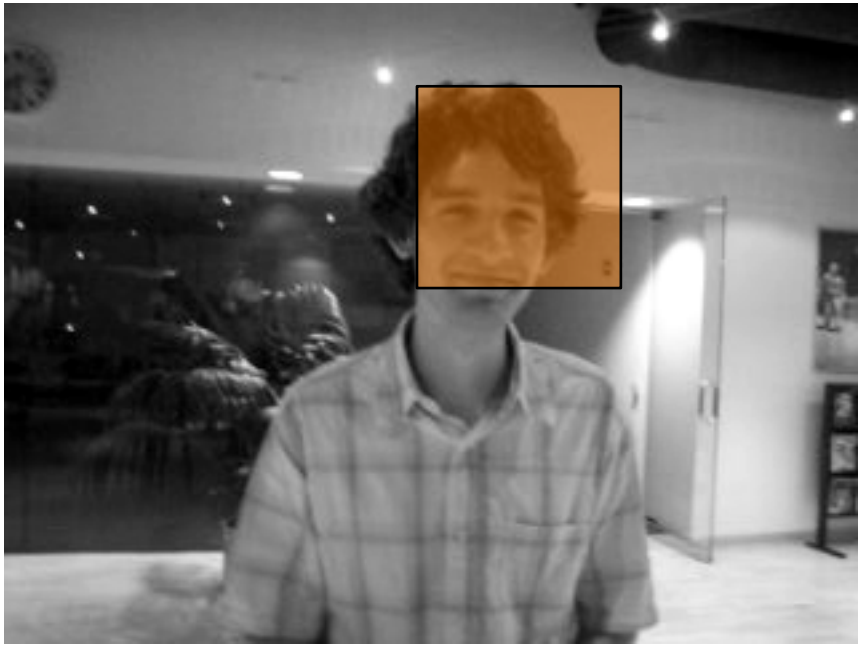
Frame T

Frame T+d

Basics of Tracking

Frame T

Frame T+d



Basics of Tracking

Frame T

Frame T+d

Basics of Tracking

Find best match of patch I to image J,
for some set of transformations.

patch I

image J

The core alignment problem

- Given a patch in one image, find the region in another image that is as similar as possible to that region.
  - What similarity function?
    - Image representation
    - Comparison function
  - What method to find the optimum?

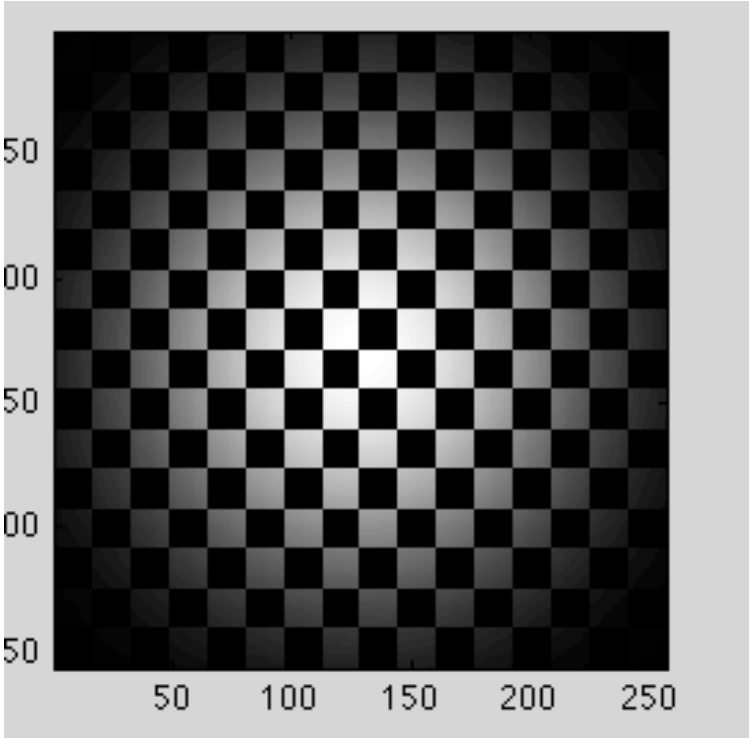
Common pixel-based similarity measures

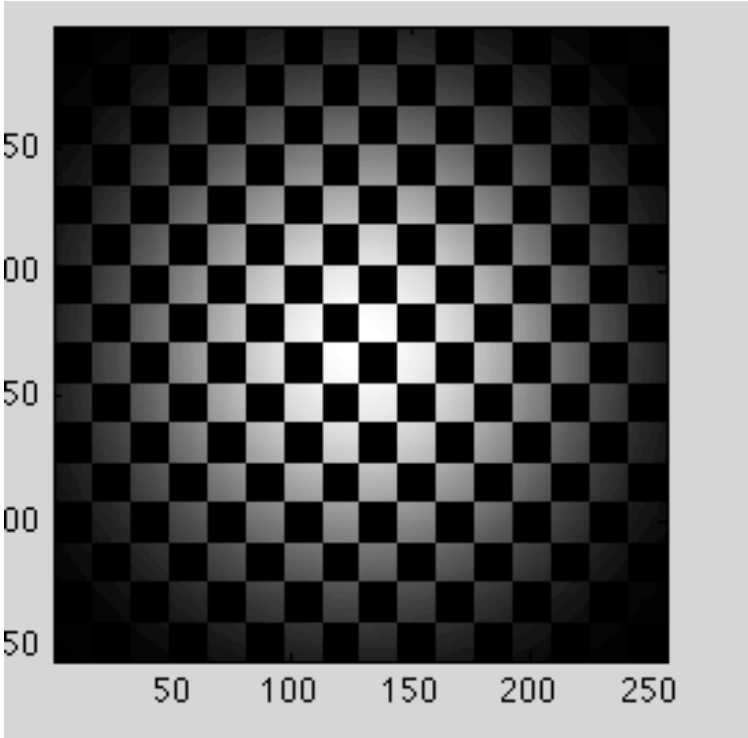
- L2 (Euclidean)
  - Square root of sum of squares differences in pixels
- L1
  - Sum of absolute value of pixel differences
- Correlation measures
  - Are brightness values in image correlated?
  - Maximum value of 1
  - Minimum value of -1
  - Value of 0 implies no linear relationship.

A Strange State of Affairs…

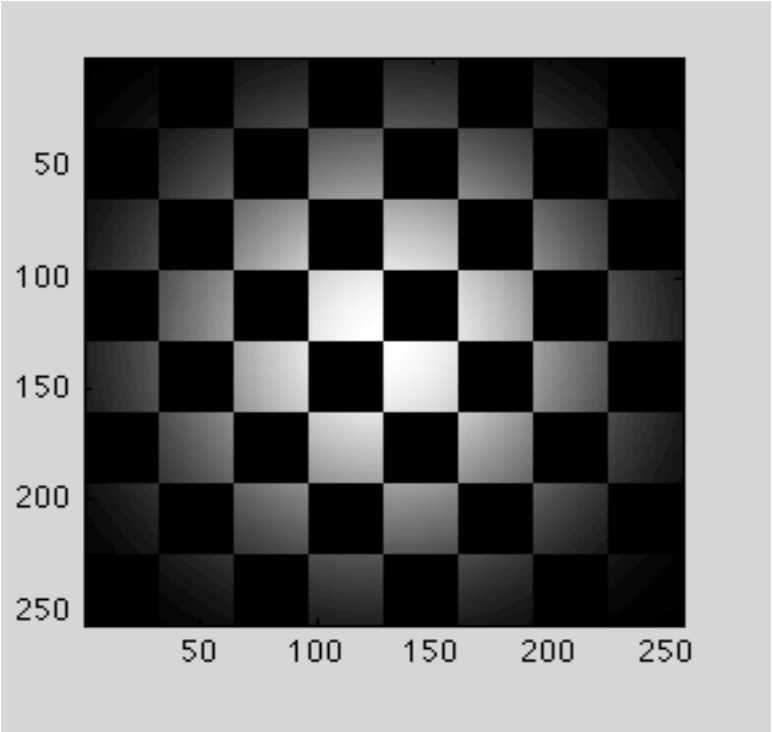
- For some pairs of images, the human notion of similarity is nearly *opposite* to common notions of similarity used in computer vision.

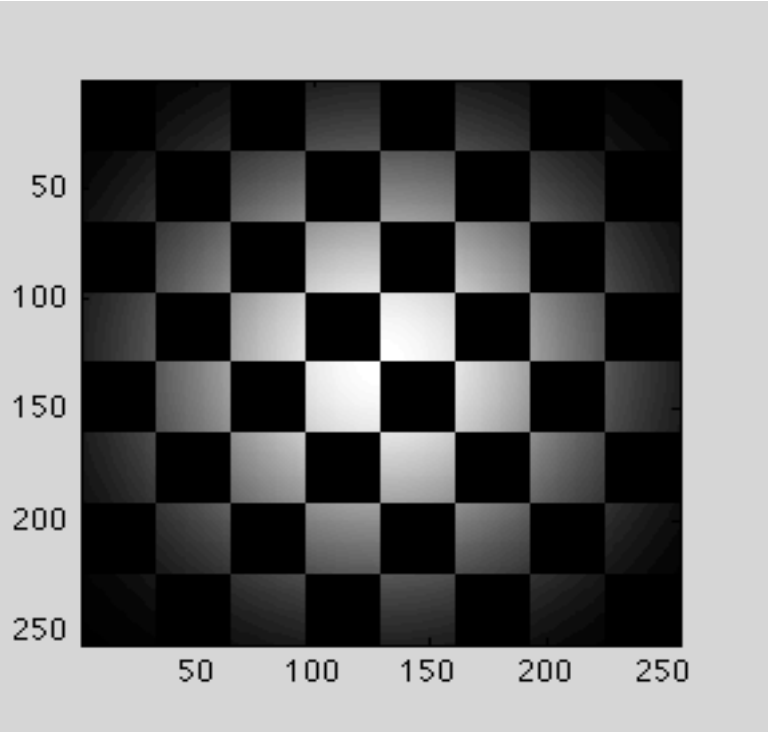
How similar are these images...



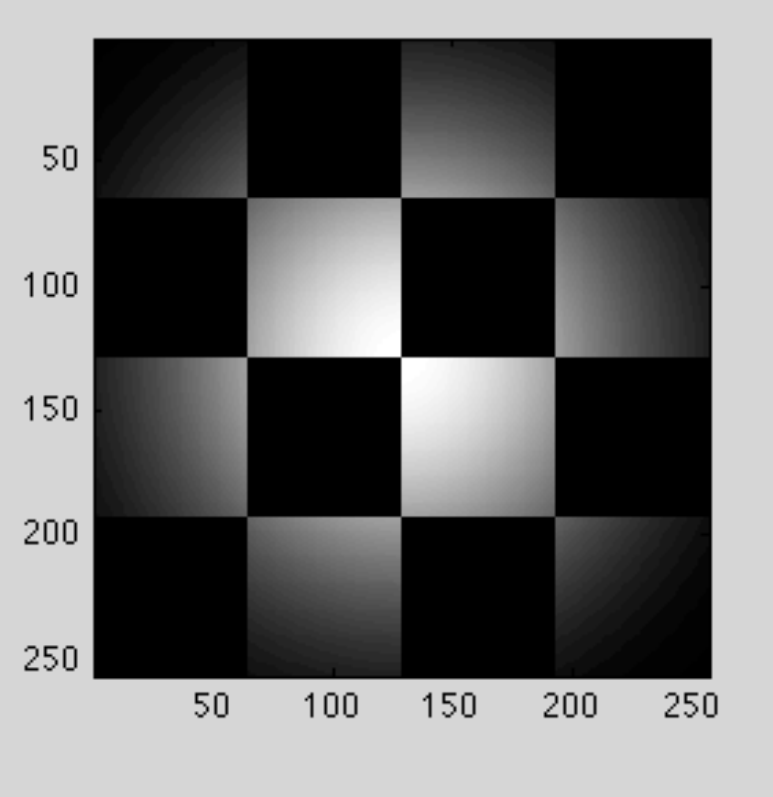


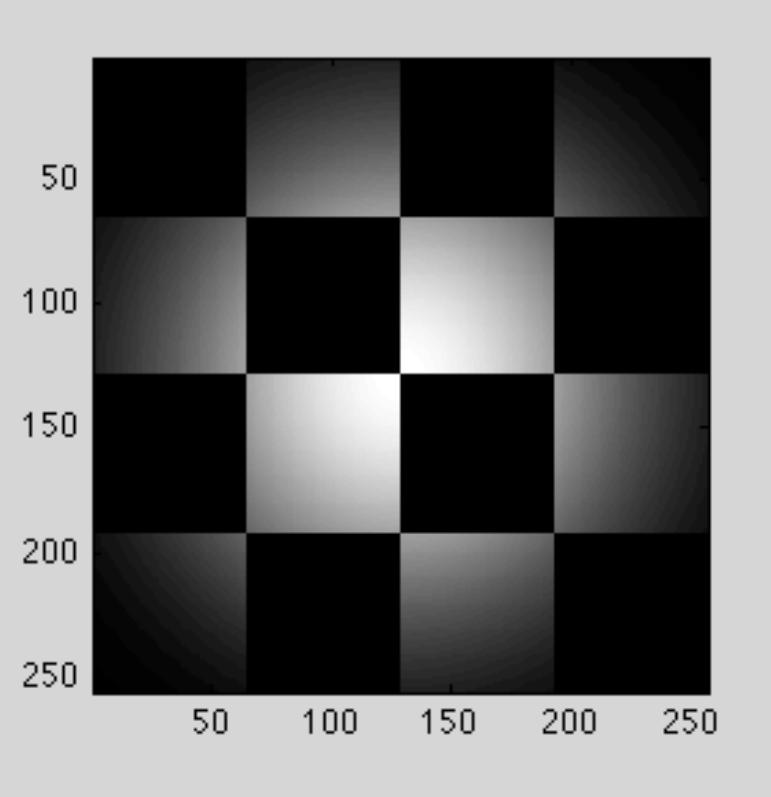
How about these?



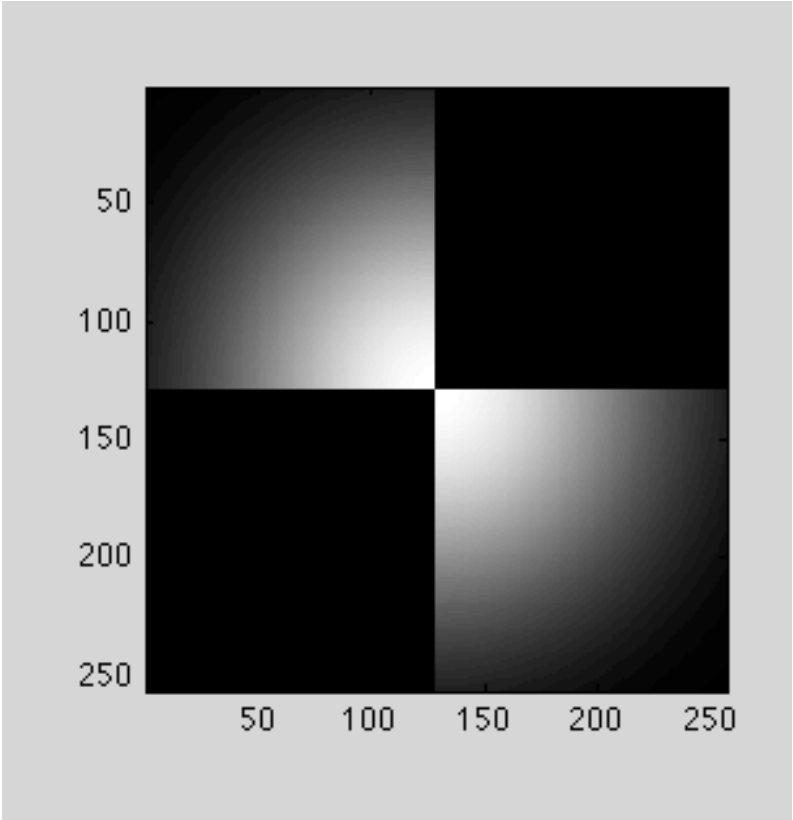


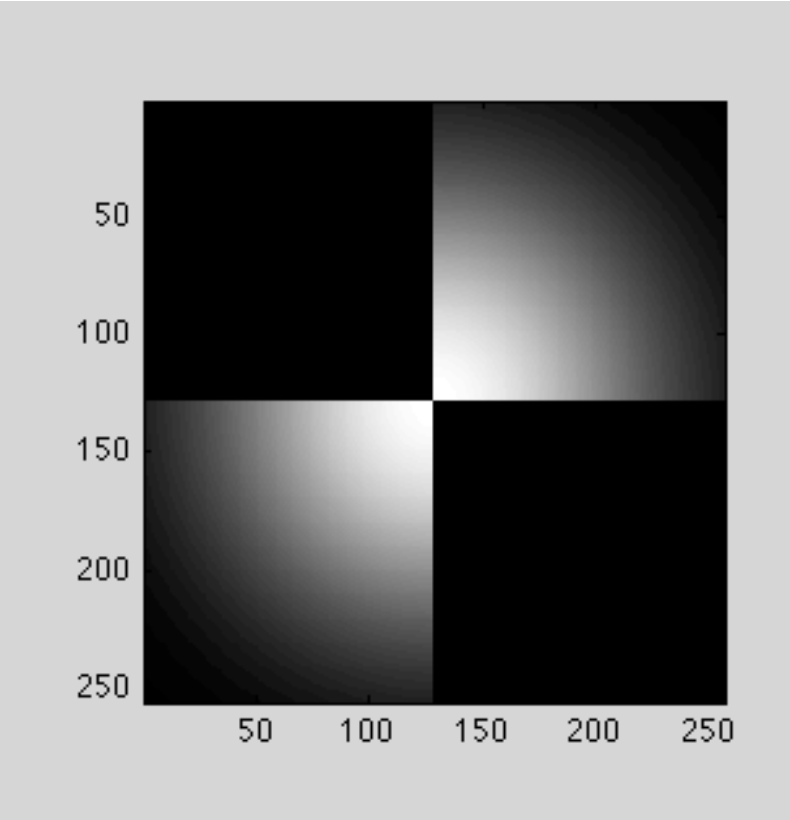
or these


50
100
150
200
250
50
100
150
200
250


50
100
150
200
250
50
100
150
200
250

or these?


50
100
150
200
250
50
100
150
200
250


50
100
150
200
250
50
100
150
200
250

Observation

- It would appear that humans don’t care about precise alignment (in all cases).

Pixel representations

Pixel representations

Squared differences

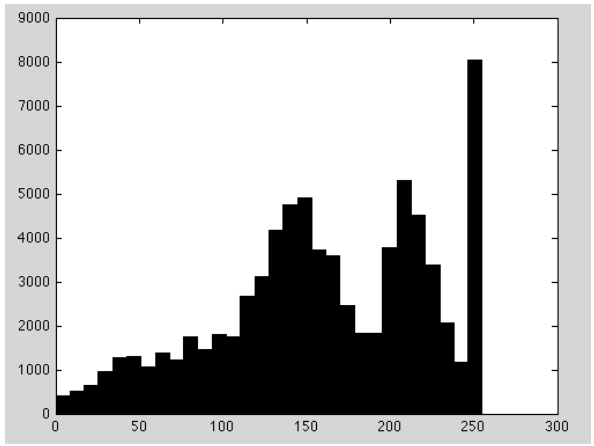
Conclusions

1. Pixelwise representations:
   overly sensitive to position

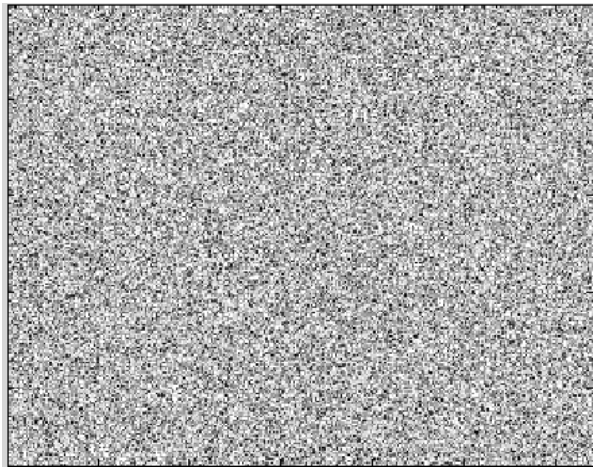
Conclusions

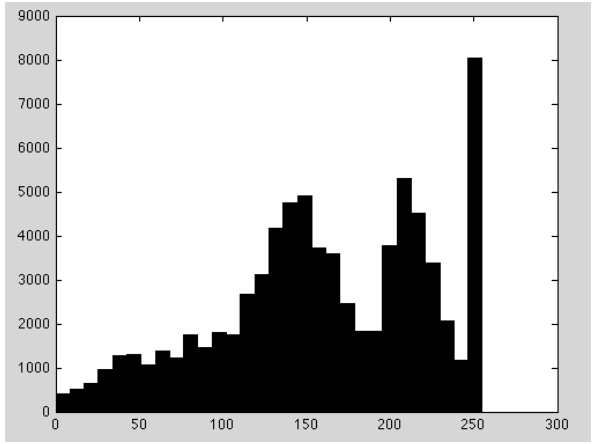
1. Pixelwise representations:
   overly sensitive to position
2. Histogram representations:

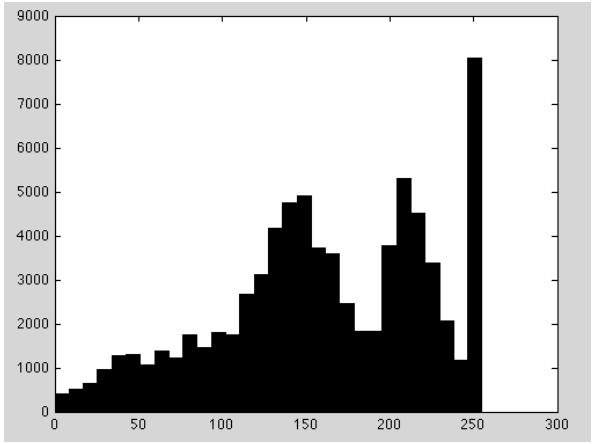
Gray value histogram comparisons



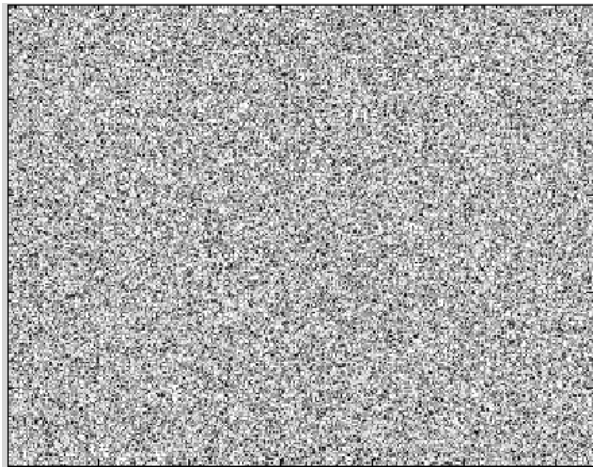
Gray value histogram comparisons







Gray value histogram comparisons

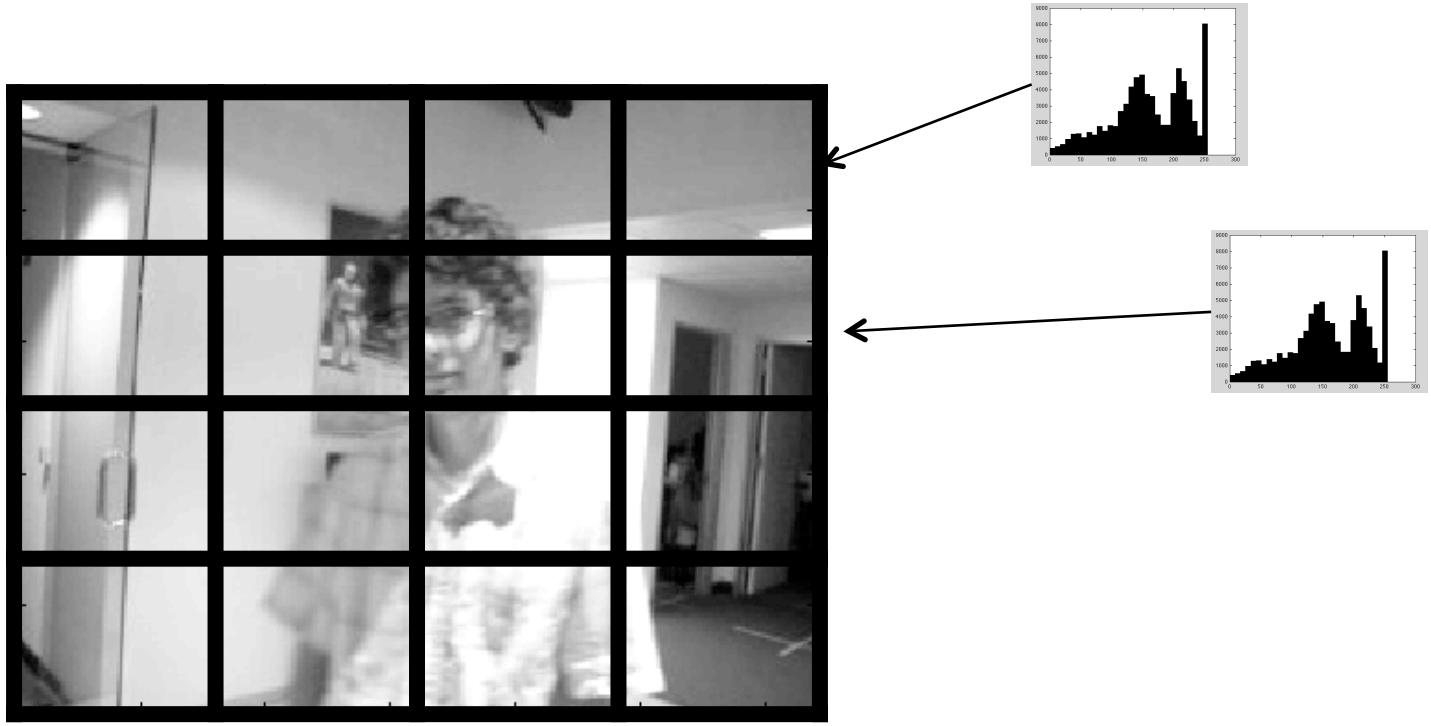


They're equal

Conclusions

1. Pixel representations:
   overly sensitive to position
2. Histogram representations:
   under-sensitive to position

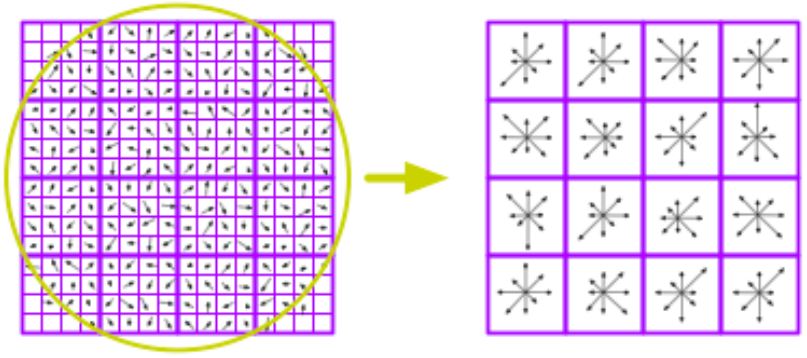
The Standard Compromise



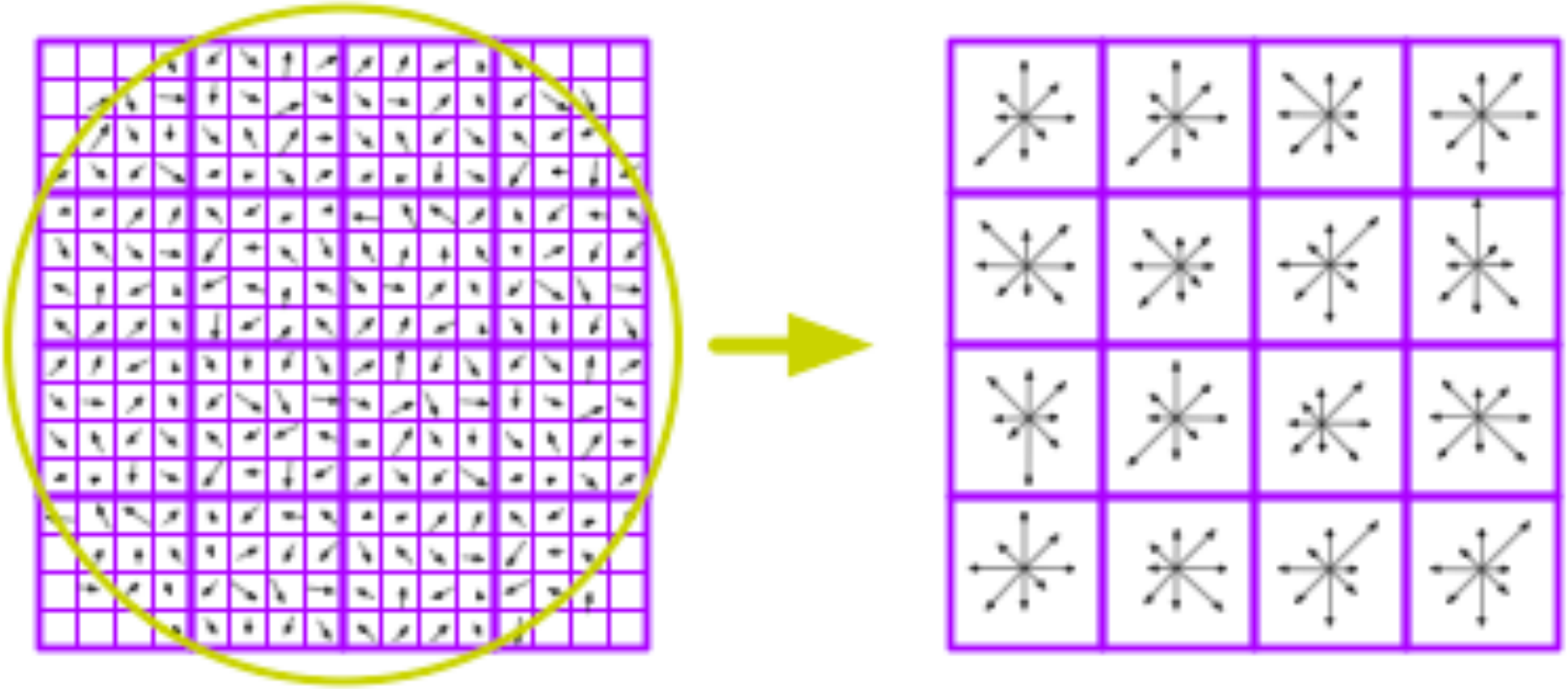
A separate histogram for each region.

Spatial invariance

- Usually achieved by integrating, averaging, or taking a max over a neighborhood
  - Binning (SIFT, HOG, histograms)
  - "max pooling" (deep belief nets)
- Array of histogram descriptors (non-overlapping)
  - SIFT, HOG, generalized shape contexts, …
  - Dominate vision apps



Quick SIFT/HOG tutorial/Shape Context tutorial



The fundamental dilemma

- Larger bins
  - more spatial invariance
  - more fundamentally different images map to the same descriptor
- Smaller bins
  - higher specificity
  - less invariance

The fundamental dilemma

- Larger bins
  - more spatial invariance
  - more fundamentally different images map to the same descriptor
- Smaller bins
  - higher specificity
  - less invariance

- *What bin size should we use?*

Suppose we are given the optimal bin size...

Suppose we are given the optimal bin size…

- Claim: the descriptor still stinks!

Suppose we are given the optimal bin size…

- Claim: the descriptor still stinks!

- We define two properties that similarity functions should have, and show that no similarity function based on an array-of-bins descriptor can have both properties.

Minimal descriptor requirements

- Property 1: "norm-like"
  - Build an image distance function using the descriptor and any standard vector distance (L1, L2, L_inf).
  - Minimum distance should be attained only when I=J.
  - We call the behavior of such an image comparison function "norm-like".

Minimal descriptor requirements

- Property 1: "norm-like"
  - Build an image distance function using the descriptor and any standard vector distance (L1, L2, L_inf).
  - Minimum distance should be attained only when I=J.
  - We call the behavior of such an image comparison function "norm-like".
  - Not satisfied by ANY histogram descriptor, since multiple images can map to the same descriptor.

Minimal descriptor requirements

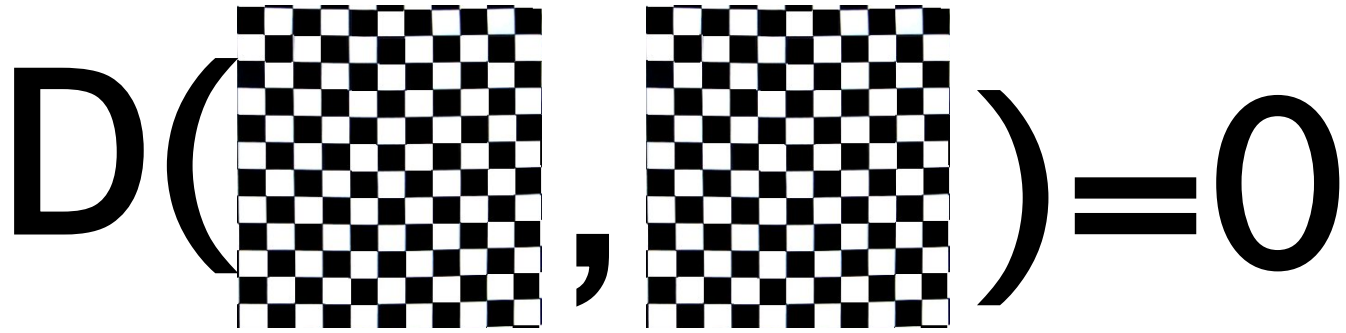
- Property 2: weak invariance to position
  - Goal: "small" translations of an image, or portion of an image, should have "small" impact on similarity function

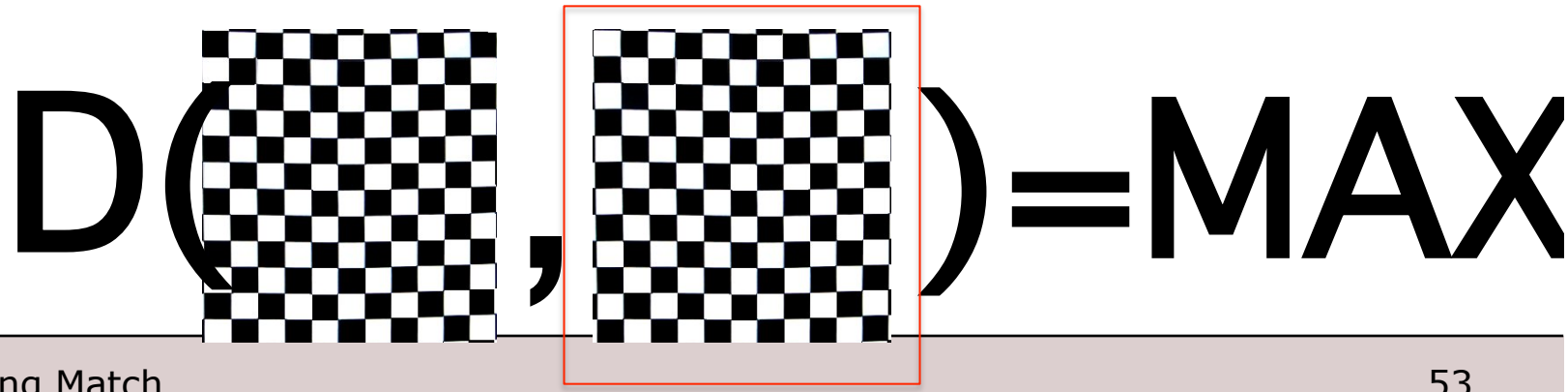
Minimal descriptor requirements

- Property 2: weak invariance to position
  - Suppose D(I,J) = 0
  - Let K be a the image J translated by a single pixel.
  - Now suppose that D(I,K)= MAX
    - MAX = maximum possible value of distance function.
  - In this case, we say that the distance function *fails to exhibit weak invariance to position*

Weak Invariance to Position: Failure

- Under L2 metric on pixel values:





Shocking Result!

- No image distance based on an array-of-histogram descriptor can satisfy BOTH properties 1 and 2.
  - If bin size > 1, property 1 fails
    - Why? Multiple images map to same descriptor. Not norm-like.
  - If bin size = 1, property 2 fails
    - Fails weak invariance test for checkerboard image.

Implications

- Using these descriptors we can either
  - A) Not tell when images are the same, or
  - B) Not consider images that are virtually equivalent (up to a 1 pixel translation) to be even remotely similar.

- What's the resolution of this problem?

Implications

- Using these descriptors we can either
  - A) Not tell when images are the same, or
  - B) Not consider images that are virtually equivalent (up to a 1 pixel translation) to be even remotely similar.

- What's the resolution of this problem?
  - *Adaptive bin sizes...*

Outline

- Similarity measures in vision—general remarks
- Piece 1: Sensitivity to position in image comparison
  - What's the right histogram bin size?
- Piece 2: Image matching with gradient descent
  - Overcoming problems with traditional blurring approaches using *distribution fields.*
- Putting the pieces together:
  the sharpening match.
- Some results
  - Basin of attraction studies
  - Tracking experiments

Finding the optimum alignment

- Exhaustive search
- Gradient descent
- Keypoint methods

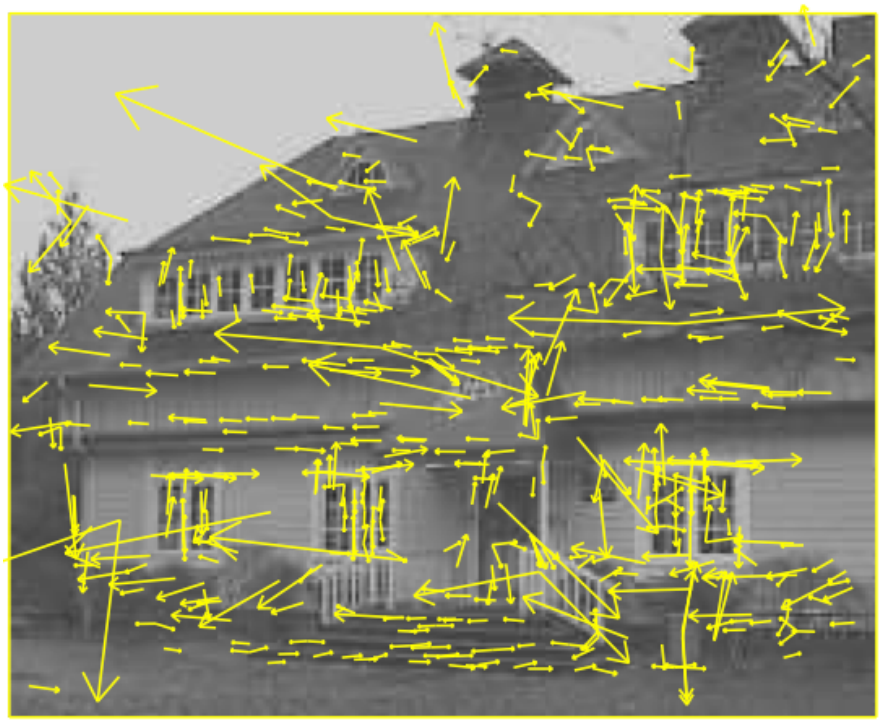
Exhaustive search

- Matching a 10x10 patch to a 100x100 image with 256 pixel values:
  - Integer-valued translations: 90x90 = 8100
  - Sub-pixel translations: 8100*256*256 = 2^29
  - Translations and rotations: about 2^40
  - Similarity: 2^50
  - Affine: 2^70

Keypoint methods

- Define "special" locations in image.
  - Local brightness extremum
  - Local edge energy extremum
  - "reddest" point locally
  - etc.
- Find all such special points in patch and image.
- Try to find a mapping from patch points to image points.

Keypoints



Keypoint matching

Finding the optimum alignment

- Exhaustive search - too slow for large sets of transformations
- Keypoints: not repeatable for far-field tracking, tracking with occlusion, or tracking low-texture objects
  - Many features are not "dense"
- Gradient descent
  - Often can't tolerate large displacements, *but good for many low level vision problems.*

Finding the optimum alignment

- Exhaustive search - too slow for large sets of transformations
- Keypoints: not repeatable for far-field tracking, tracking with occlusion, or tracking low-texture objects
  - Many high level features are not "dense"
- Gradient descent
  - Often can't tolerate large displacements, *but good for many low level vision problems.*

Gradient descent and human vision

- Human vision has two basic modes of object search:
  - Iterative saccades
  - Smooth pursuit
- Gradient descent is analogous to smooth pursuit, which most intelligent animals are very good at.

Gradient Descent Alignment

Find best match of patch I to image J,
for some set of transformations.

patch I

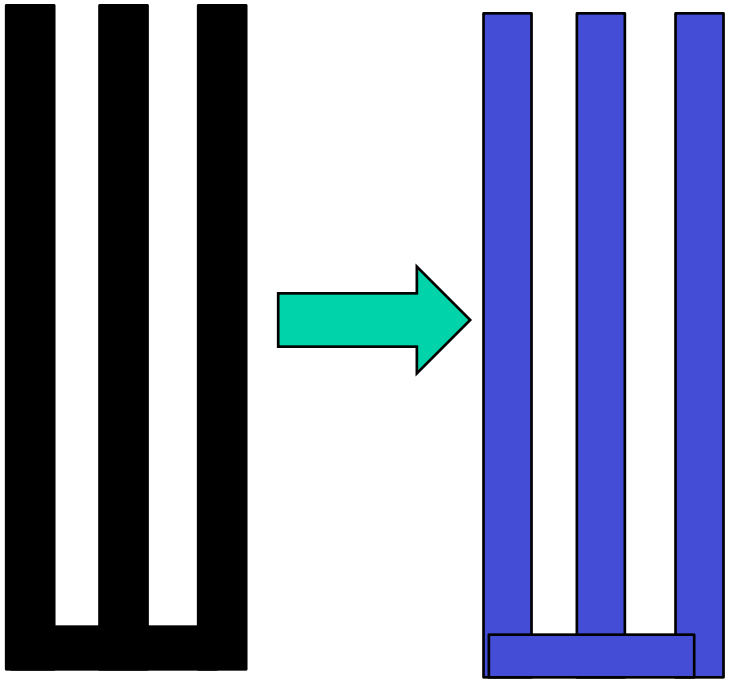
image J

Difficulties with gradient descent (minimizing distance)

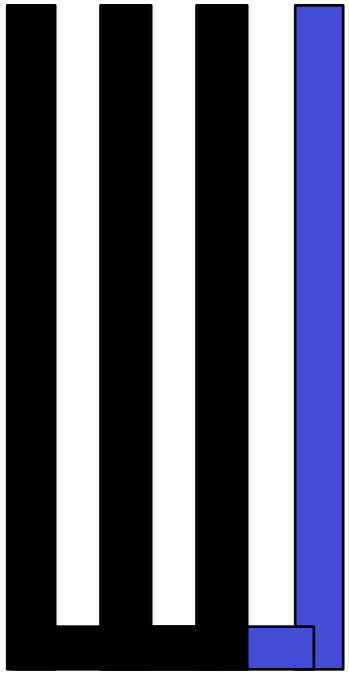
- Zero gradient problem:
  - Moving patch I doesn't change similarity function.
- Local optima:
  - We're at a minimum, but it's the wrong one.

Local optimum problem in alignment

Unaligned



Stuck in a local optimum



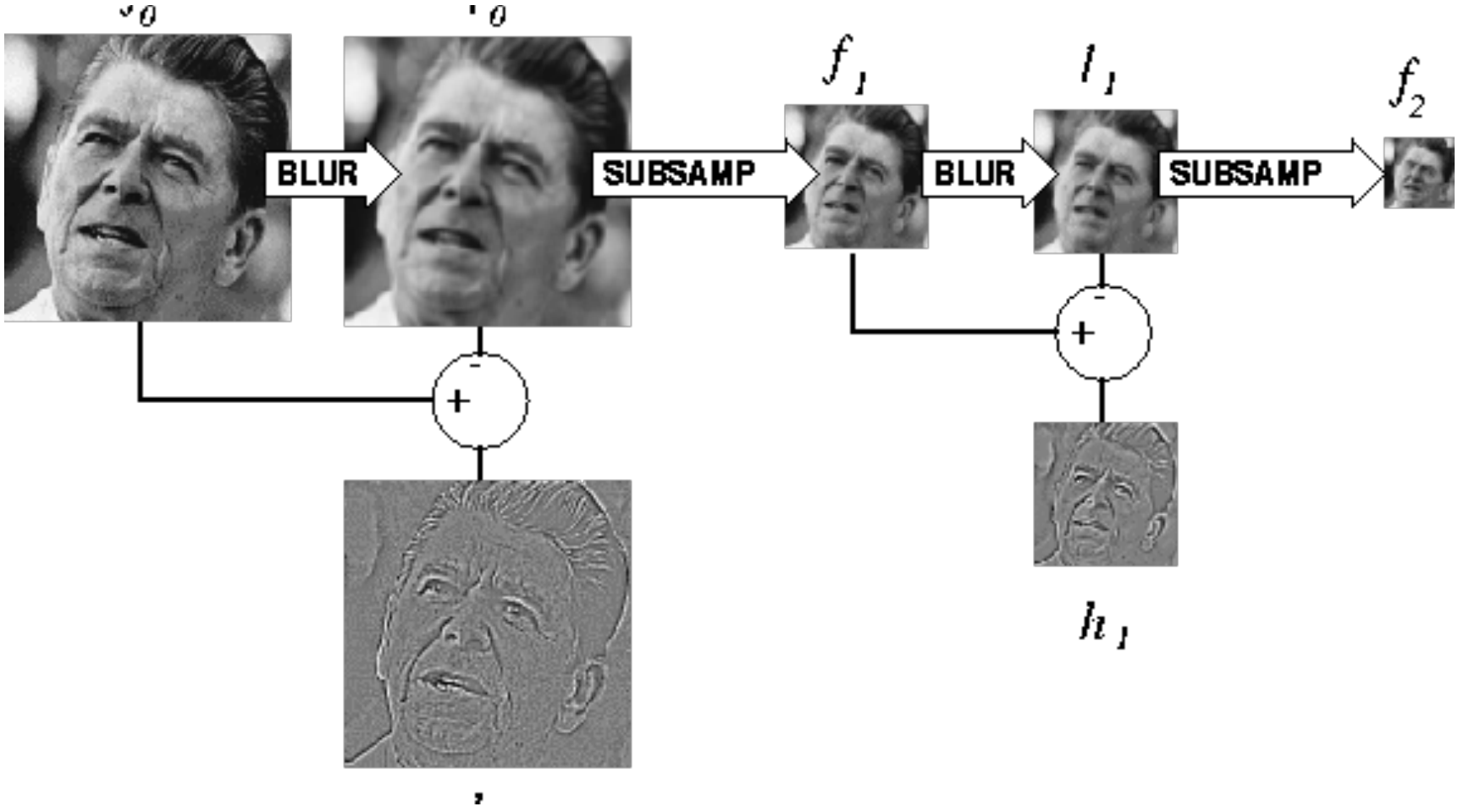
Common solution to gradient descent matching

- Blur images?
  - "Spreads information"
  - Also destroys information through averaging

Image Pyramids

- Basic pyramid:
  - Half the resolution (via sampling or interpolation) at each level. Number of levels: Log(n).
- Gaussian pyramid:
  - Gaussian blur the image, then subsample.

Gaussian Pyramid



Gaussian Pyramid

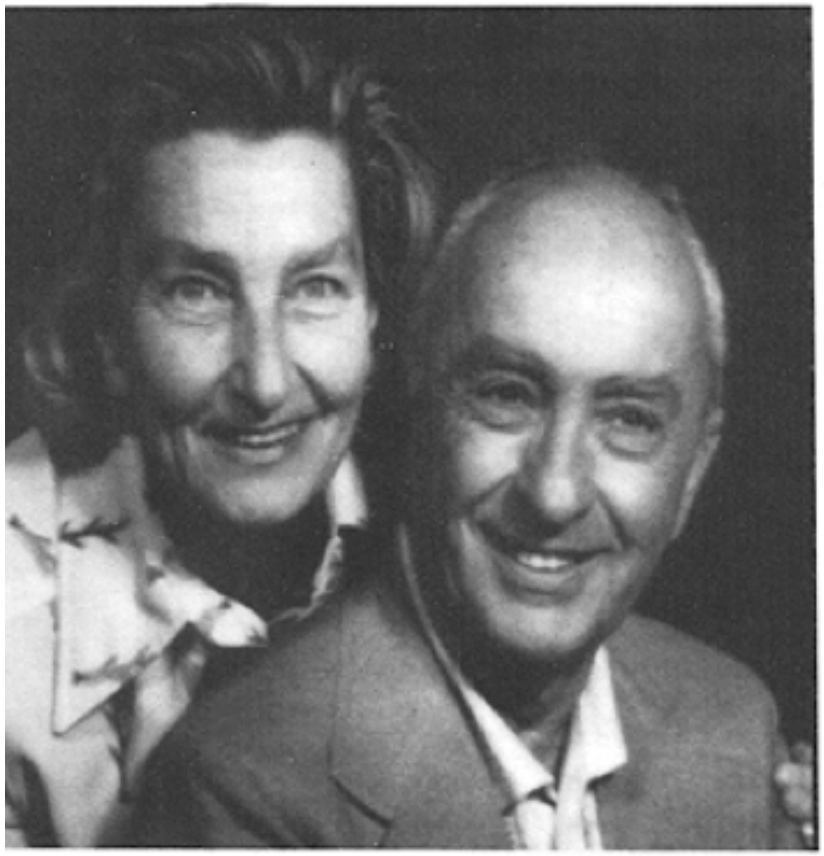


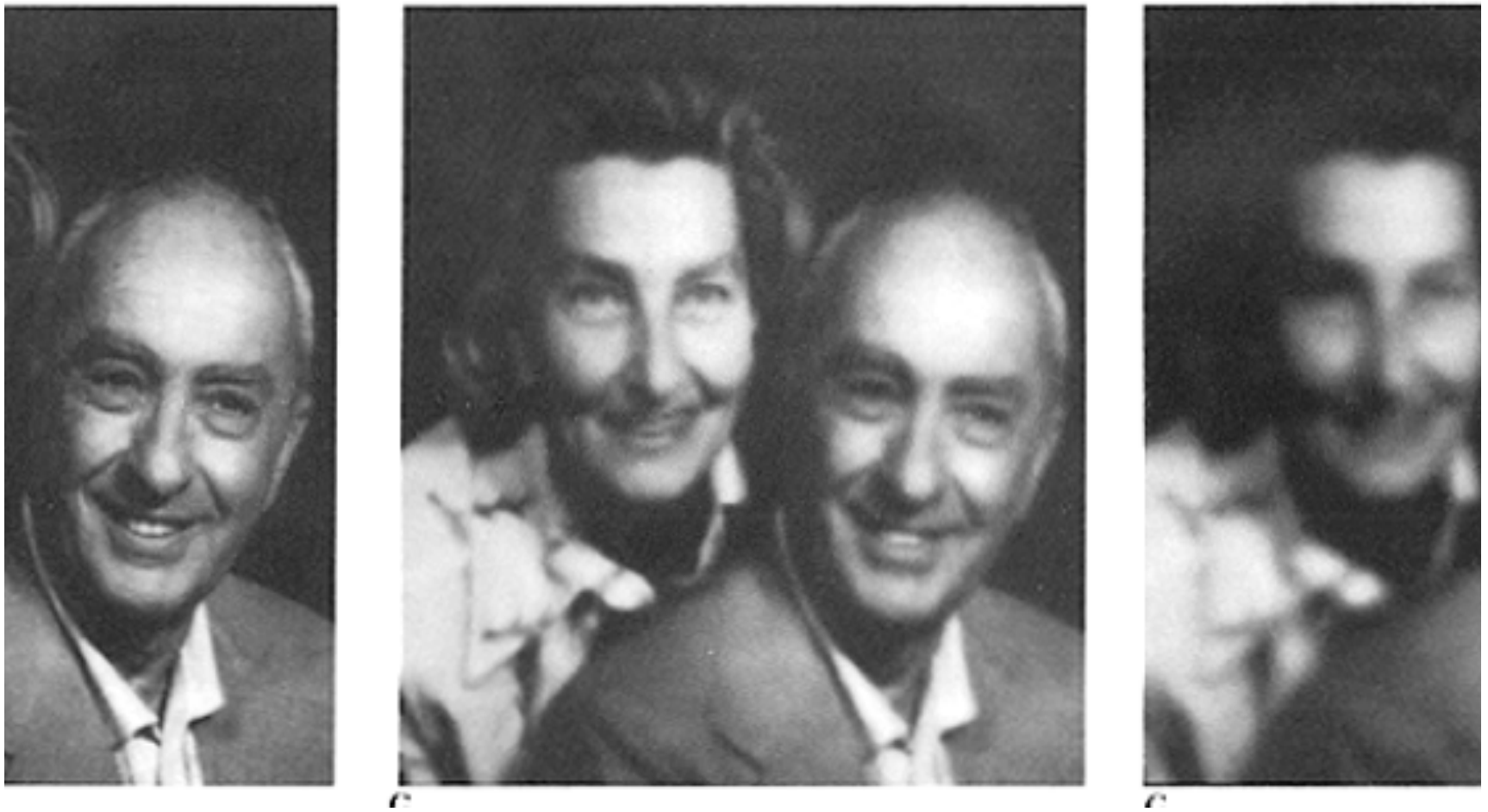
Fig. 2a. *The Gaussian pyramid. The original image, G repeatedly filtered and subsampled to generate the seque of reduced resolution image G_1, G_2, etc. These compris set of lowpass-filtered copies of the original image in wi the bandwidth decreases in one-octave steps.*

G_1

G_2

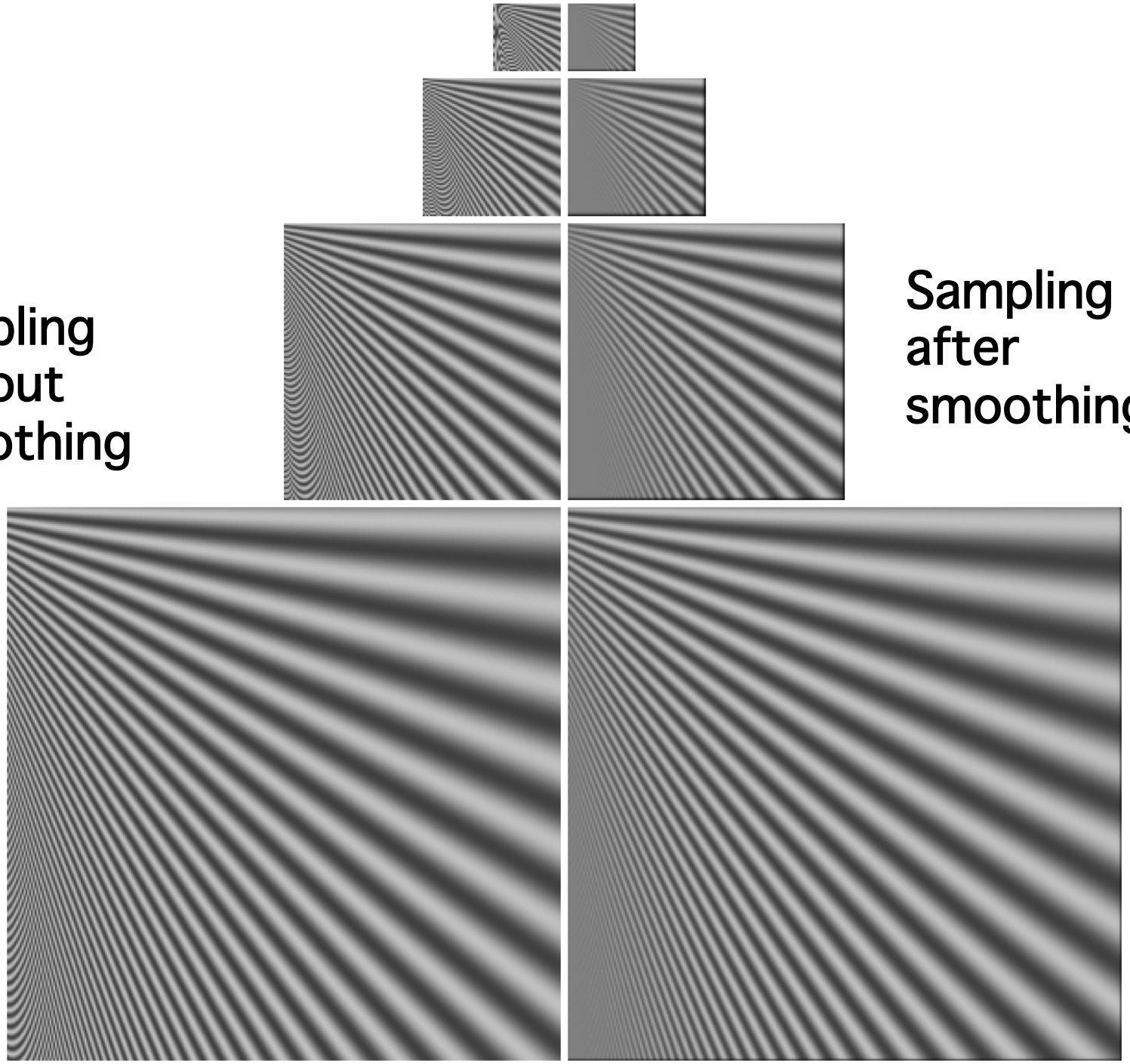
G_3

Gaussian Pyramid



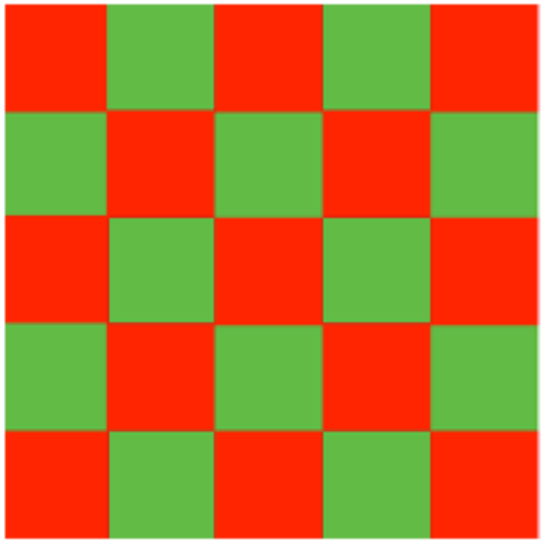
Sampling without smoothing

Sampling after smoothing

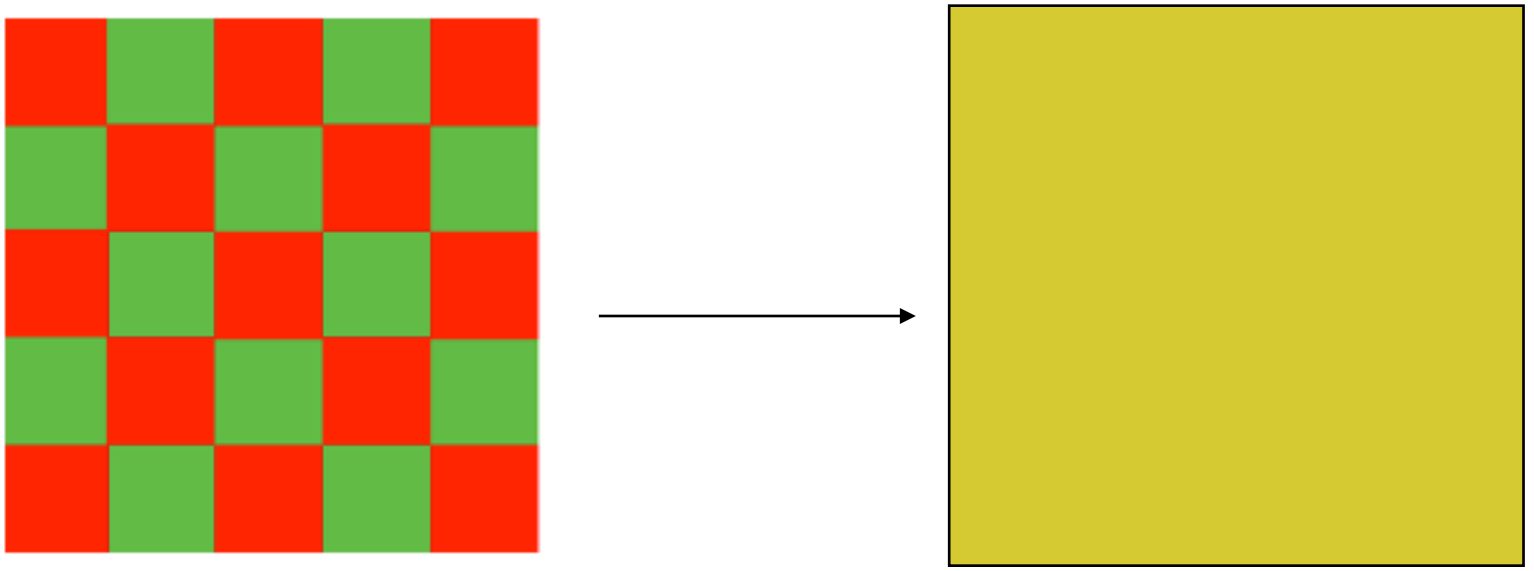


When does blurring lose the target?

What happens to this under blurring?



What happens to this under blurring?



Observation

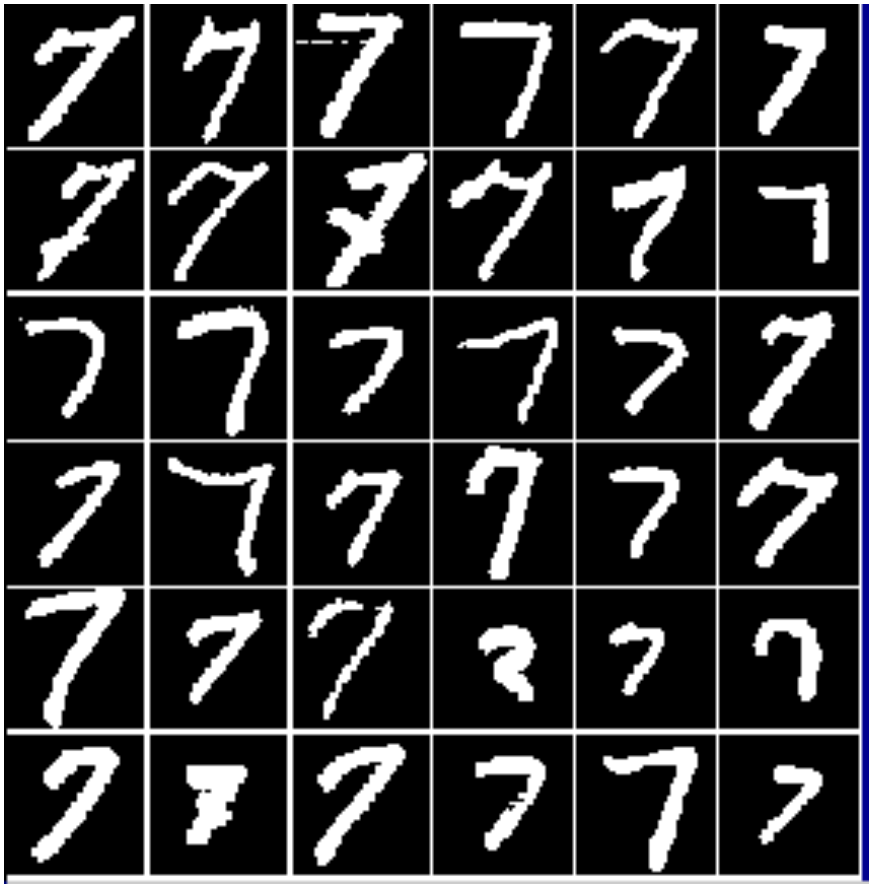
- Changing the representation to help find the optimum can make the representation significantly worse.

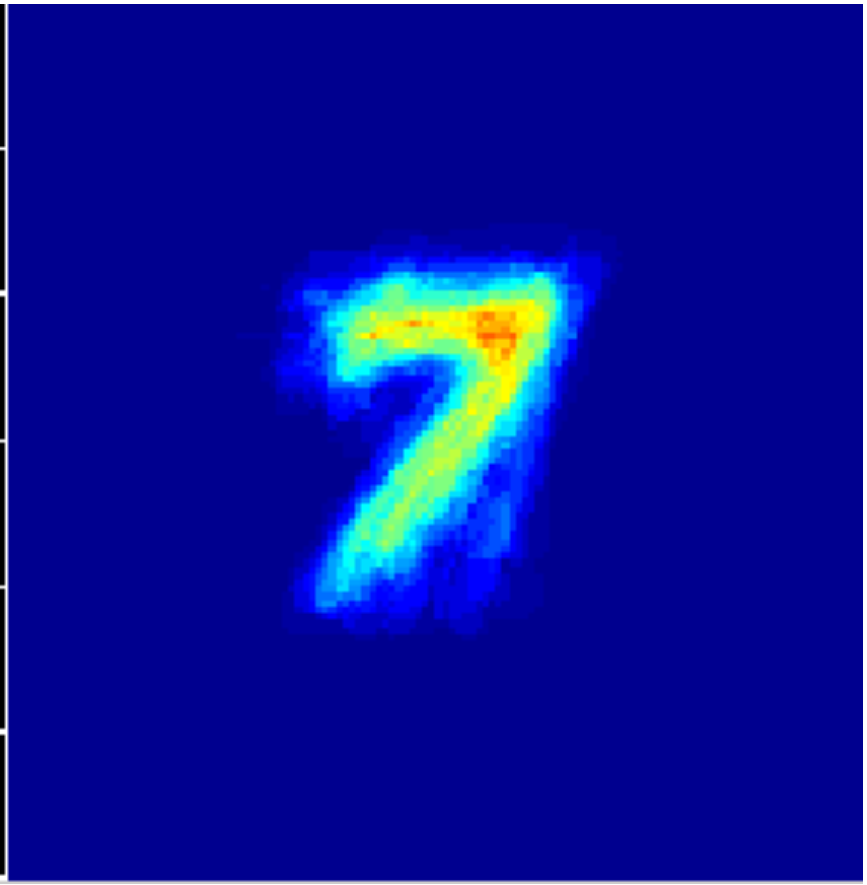
Observation

- Changing the representation to help find the optimum can make the representation significantly worse.

- Question: Can we smooth the optimization landscape without destroying image information?

Congealing (CVPR 2000)





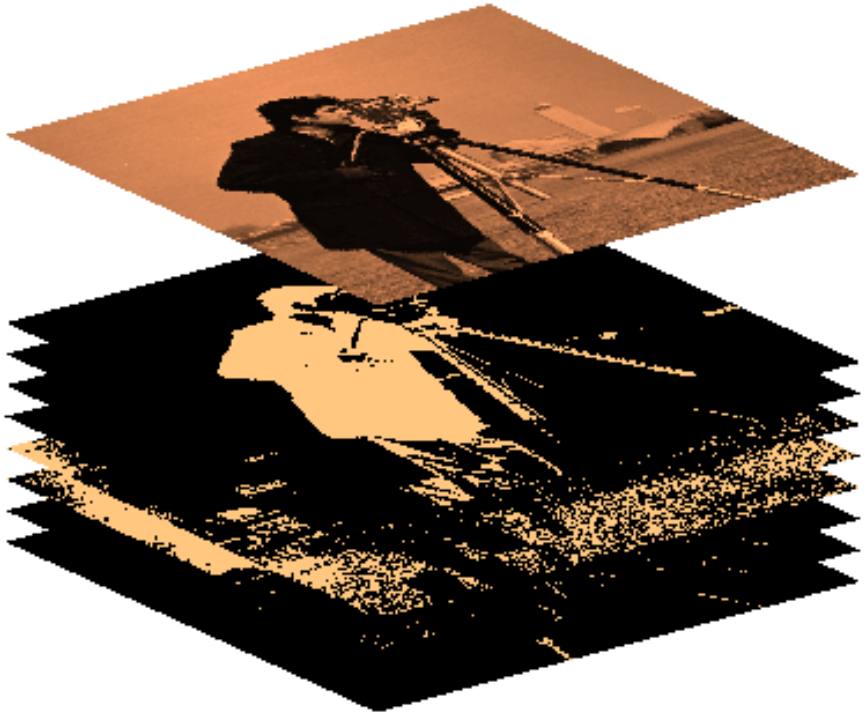
Properties of Congealing

- Smooths the optimization landscape without smoothing individual images.
- Has large "basin of attraction".
  - Few images get stuck in local minima
  - Few images have zero-gradient problem

Congealing with 2 images?

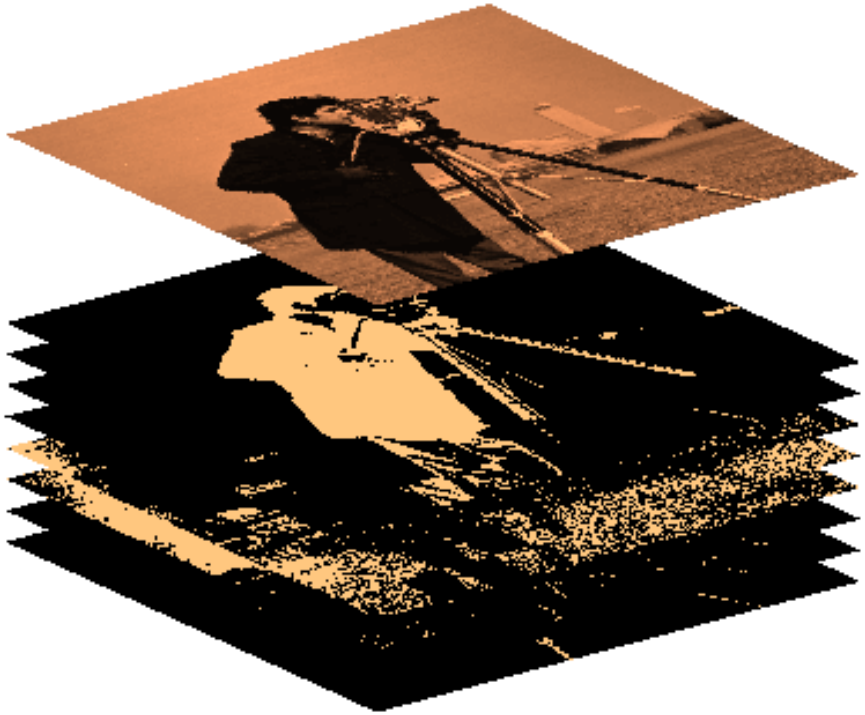
- How can we get the benefits of congealing without a large stack of images?

Exploding an image

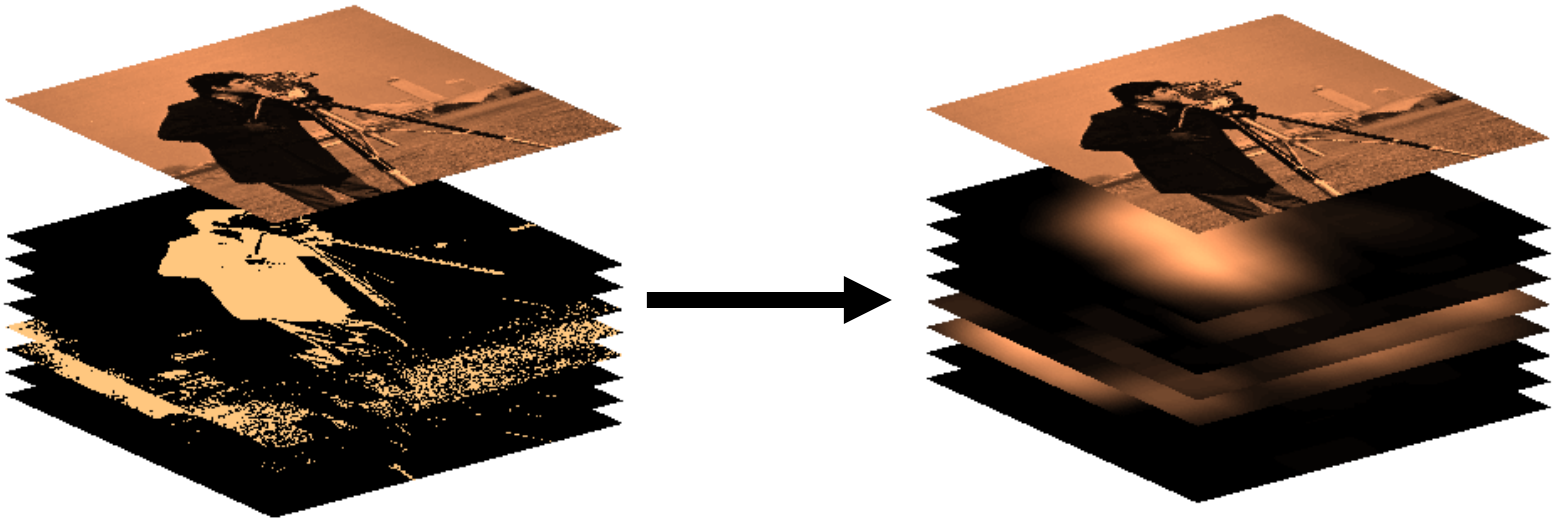


Exploding an image

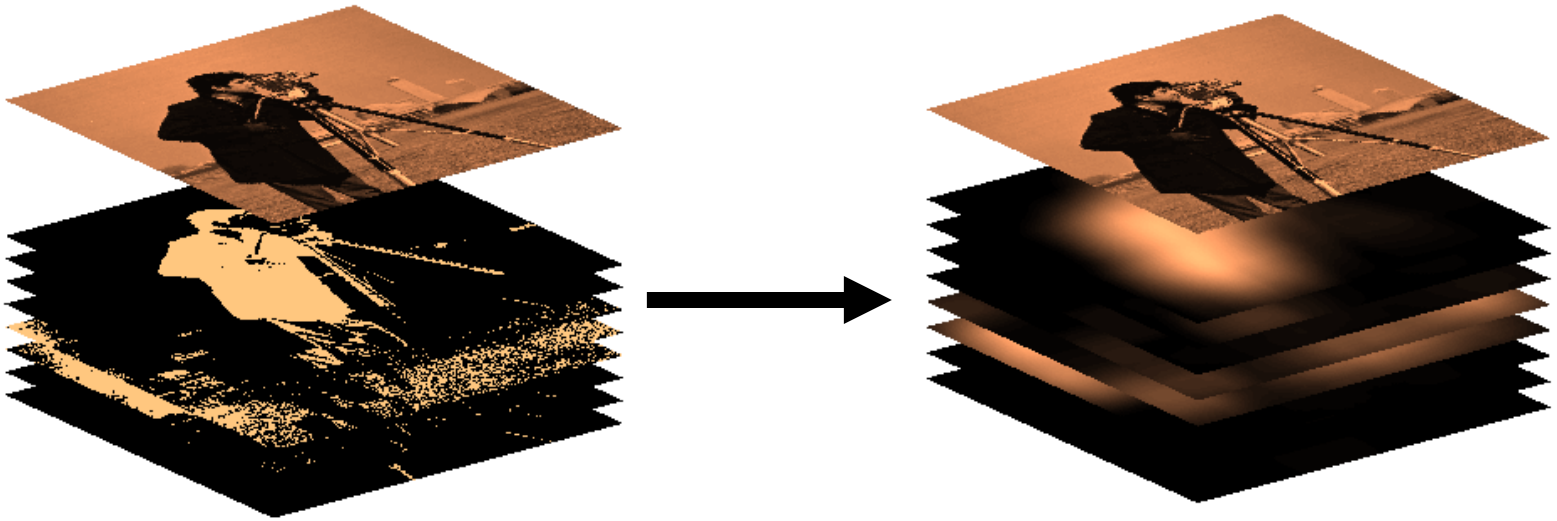
Why?



Spatial Blur: 3d convolution with 2d Gaussian



Spatial Blur: 3d convolution with 2d Gaussian



KEY PROPERTY: doesn't destroy information through averaging

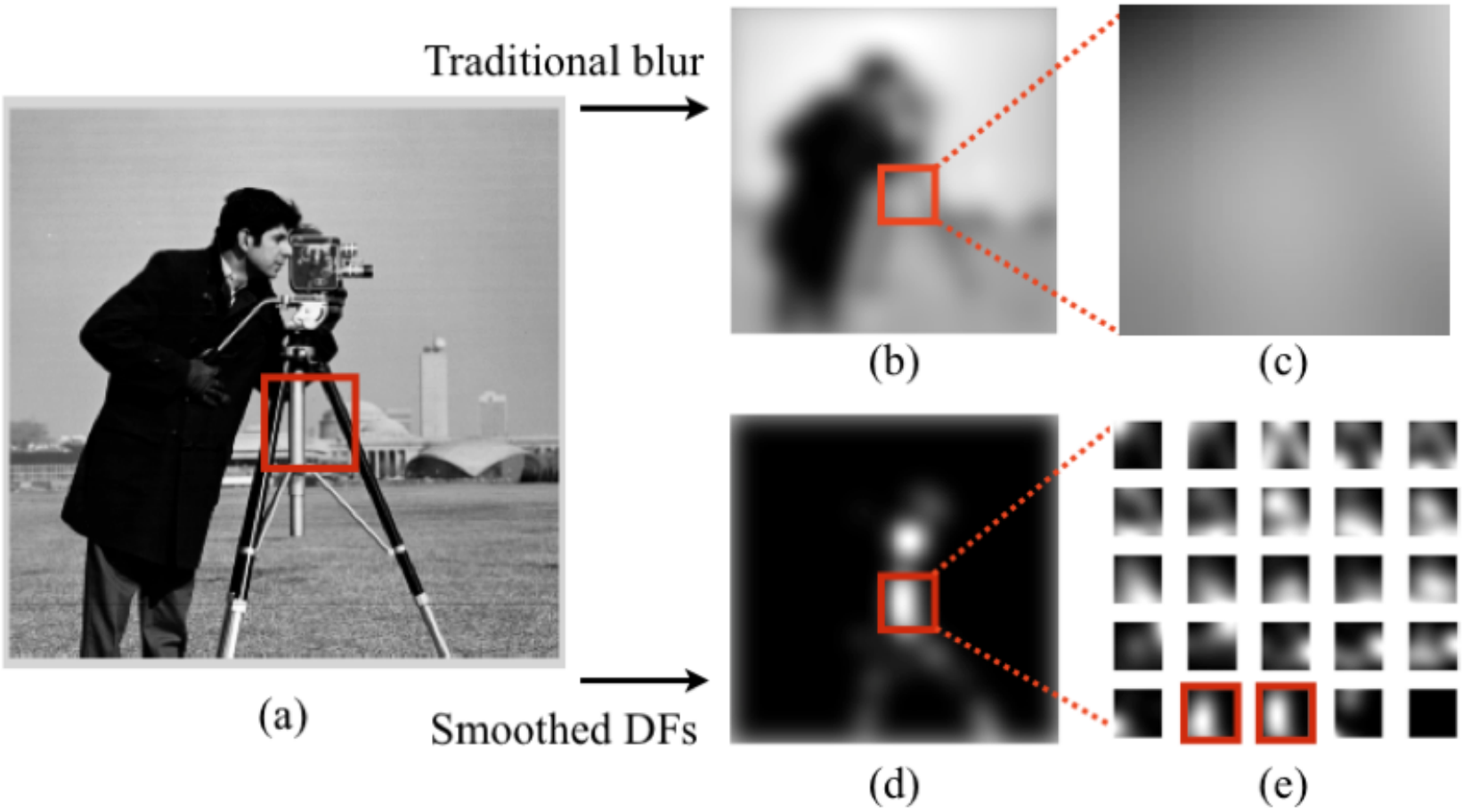
Benefits of congealing without all the images

- Instead of having hundreds of images, just "invent" hundreds of images by perturbing a couple of images.
- SAME as convolving an exploded distribution field with a 2D Gaussian.
- Produces smooth landscape for alignment!

Similar representations

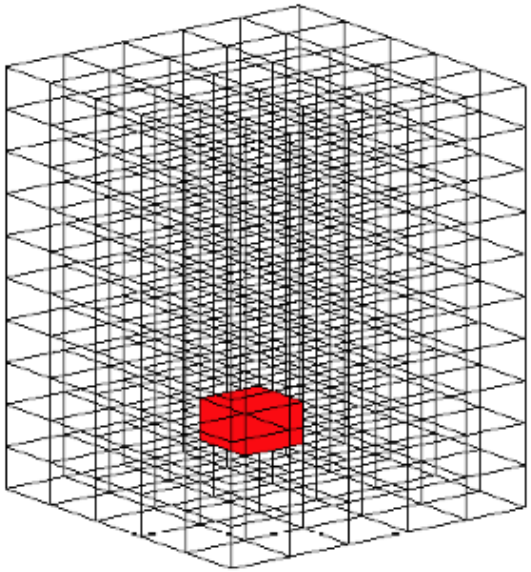
- SIFT (Lowe)
  Generalized Shape Context (Belongie)
  - integrates sparse feature information over blocks
- Geometric blur (Berg)
  - spreads edge information in sparse feature space

Blurring while preserving information

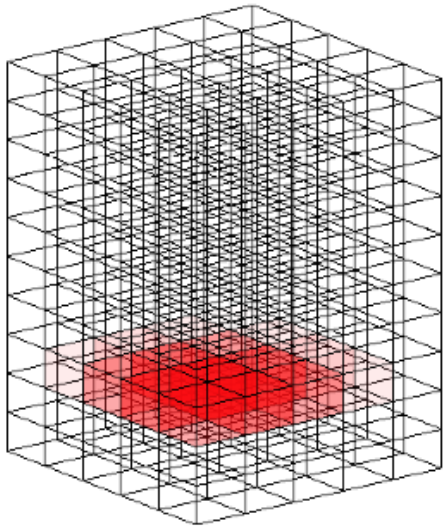


Feature space blur

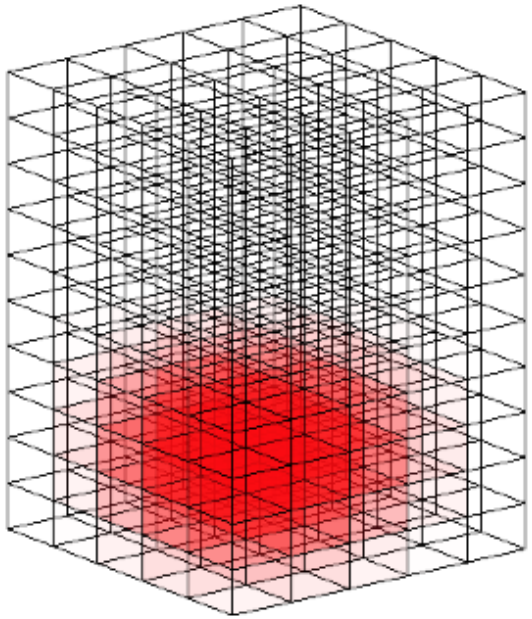
Delta function at one pixel



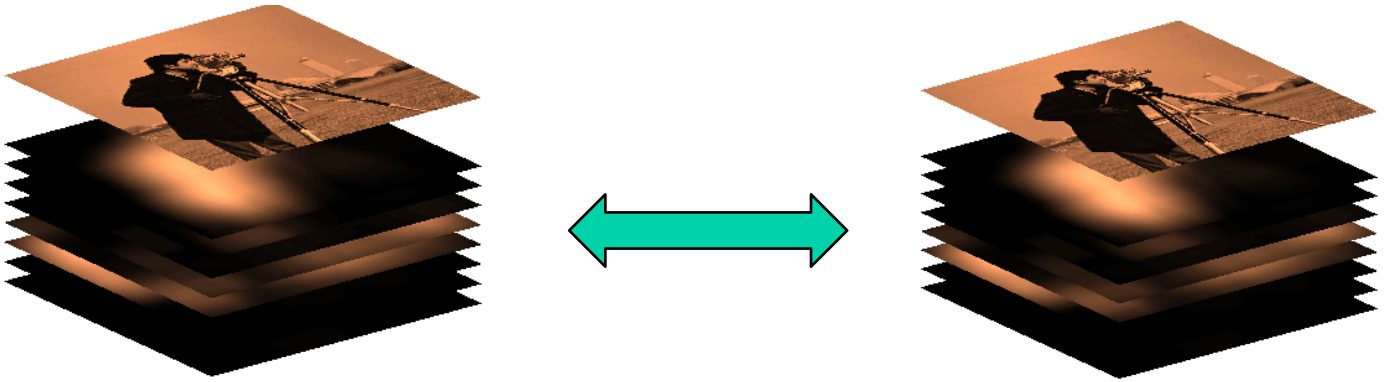
Spatial blur



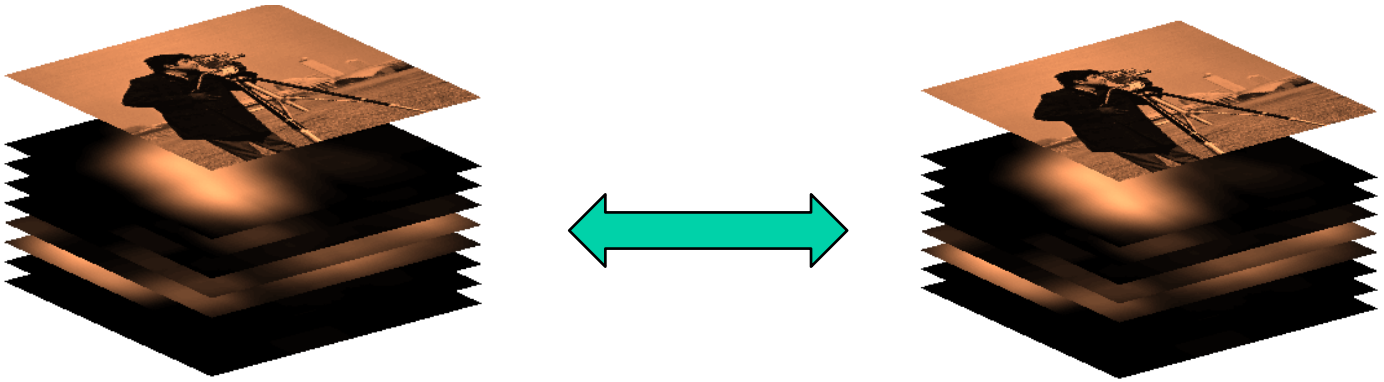
Spatial and feature-space blur



How to compare?



How to compare?

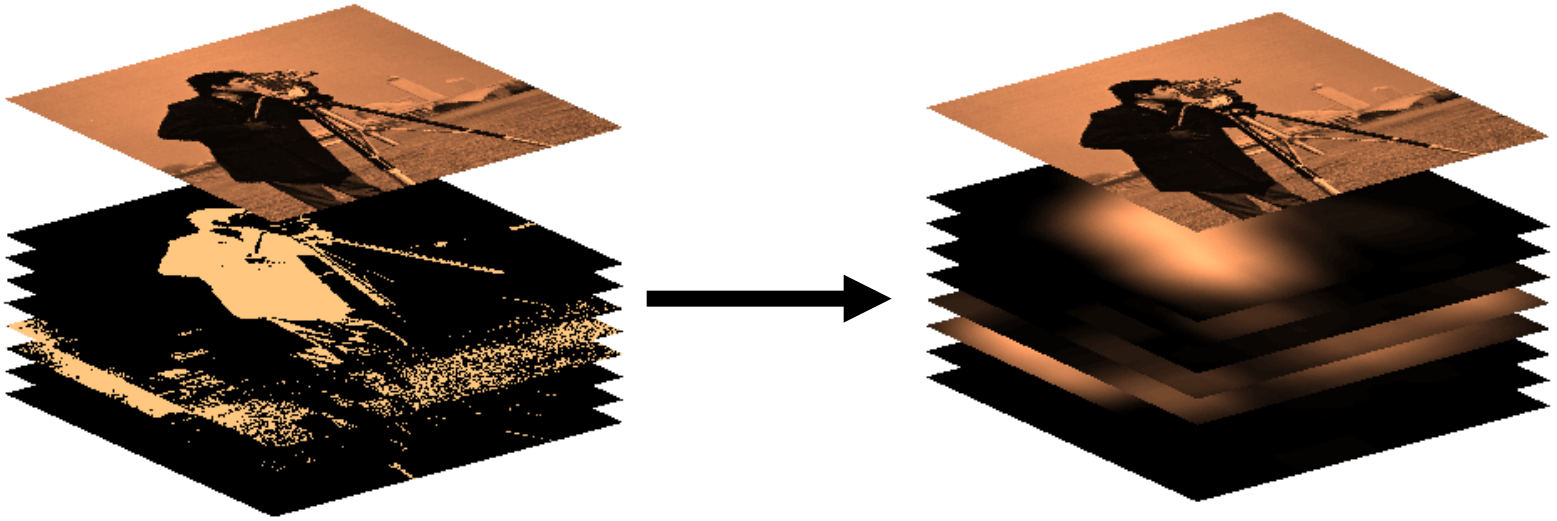


- L1 distance?
- L2 distance?
- KL divergence?

Outline

- Similarity measures in vision—general remarks
- Piece 1: Sensitivity to position in image comparison
  - What's the right histogram bin size?
- Piece 2: Image matching with gradient descent
  - Overcoming problems with traditional blurring approaches using *distribution fields.*
- Putting the pieces together: the sharpening match.
- Some results
  - Basin of attraction studies
  - Tracking experiments

Distribution Fields: Invariance through blurring

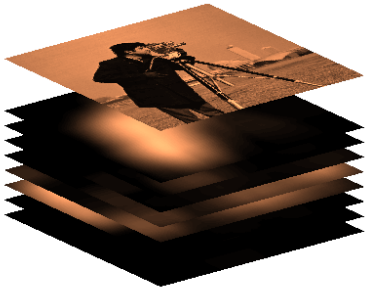


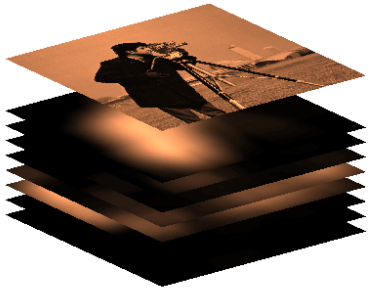
Each pixel location becomes a distribution of the local distribution of brightness values.

The width of the blur kernel determines how wide the neighborhood is.

Comparing Images with Dist. Fields

Given two images I  and J: 

1.  → 

2.  ()

The Likelihood match

$$p(J|I,\sigma) = \prod_{i=1}^{N} d_i(J_i;\sigma),$$

The Likelihood match

ith distribution in distribution field

$$p(J|I,\sigma) = \prod_{i=1}^{N} d_i(J_i;\sigma),$$

ith pixel in image J

The Sharpening Match

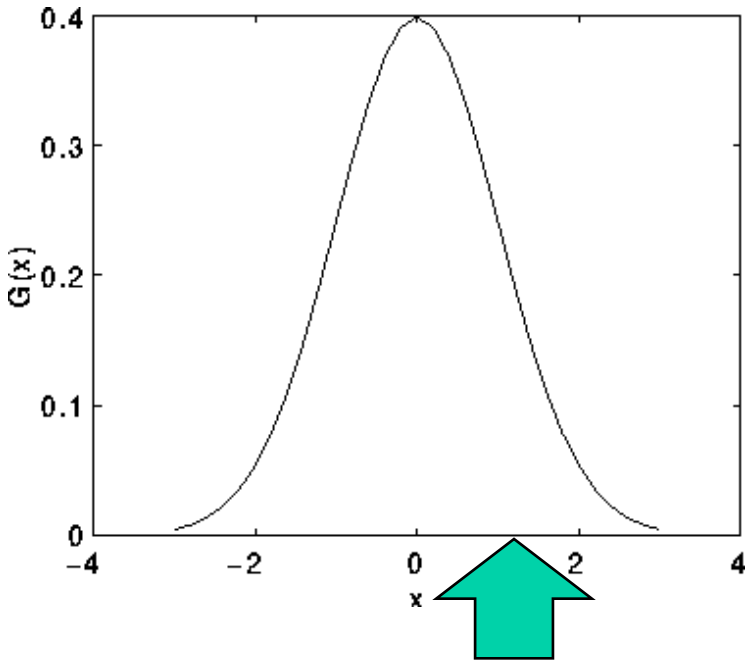
$$\sigma^* = \arg\max_{\sigma} \; p(J|I,\sigma) = \arg\max_{\sigma} \; \prod_{i=1}^{N} d_i(J_i;\sigma)$$

The Sharpening Match

$$\sigma^* = \arg\max_{\sigma} \; p(J|I,\sigma) = \arg\max_{\sigma} \; \prod_{i=1}^{N} d_i(J_i;\sigma)$$

Pop Quiz: What happens when I = J?

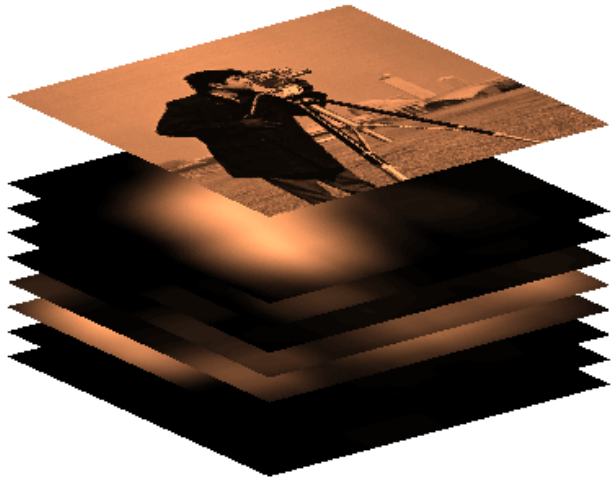
Understanding the sharpening match



What standard deviation maximizes the likelihood of a single point x under a zero-mean Gaussian?

Intuition behind sharpening match

- Increase standard deviation until it matches "average distance" to matching points.



Properties of the sharpening match

- An image has sigma = 0 under its own distribution field.
  - Satisfies property 1 (!!!)
- Probability of an image patch degrades gracefully as it is translated away from best position.
  - Satisfies property 2 (!!!)
- Optimum sigma value gives a very intuitive notion of the quality of the image match.

The likelihood match

- Recall image I and patch J.
- Make a distribution field out of I and evaluate the likelihood of J under the field.

Image I

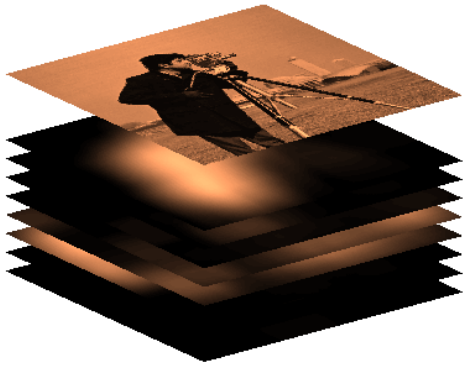
Patch J

What happens in gradient descent

- 1) Build DF with one image
- 2) Expand kernel until patch likelihood is maximized: tends to be a big kernel
- 3) Update position of patch
- 4) Adjust kernel size to match likelihood again
  - Tends to be a smaller kernel
- 5) When you're done, the remaining kernel size gives you the quality of the match!

Intuition behind sharpening match

- Increase standard deviation until it matches "average distance" to matching points.



Properties of the sharpening match

- A patch has probability of 1.0 under its own distribution field.
- Probability of an image patch degrades gracefully as it is translated away from best position.
- Optimum "sigma" value gives a very intuitive notion of the quality of the image match.

Outline

- Similarity measures in vision—general remarks
- Piece 1: Sensitivity to position in image comparison
  - What's the right histogram bin size?
- Piece 2: Image matching with gradient descent
  - Overcoming problems with traditional blurring approaches using *distribution fields.*
- Putting the pieces together:
  the sharpening match.
- Some results
  - Basin of attraction studies
  - Tracking experiments

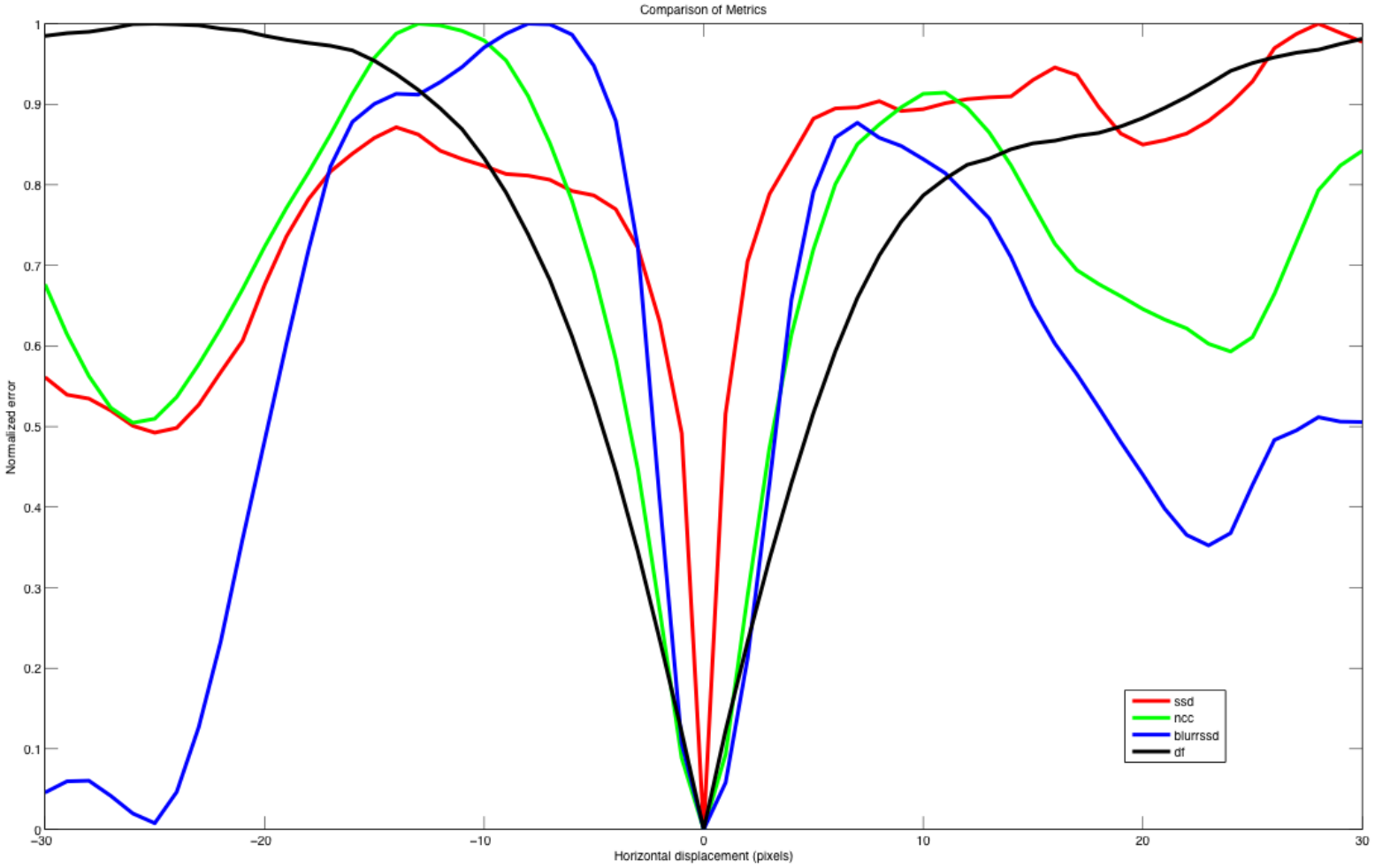
Basin of attraction studies

Basin of attraction studies

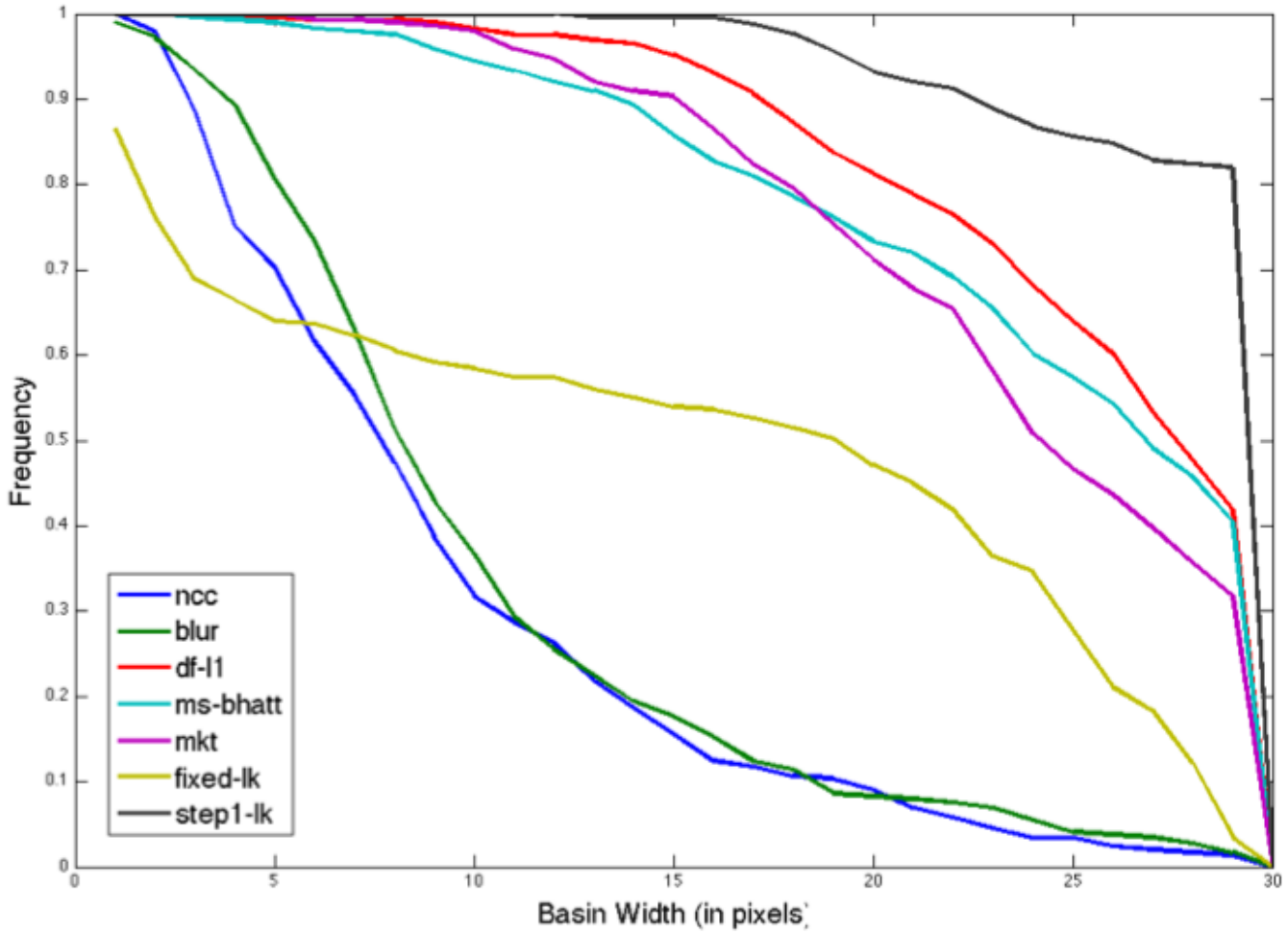
Basin of attraction studies

Basin of attraction studies

Basin of attraction studies



Basin of attraction results (CVPR 2012)

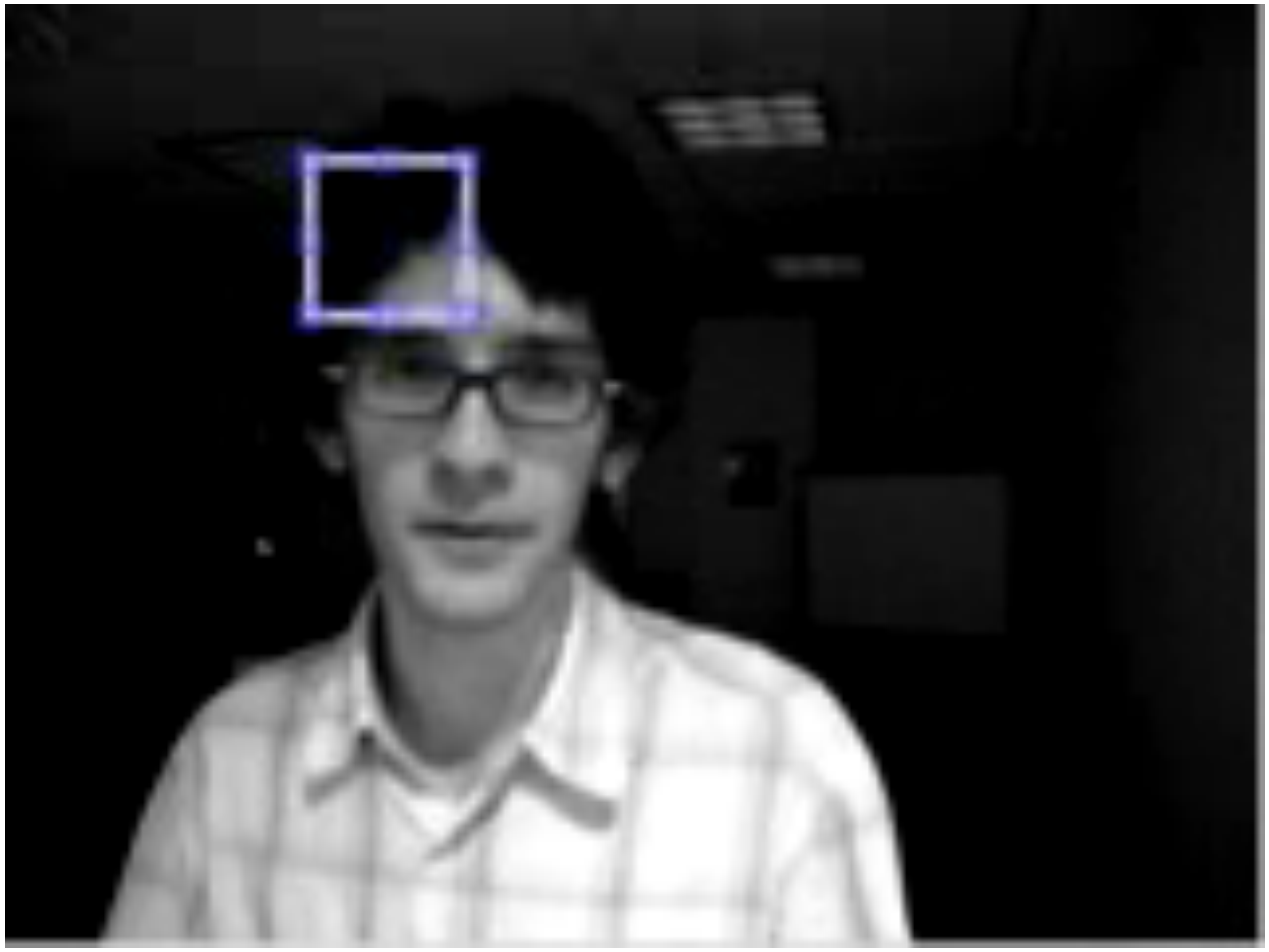


Lukas-Kanade etc..

Distribution Field alignment

Tracking results

- State of the art results on tracking with standard sequences
  - Very simple code
  - Trivial motion model
  - Simple memory model



It's not perfect…

Conclusion

- The sharpening match addresses
  - The difficulties in developing a matching function which can tolerate positional differences
  - The difficulties of doing gradient descent alignment

Related work

- Mixture of Gaussian backgrounding (Stauffer...)
- Shape contexts (Belongie and Malik)
- Congealing (me)
- Bilateral filter
- SIFT (Lowe), HOG (Dalal and Triggs)
- Geometric Blur (Berg)
- Rectified flow techniques (Efros, Mori)
- Mean-shift tracking
- Kernel tracking
- and many others...

Thanks!