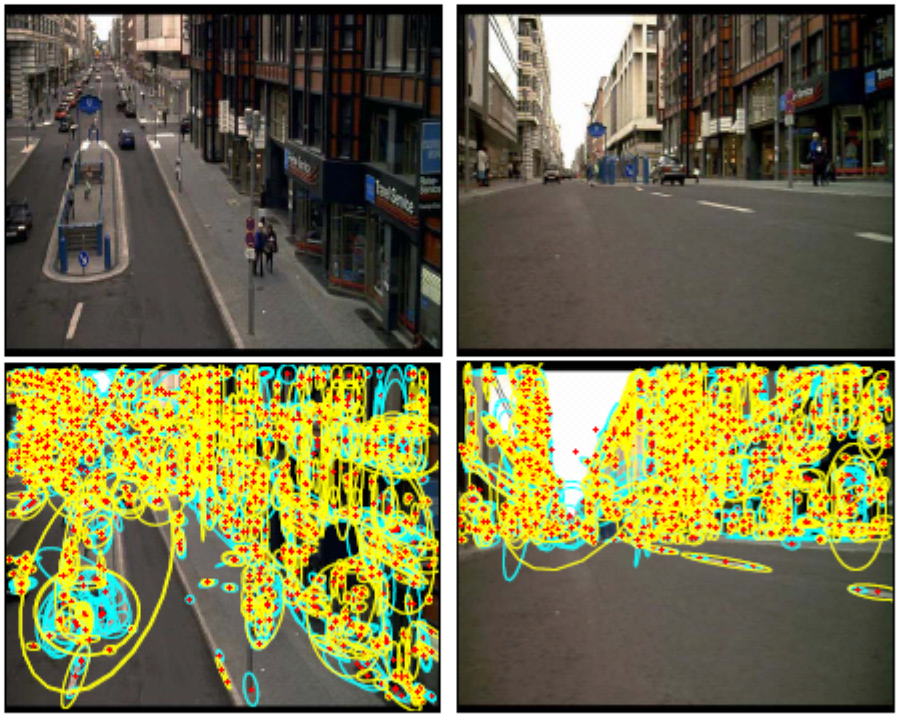
Sampling Strategies for Object Classification

Gautam Muralidhar

Reference papers

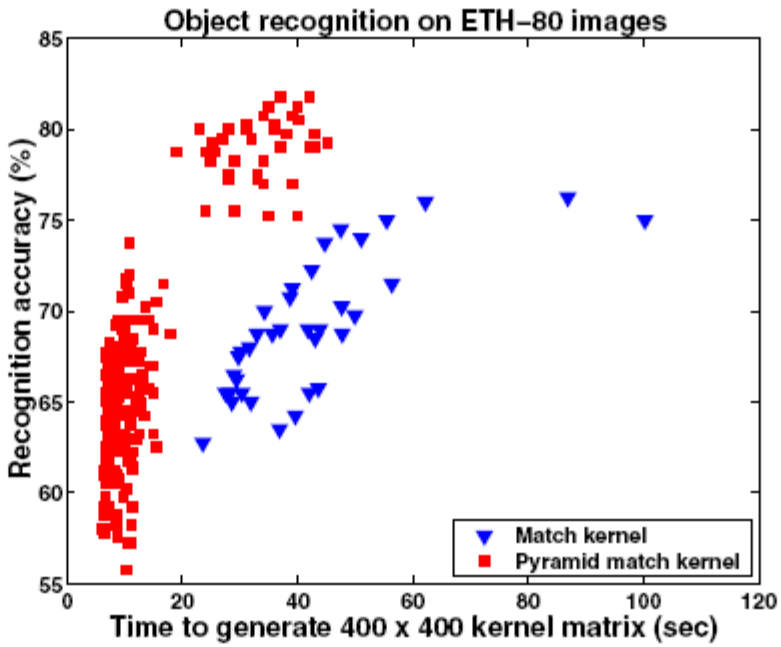
- The Pyramid Match Kernel – Grauman and Darrell
- Approximated Correspondences in High Dimensions – Grauman and Darrell
- Video Google – Sivic and Zisserman
- Scale and Affine Interest Point Detectors – Mikolajczyk and Schmid
- Robust Wide Baseline Stereo from Maximally Stable Extremal Regions – Matas *et al*
- Sampling Strategies for Bag of Features Image Classification – Nowak, Jurie and Triggs
- Object Recognition from Local Scale Invariant Features - Lowe

Motivation



In Sivic & Zisserman's Video Google paper, two operators are used to capture complementary region types (blobs, corners), and thereby make a fuller vocabulary.

Further, recent work on Sampling Strategies for Bag of Features Image Classification suggest that classification performance is best with random sampling than with the use of sophisticated multi-scale interest operators.



In Grauman & Darrell's Pyramid Match paper, we see that generating more features per image yields better classification accuracy.

Slide borrowed from K. Grauman

Main Goals

- The goal of my study was to explore the effect of various interest point operators and uniform dense sampling on the classification performance.

- The hypothesis was that dense uniform sampling of the image space results in better classification than interest point operators.

- The intuition behind this being more spatial coverage provides semantic information that can be utilized for better decision making.

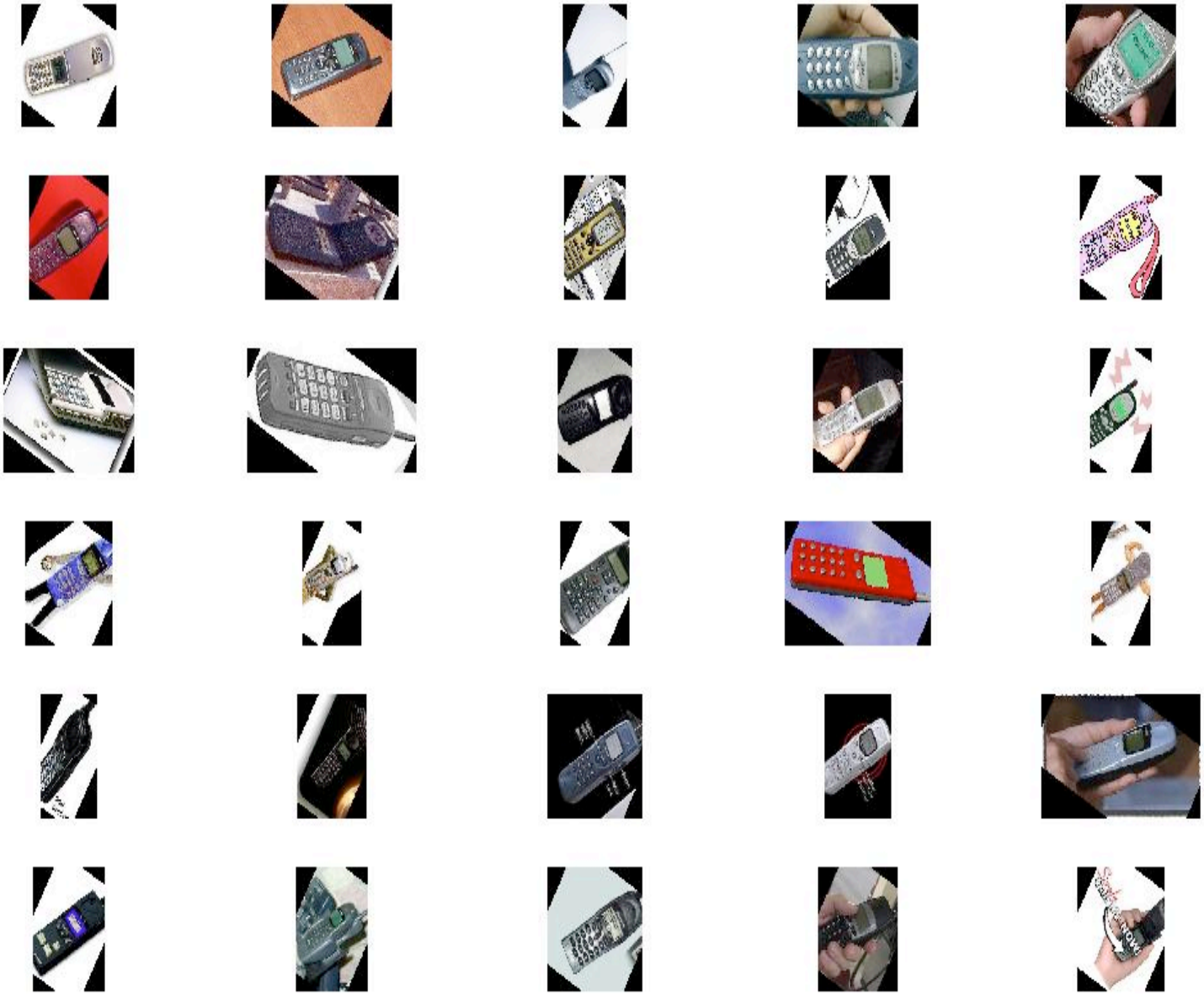
Dataset

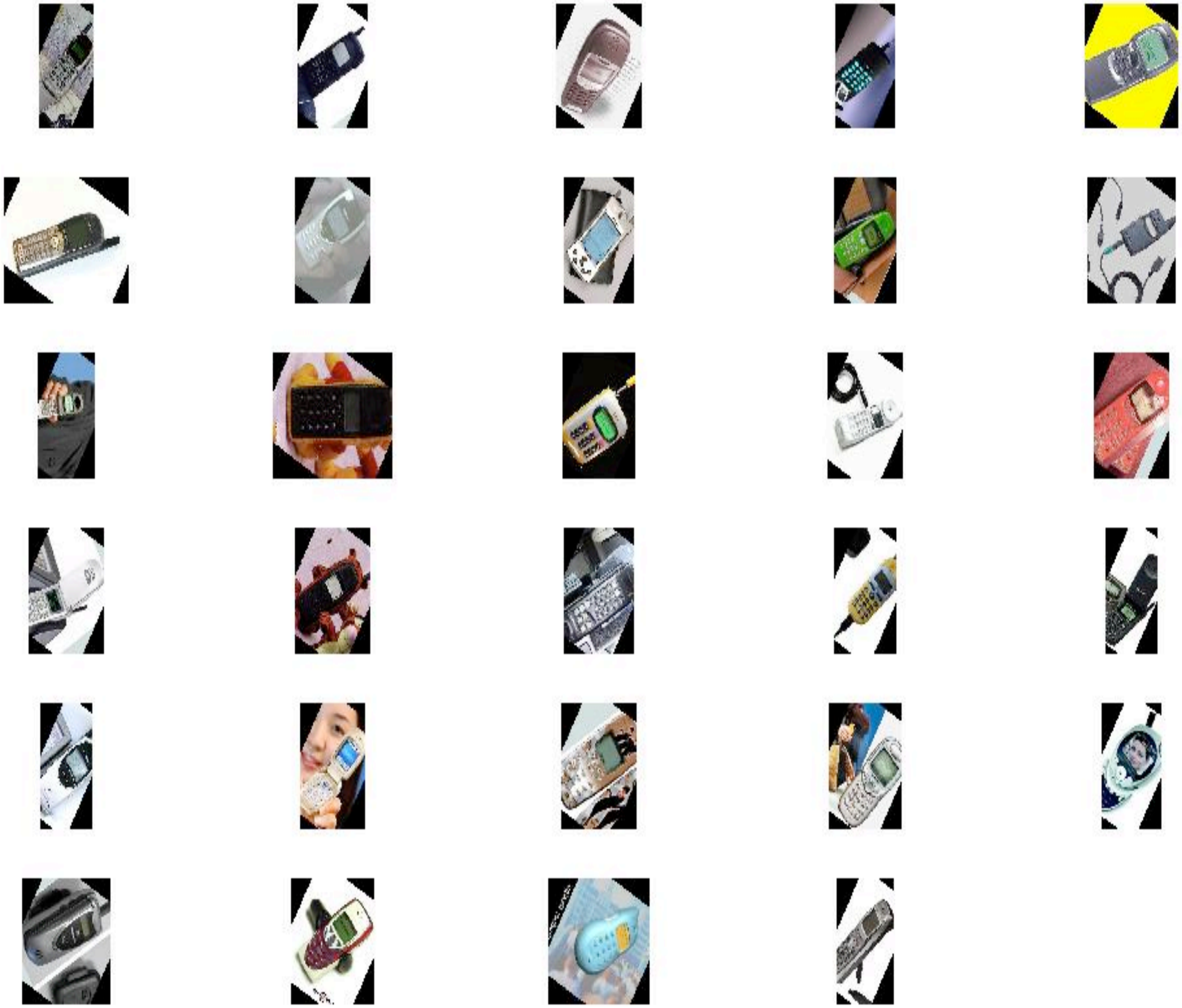
- Caltech 101 – dataset - http://www.vision.caltech.edu/Image_Datasets/Caltech101/

- This has a total of 101 object categories with 30 to 800 images under each category.

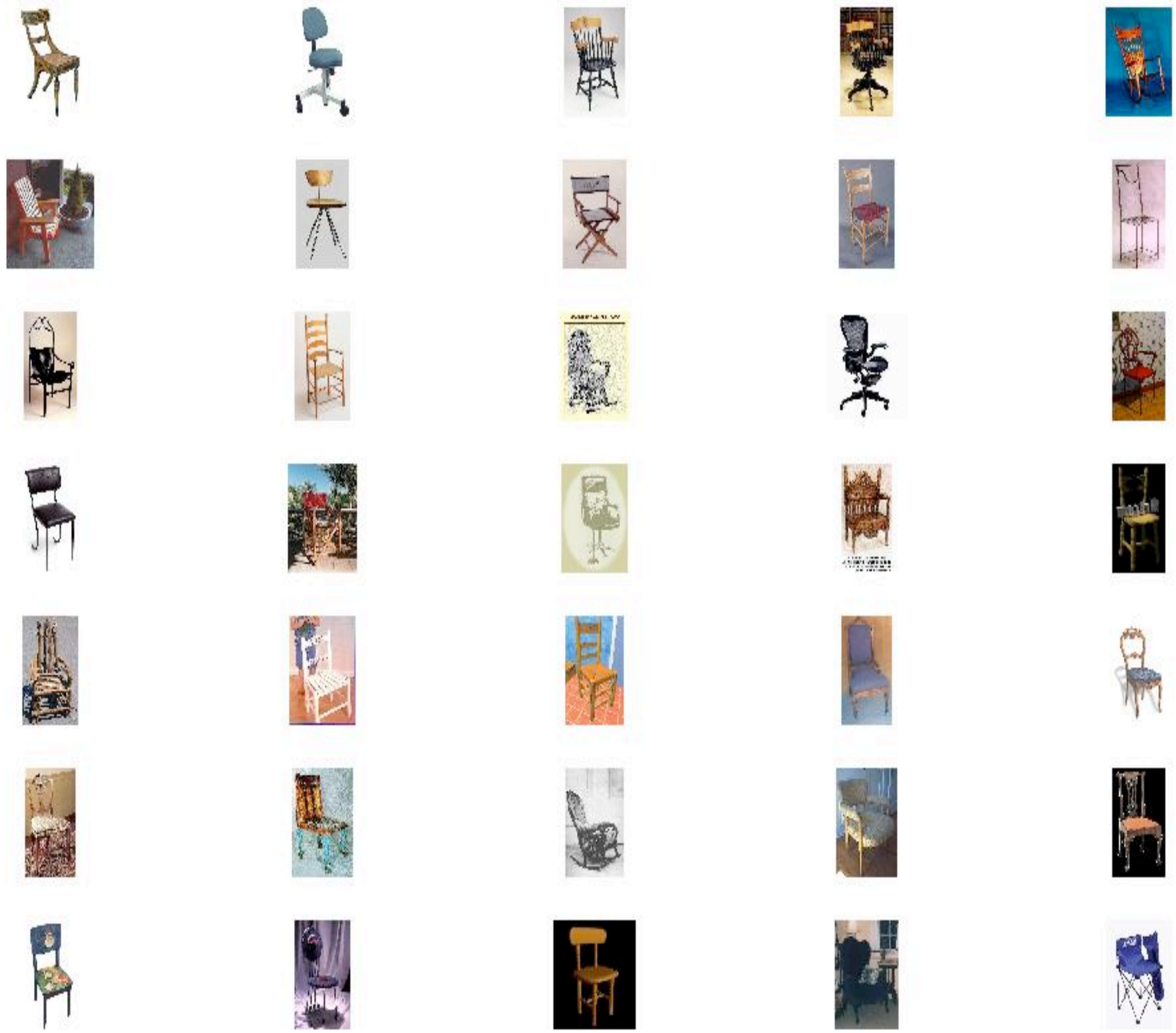
- 5 categories were used in this study – Cell phone, Chair, Lobster, Panda and Pizza to give a total of 253 images.

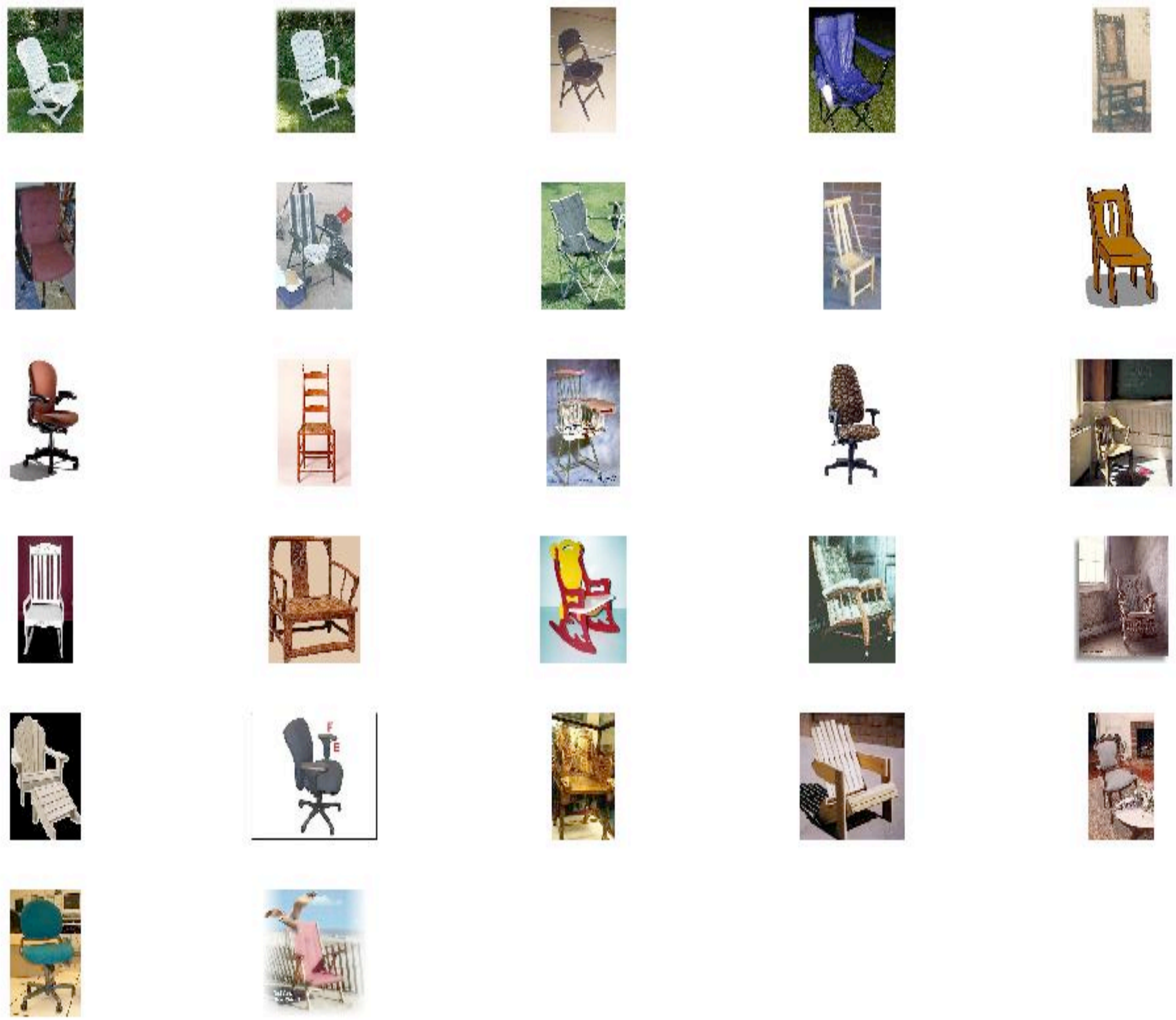
Cell phone– 59 Images



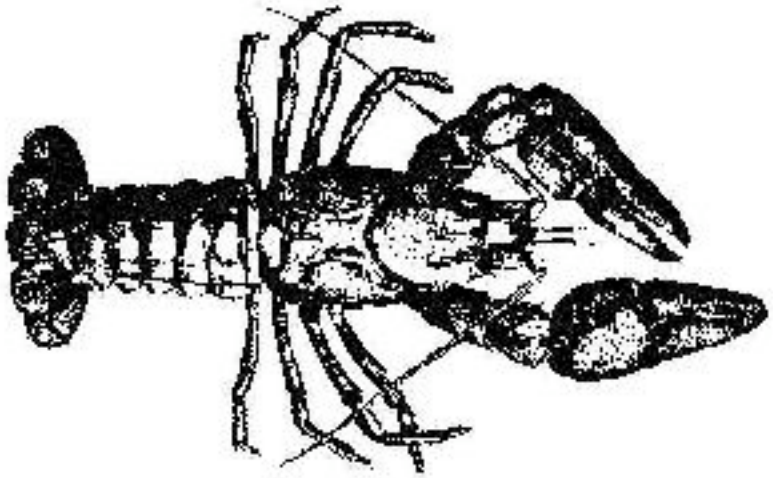


Chair – 62 Images





Lobster – 41 images







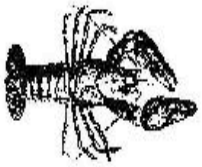













OF SUPER IDEAS





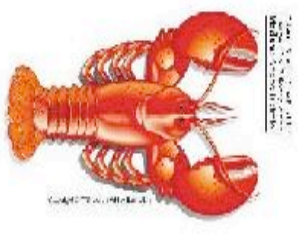






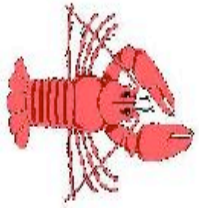




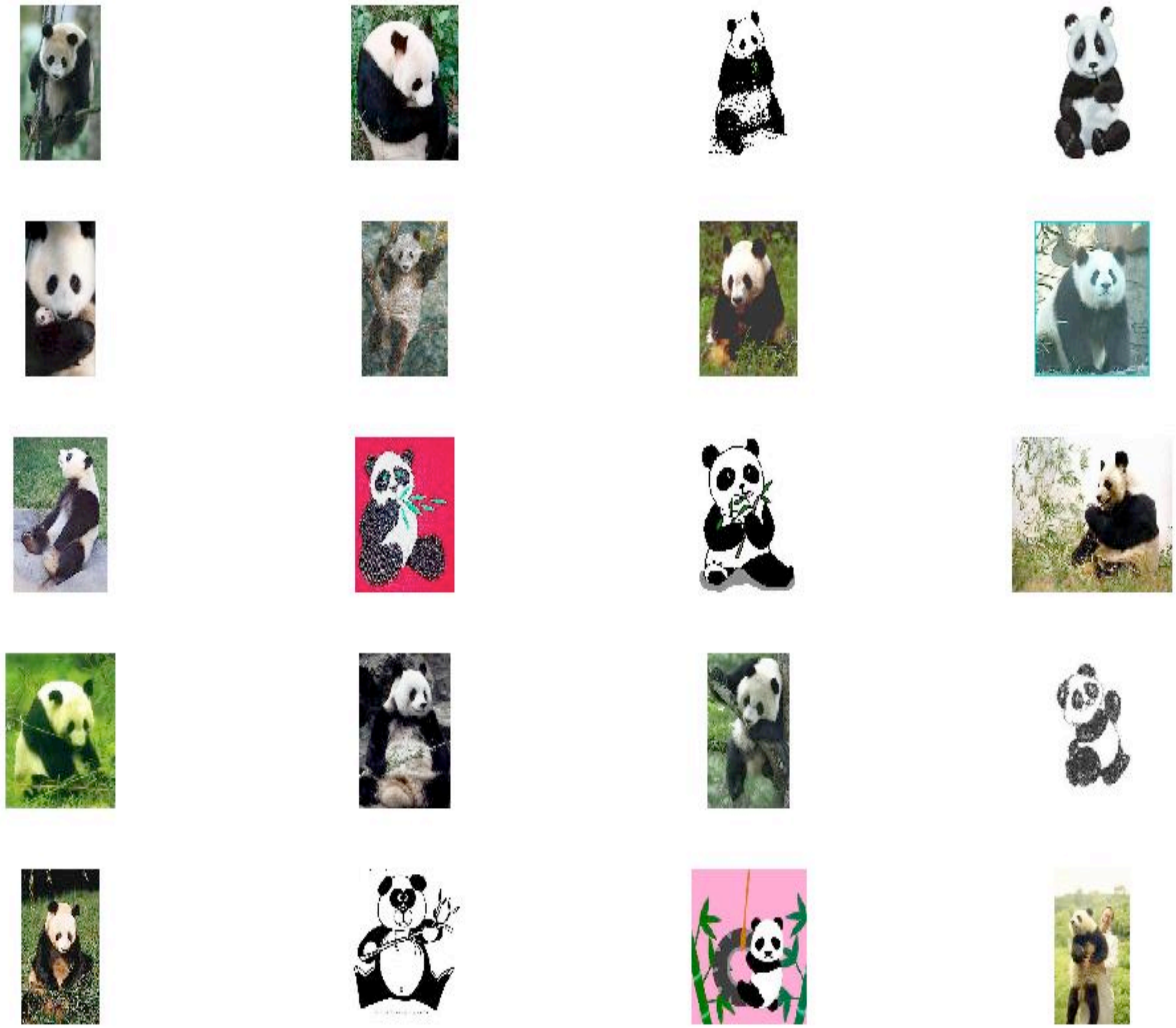



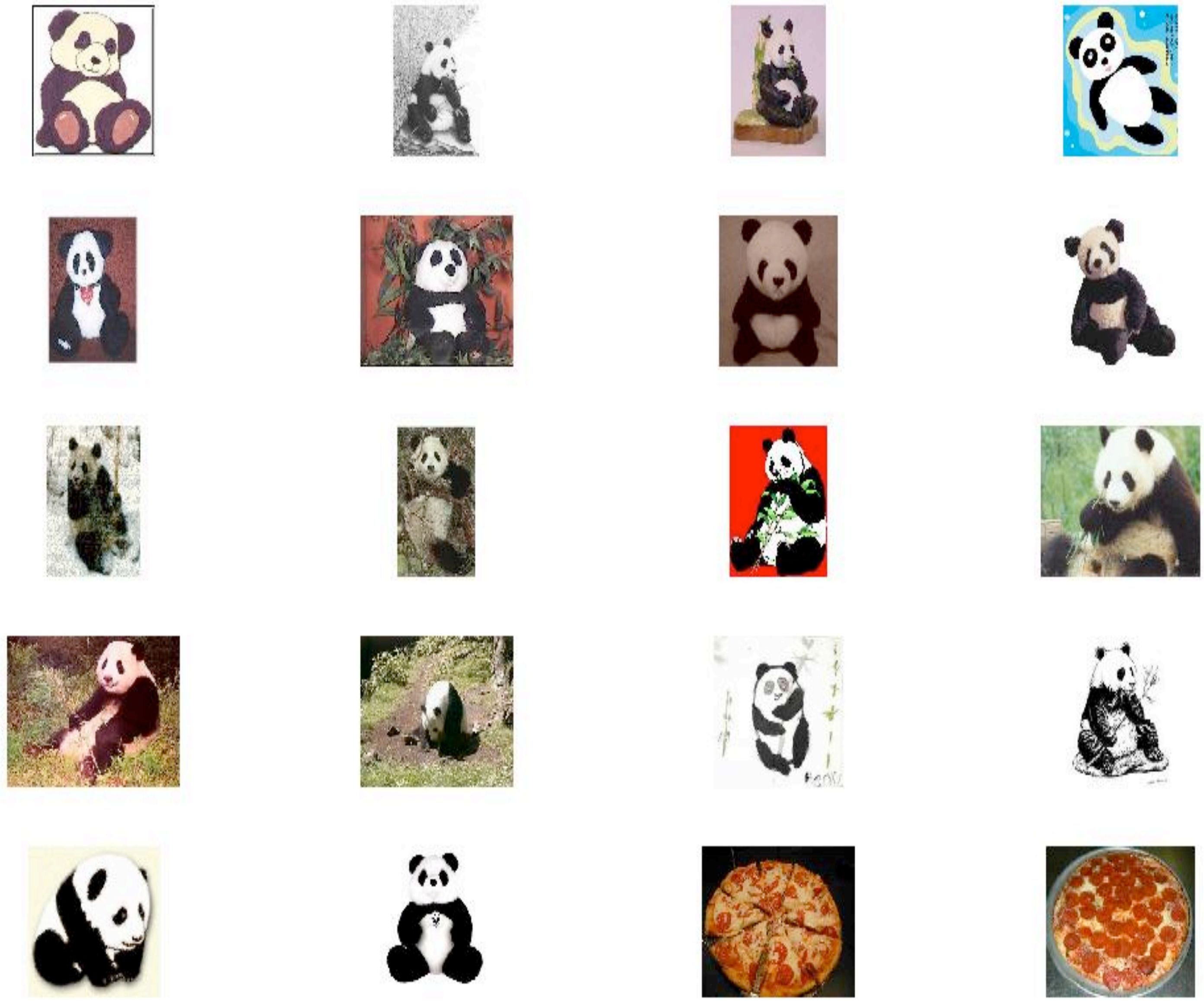




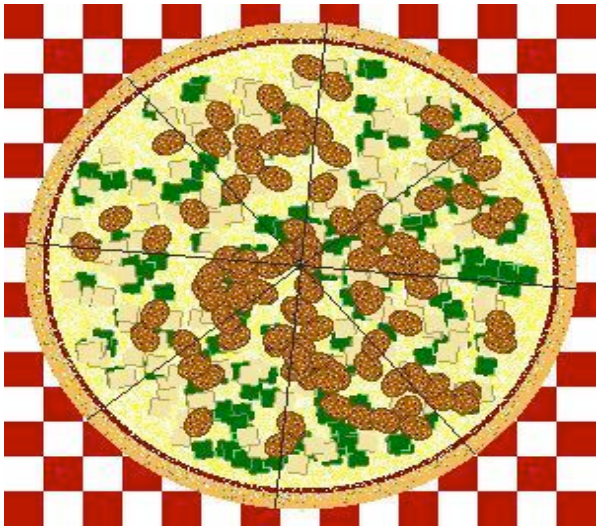


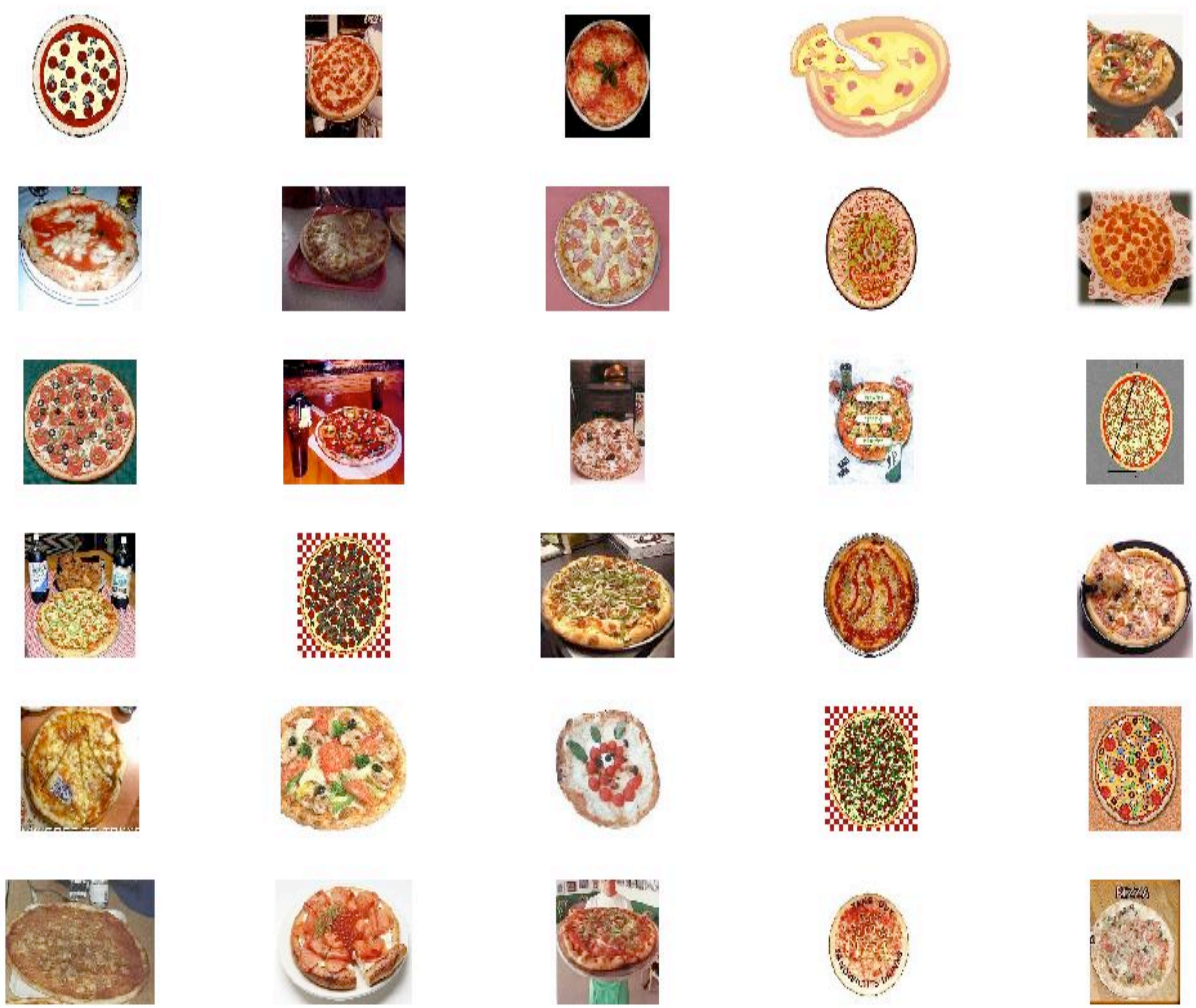
Panda – 38 Images

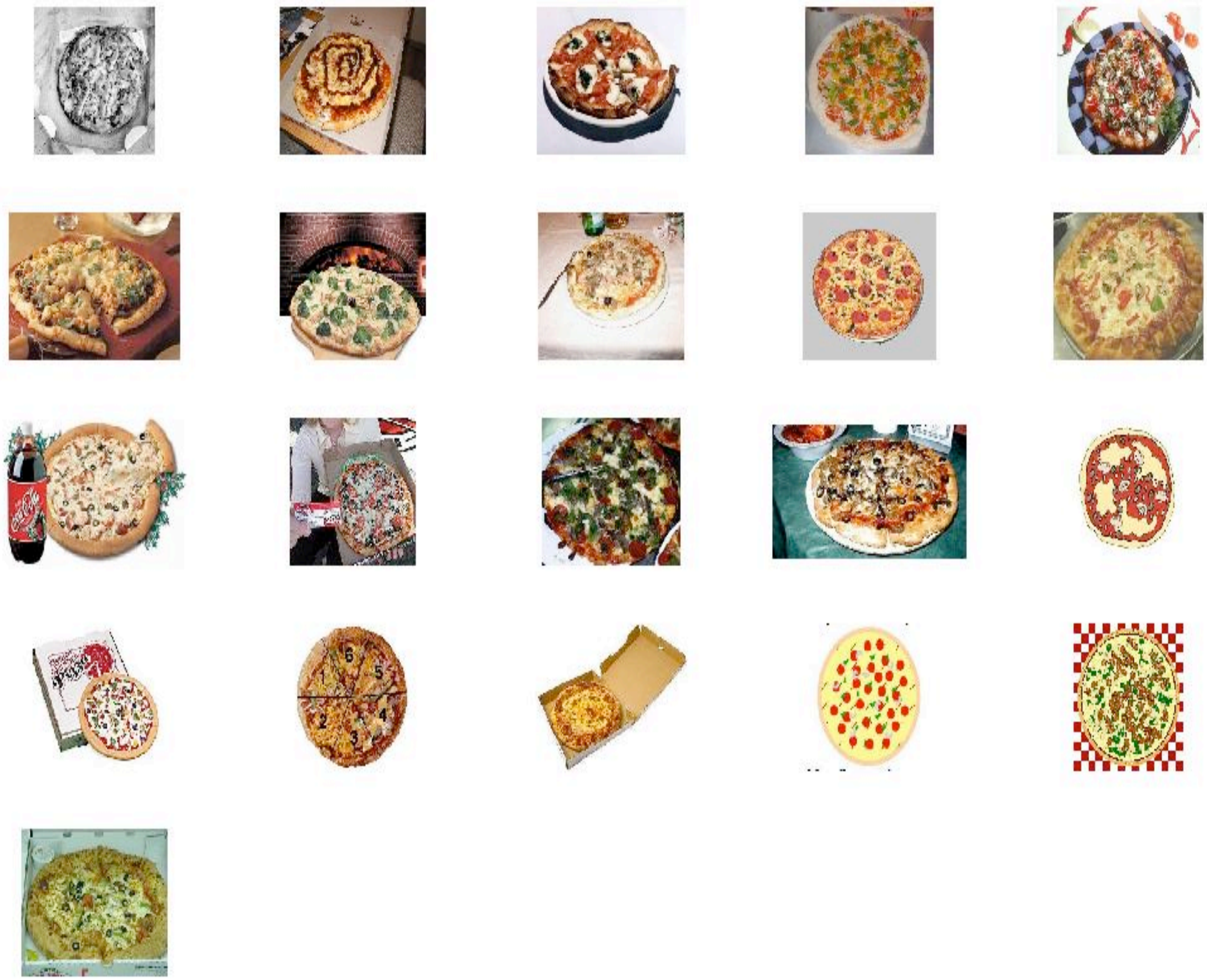




Pizza – 53 images







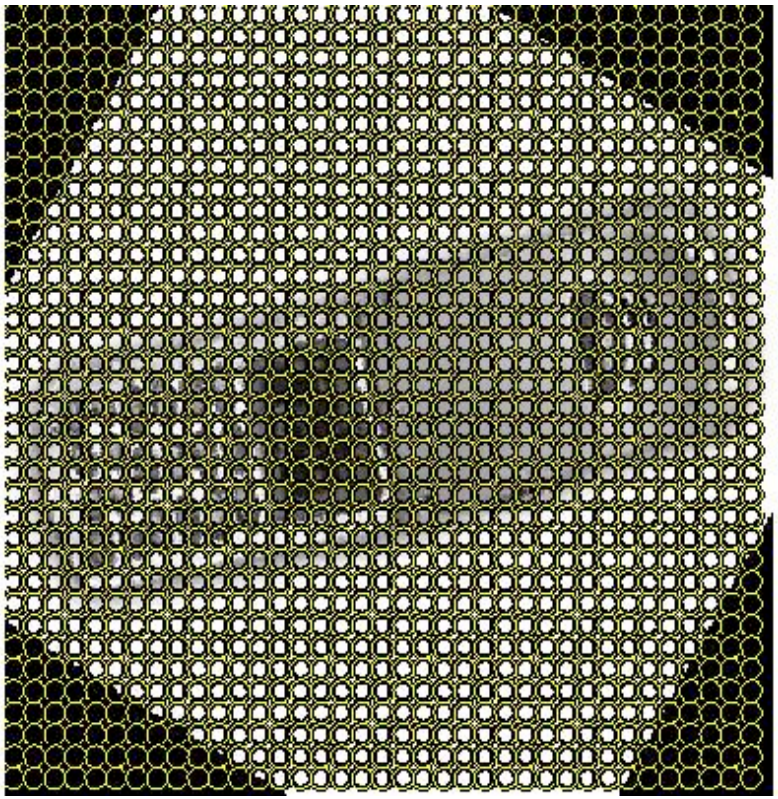
Experiments

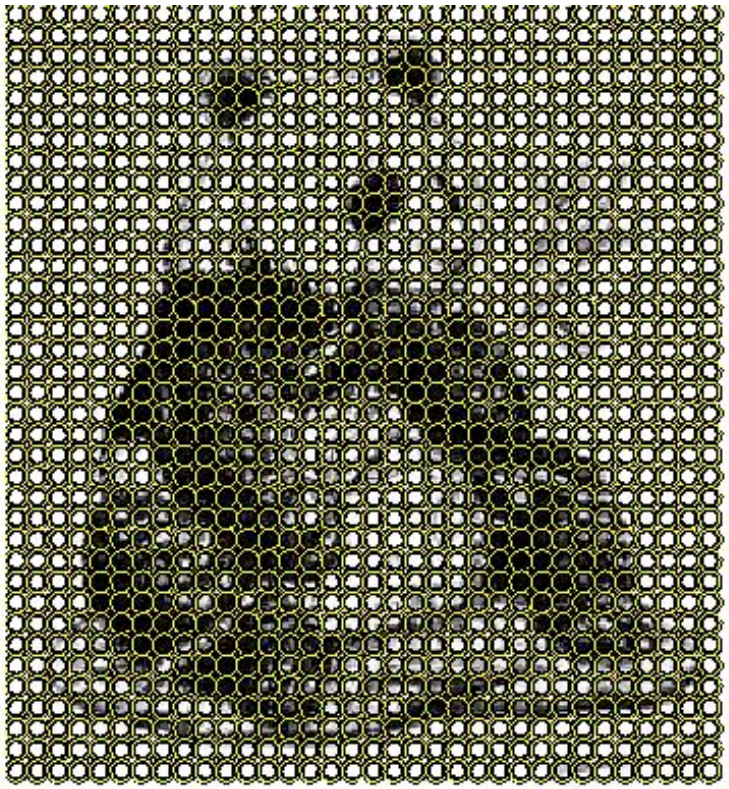
- Dense uniform sampling of image space – vertical and horizontal pixel spacing – 8 pixels.

- Harris affine interest points.

- Combination of Harris Affine and Blob based interest point detector (MSER).

Dense Uniform Sampling





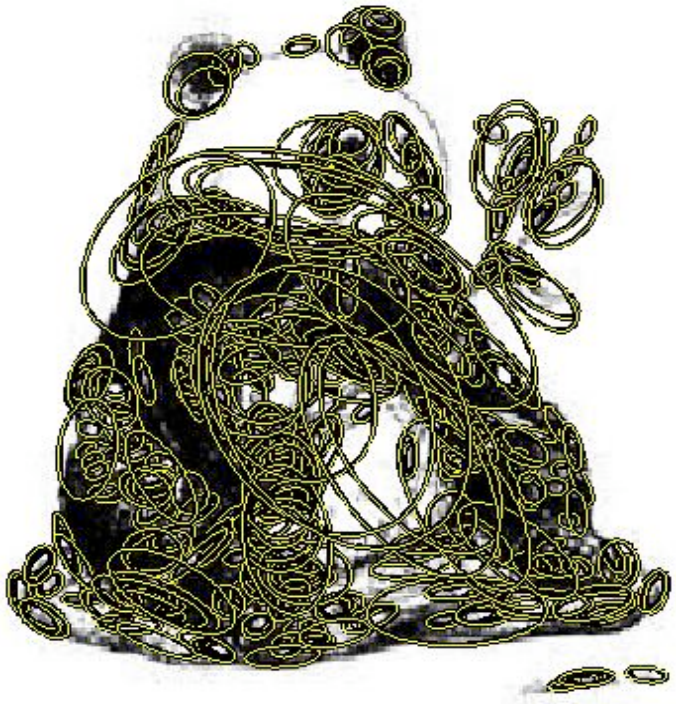
Horizontal and Vertical Pixel spacing – 8 pixels

Harris Affine Interest Point Detector

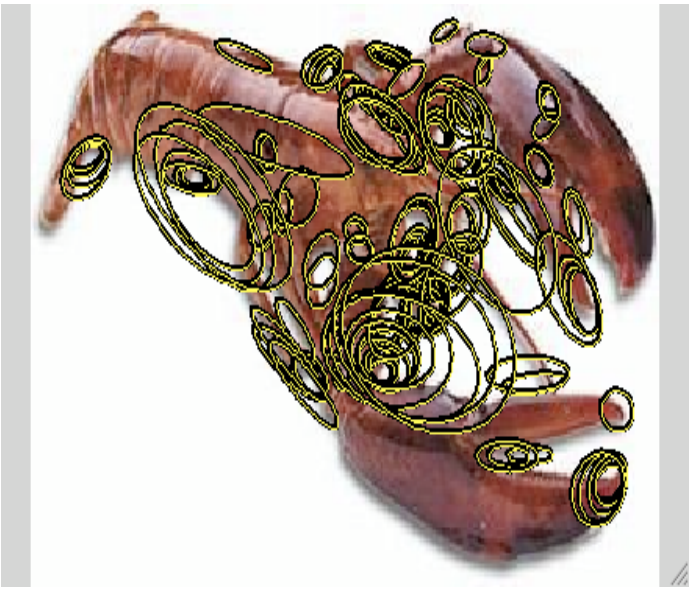
- Proposed by Mikolajczyk and Schmid.
- Adapts the Harris detector proposed by Harris and Stephens (1988) for Scale and Affine invariance.
- The Harris detector is regarded as an ‘edge’ and ‘corner’ detector – detects points in images where intensity changes exist along multiple directions.
- Scale and Affine invariance is achieved via LOG extrema detection at Harris interest points in scale-space followed by shape adaptation.

Harris Affine Detections

- Focus on regions of curvature (corner regions)

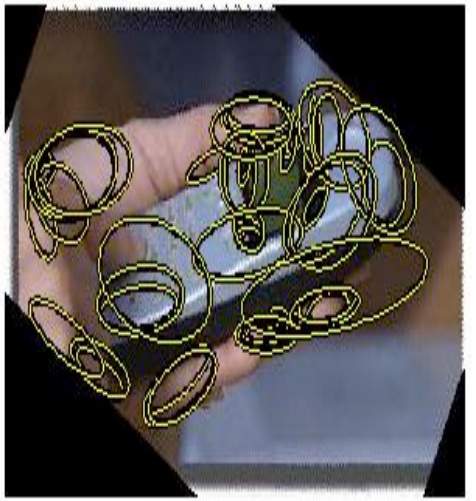
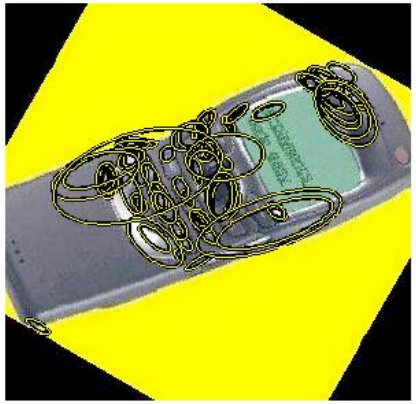


Harris Affine Detections



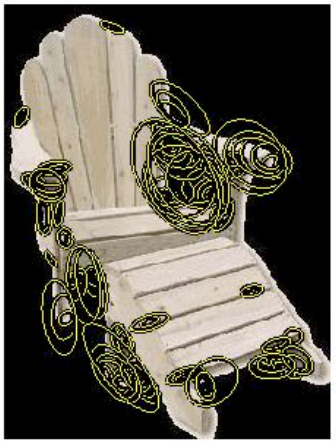
Commonality in Harris Affine Detections

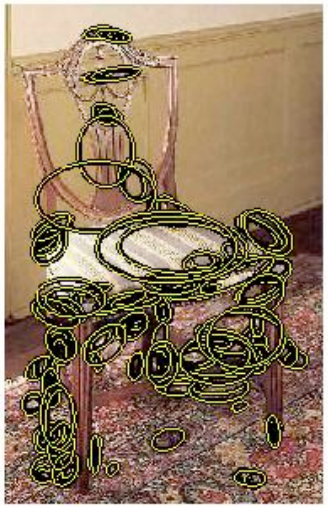


- Cell phone buttons, display in some cases, human hand!

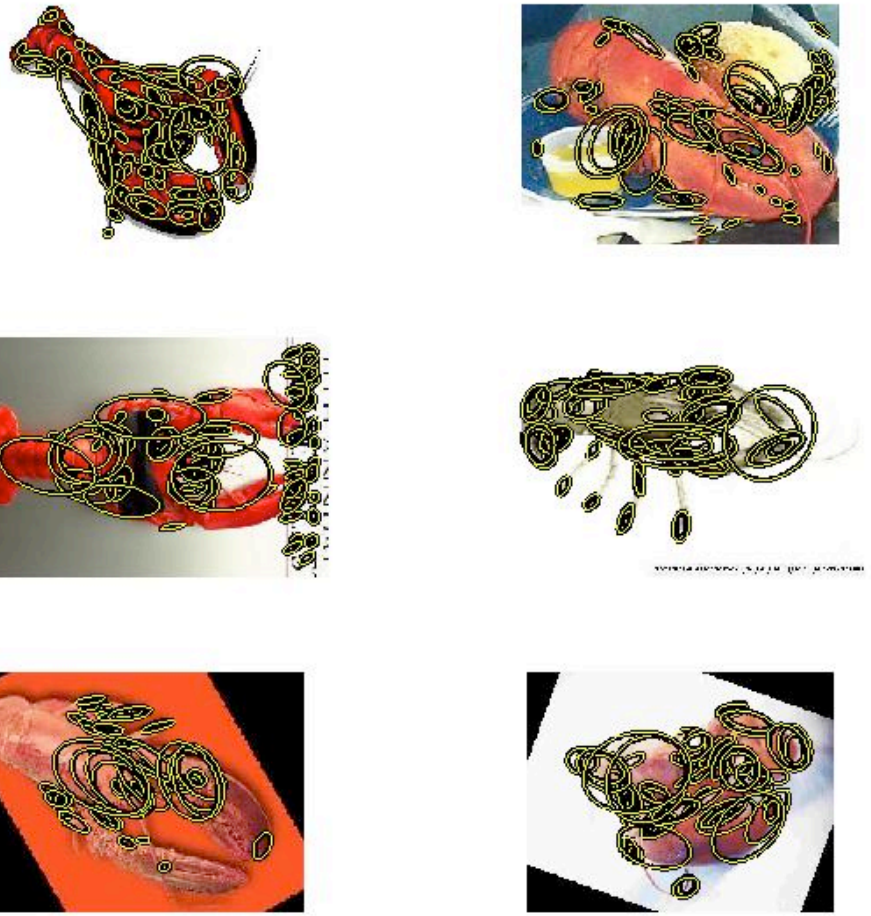
Commonality in Harris Affine Detections



- Corner between legs and seating area, back rest ….

Commonality in Harris Affine Detections



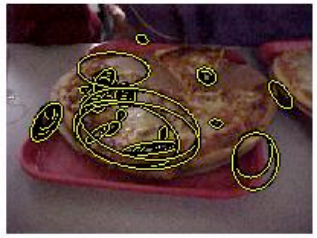
Commonality in Harris Affine Detections

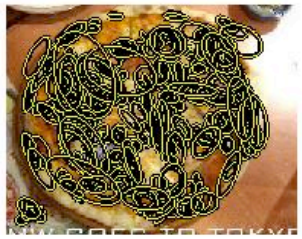





•Ears, nose, eyes, paws…

Commonality in Harris Affine Detections

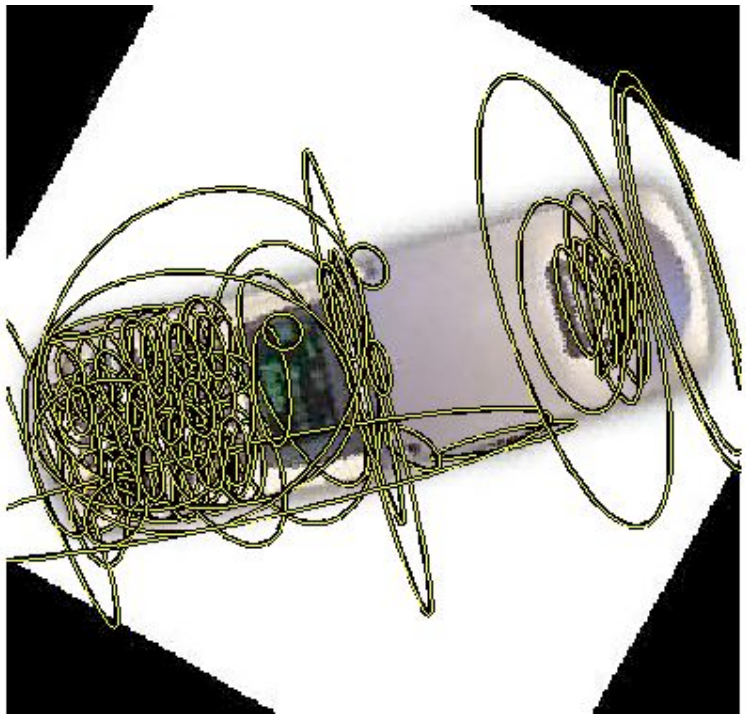




•Pizza toppings!

Maximally stable external regions (MSER)

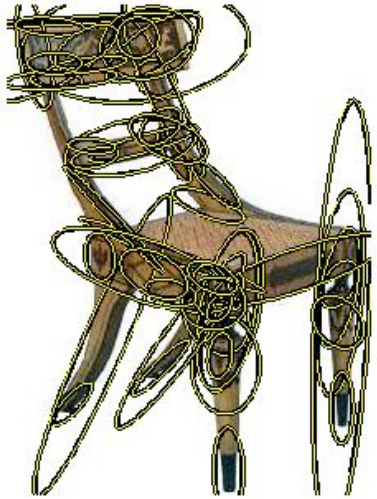
- Proposed by Matas *et al* to find correspondences between two different view points of the same image.
- The basic idea is to threshold the image I with intensity threshold I_O
- For each threshold, extract connect components that are called “Extremal Regions”.
- Extract the maximally stable extremal regions by finding the regions whose support is nearly the same over a range of thresholds.
- MSER provides invariance to affine transformation of image intensities and multi-scale detection without smoothing as both large and fine structures are detected.

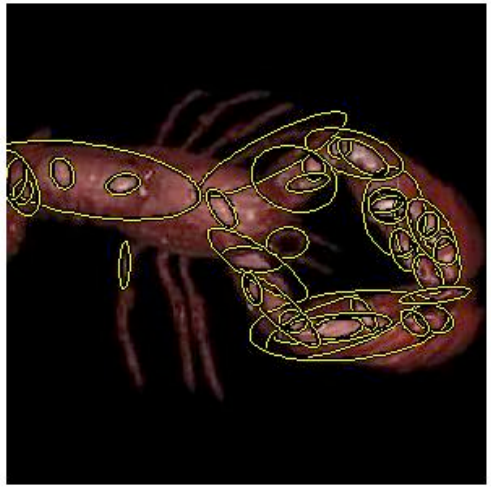
MSER detections

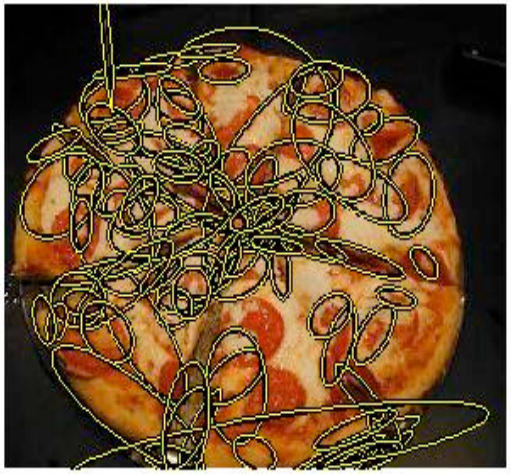


•MSER detection regions approximated as ellipses.
•The Panda is a good example for it clearly shows the ‘blob’ based detections around the ears and the eyes- blobs of high contrast wrt surrounding.

MSER Detections

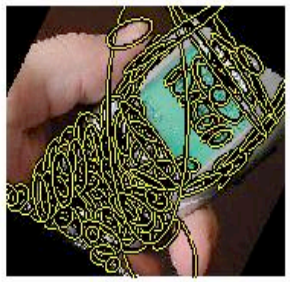






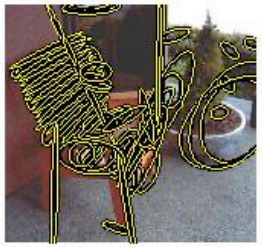
Its clear on the lobster that blobs of high contrast are picked out

Commonality in MSER Detections



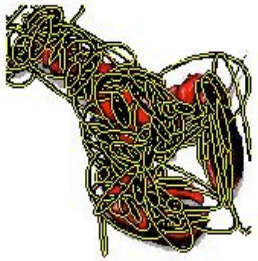



Commonality in MSER Detections






Commonality in MSER Detections



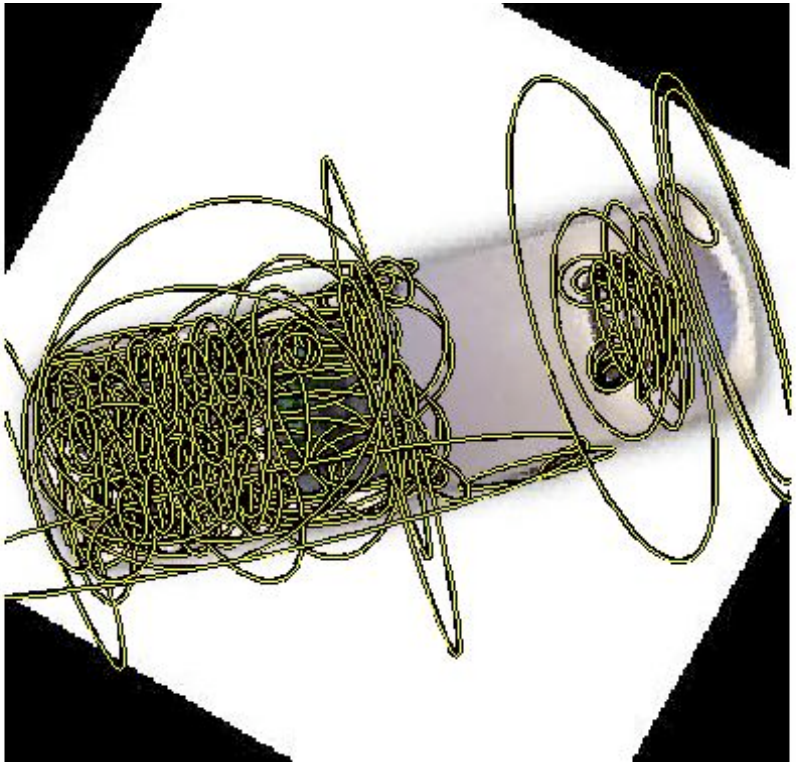
Commonality in MSER Detections

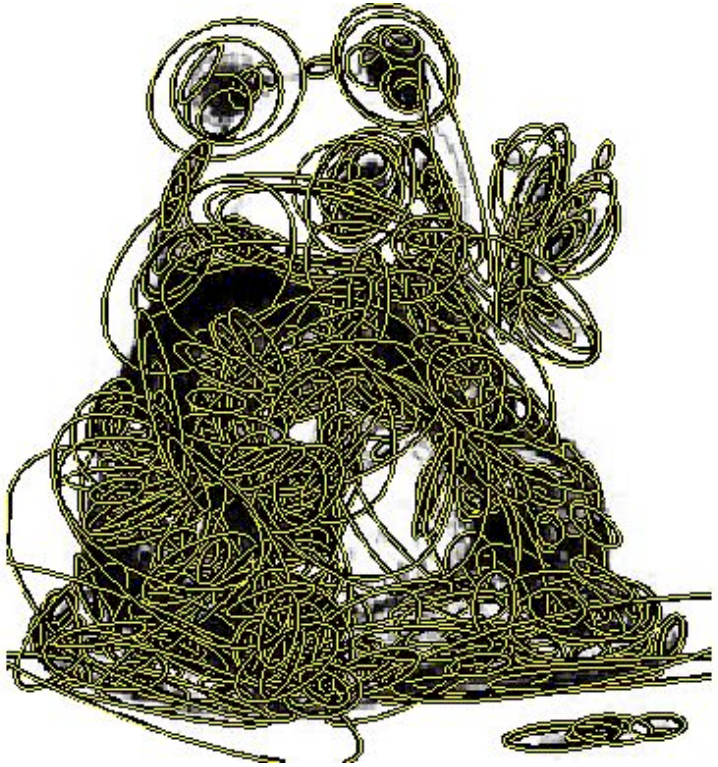
Commonality in MSER Detections





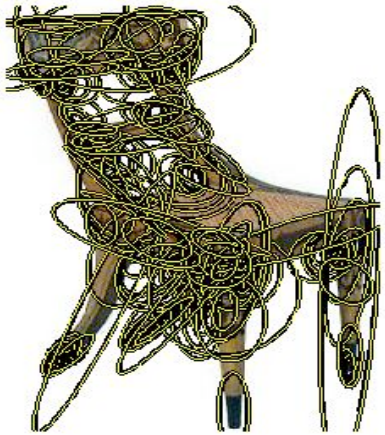

Harris + MSER combined detections

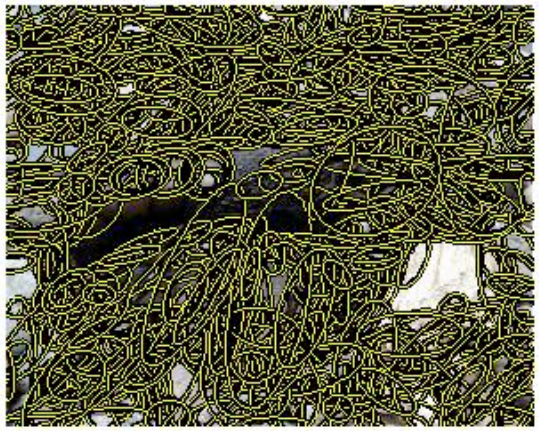




Complementary regions of an image are detected – This point was noted in the video Google paper too

Harris + MSER combined detections





•Dense coverage when compared to just Harris and MSER

Methods

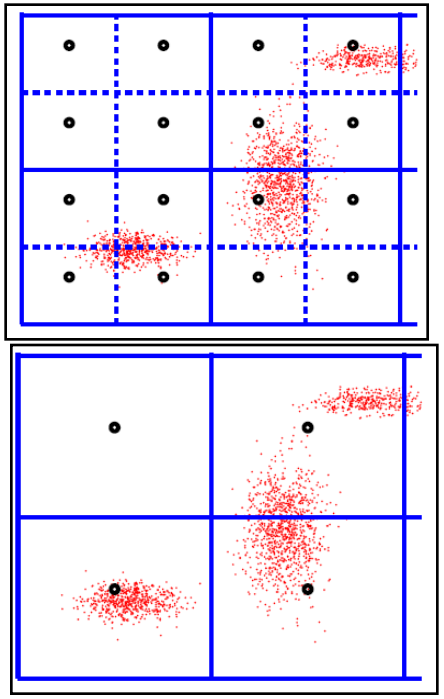
- 128 dimension SIFT description vectors were computed at each interest points.
- The kernel matrix for SVM was generated using the Pyramid Match Kernel (PMK).
- Instead of using uniform bins to build the multi-resolution histogram, a vocabulary guided tree was used.

Vocabulary Guided Tree

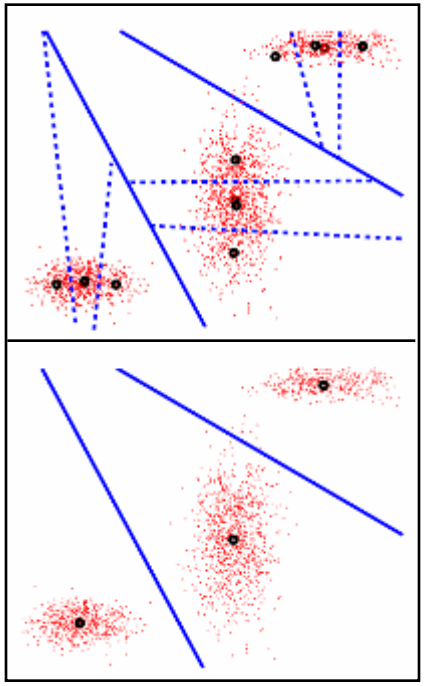
- Proposed by Grauman and Darrell for approximate matching of correspondences in high dimensions.
- Employs hierarchical clustering to group feature vectors into non uniform bins.
- A significant advantage of the VG approach is that it scales with large dimensions of feature vectors unlike the pyramid match kernel with uniform bins.

Comparing uniform bins and VG tree pyramids

Uniform bins



***Vocabulary-guided* bins**



- More accurate in high dimensions ($d > 100$)
- Requires initial corpus of features

Slide from Grauman and Darrell NIPS 2006

Classifier

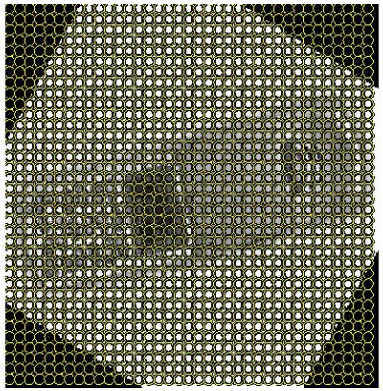
- SVM with a leave-one-out cross validation strategy.
- Each image served as a testing example while the rest served as training examples for a total of 253 test runs in one experiment.
- Classification performance was analyzed via reported accuracy and confusion matrices.

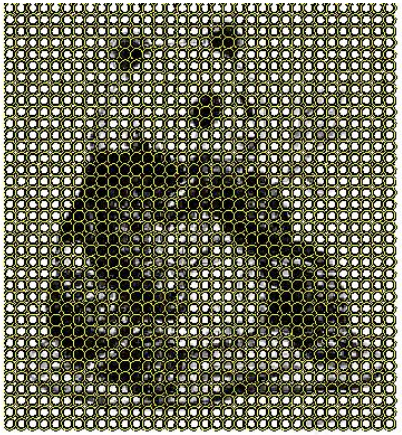
Results

Sampling Strategy	Accuracy
Harris-Affine Interest Points	0.65-0.67
Dense Uniform sampling	0.69-0.73
Harris + MSER combined	0.73 – 0.75

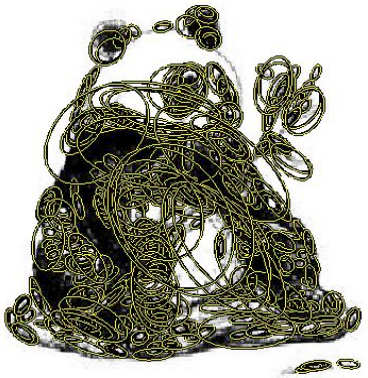
- Classification accuracy of Harris + MSER interest points looks to be the best of the three sampling strategies.

Revisiting the detections

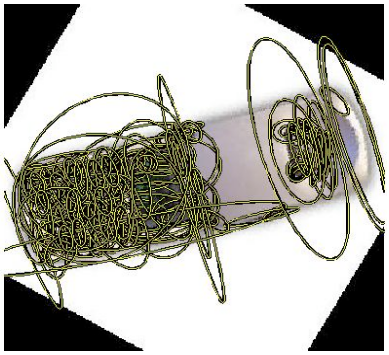


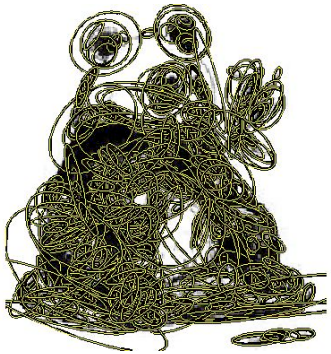


Uniform sampling



Harris affine





Harris + MSER

What do the results and detections suggest?

- Dense sampling is good – provides semantic content often missed with sparse interest point detections.

- However in uniform dense sampling, the regions were too local and non-overlapping.

- In contrast, Harris + MSER detections were sufficiently dense and multiscale, thereby suggesting that it could have provided more semantic information required for object classification.

Confusion matrix – uniform sampling

Classifier result / Truth	Cell Phone	Chair	Lobster	Panda	Pizza	Total
Cell phone	58	1	0	0	0	59
Chair	2	47	5	1	7	62
Lobster	4	6	13	5	13	41
Panda	0	3	5	28	2	38
Pizza	1	9	10	1	32	53

- The classification performance of Cell phone is close to100% while lobster is less than 50%

Confusion matrix – Harris Affine

Classifier result / Truth	Cell Phone	Chair	Lobster	Panda	Pizza	Total
Cell phone	42	11	3	1	2	59
Chair	8	44	5	2	3	62
Lobster	6	5	18	2	10	41
Panda	0	2	3	26	7	38
Pizza	4	0	2	4	43	53

- With the Harris-Affine detections, classification performance of the pizza is much better than the uniform sampling and the classification performance of the lobster shows improvement too. However, the classification performance of the cell phone has dropped significantly when compared to the uniform sampling case.

Confusion matrix – Harris + MSER combined

Classifier result / Truth	Cell Phone	Chair	Lobster	Panda	Pizza	Total
Cell phone	46	7	2	2	2	59
Chair	9	45	5	1	2	62
Lobster	4	7	22	2	6	41
Panda	1	2	3	31	1	38
Pizza	3	1	2	1	46	53

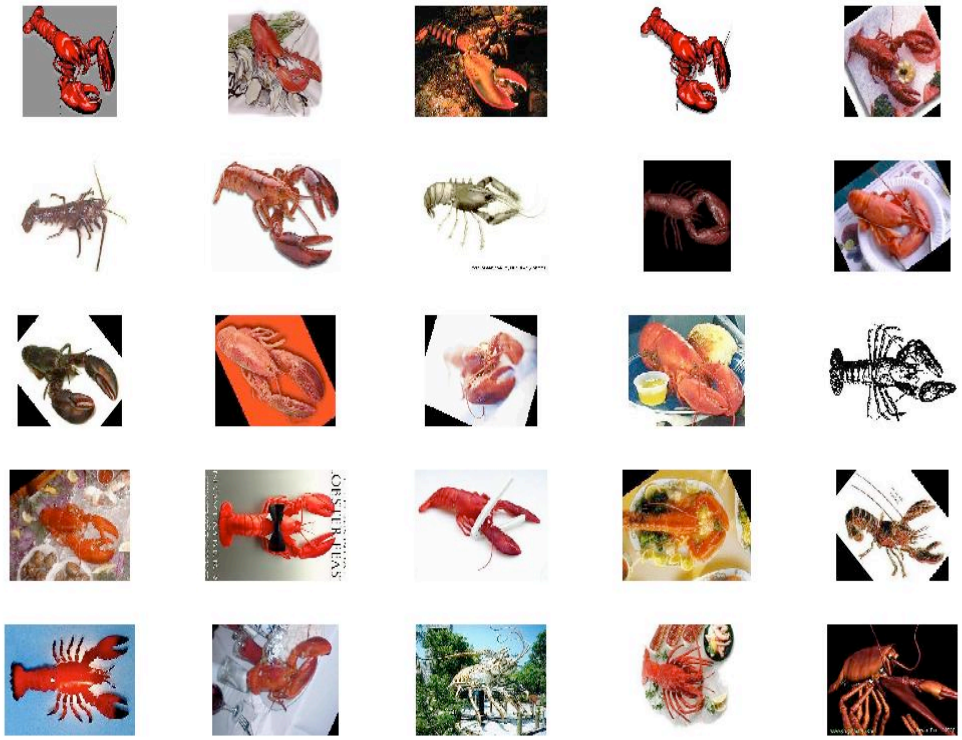
•With the combined detections, classification performance of pizza is better than the other two.

•The classification performance of the lobster and panda are highest with the combined detections – dense overlapping regions provides better semantic context.

•But the cell phone performs poorly when compared to the uniform sampling strategy.

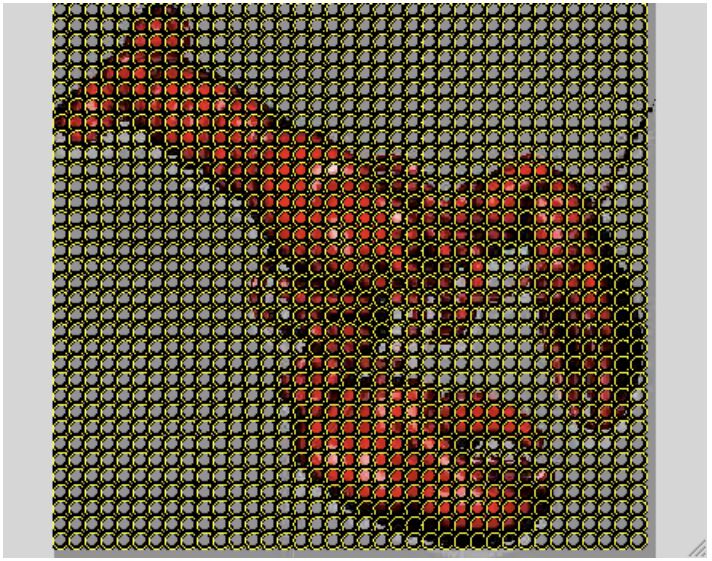
Observations from the Confusion Matrices

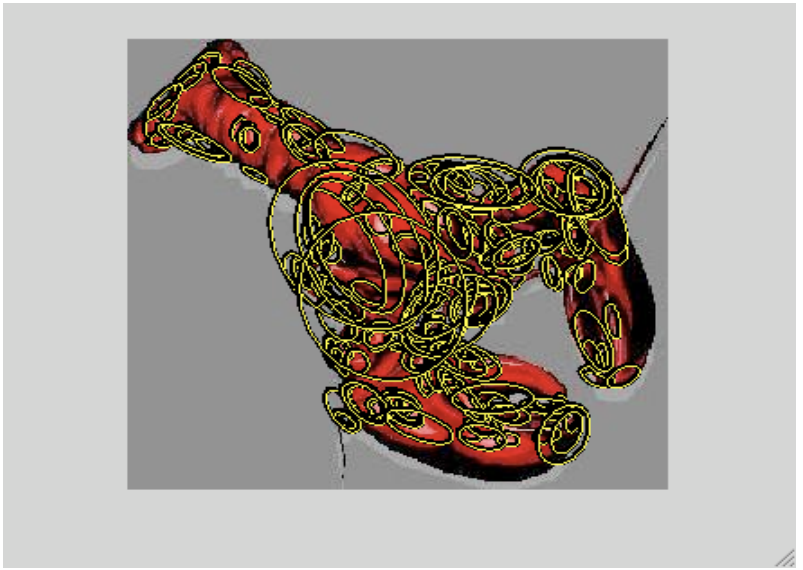
- Notice that the classification performance of the lobster improves from uniform -> Harris-Affine-> Harris + MSER

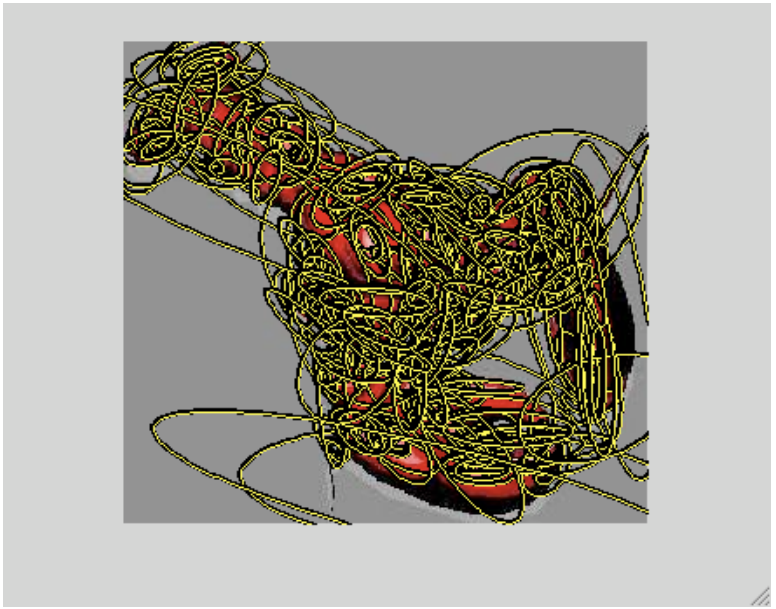


The lobster has probably many more view points than the panda (predominantly frontal pose) or the pizza (predominantly top down)

Analyzing the Lobster



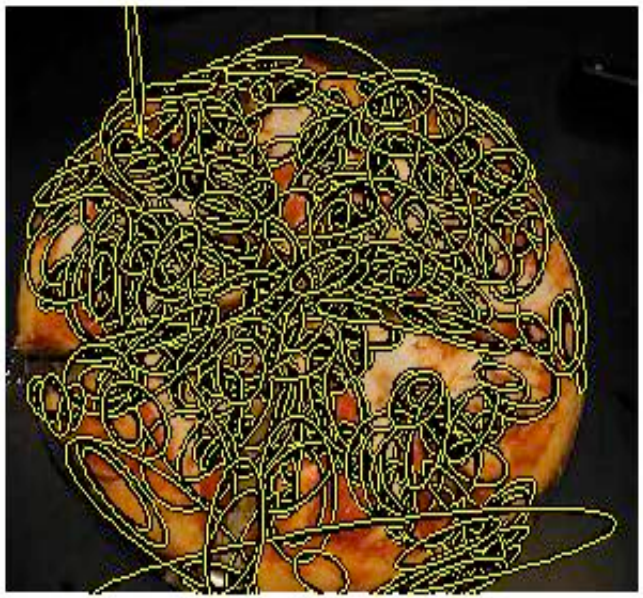




For a lobster, the semantic information pertaining to the relative placement of the whiskers, the legs etc are extremely crucial for classification. Uniform sampling with too small a region(and non-overlapping) does not quite encode this information and hence we see an improvement in performance from uniform -> Harris -> Harris + MSER.

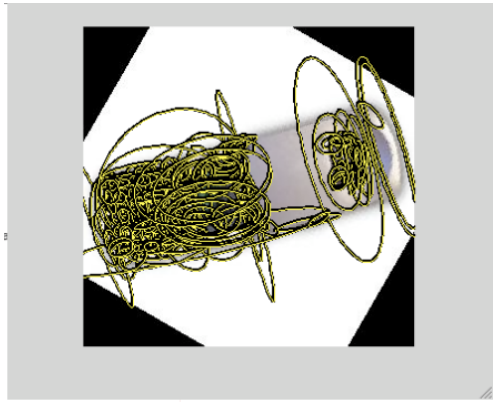
Analyzing the Pizza

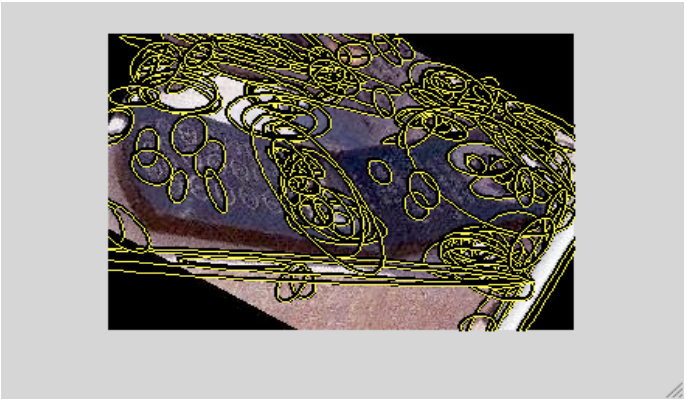
•Likewise, pizza classification is best with the combined detector primarily because a normal pizza is composed of circular regions having a good contrast against the surrounding and the Harris + MSER detector does well on such images.

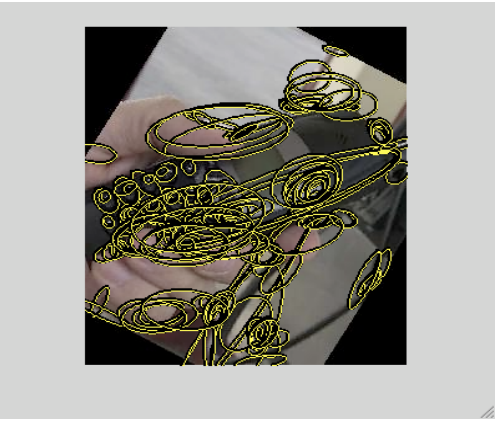


Cell phone performance degradation

- The degradation in the classification performance of the cell phone from uniform dense -> Harris -> Harris + MSER is intriguing.
- Region of uniform intensity between the keypad and display is not picked up by the combined detector.
- Uniform sampling on the other hand picks out each and every region in the image and even though the regions are small, they might be enough to encode the semantic content required to classify a cell phone.







Confusion example!

- This pizza was classified as a cell phone (presumably due to the box flipped open!) in all the 3 cases.

Additional comments

- None of the interest point detectors are biologically motivated (the SIFT interest point detector comes closest primarily due to DOG filtering).

Technical details

- Libpmk - http://people.csail.mit.edu/jjl/libpmk/
- Libpmk feature extraction framework- dependency on ImageMagick++
- Interest point detectors and descriptors - http://www.robots.ox.ac.uk/~vgg/research/affine/descriptors.html#binaries

Acknowledgement

- Thanks to Kristen for her technical inputs and help in putting together this demo.

Questions