

Contents

Preface	xxi
Notation	xxvii
1 The Challenge of Vision	1
1.1 Introduction	1
1.2 Vision	1
1.3 Theories of Vision	4
1.4 What's Next?	31
1.5 Concluding Remarks	32
I FOUNDATIONS	33
2 A Simple Vision System	35
2.1 Introduction	35
2.2 A Simple World: The Blocks World	35
2.3 A Simple Image Formation Model	36
2.4 A Simple Goal	38
2.5 From Images to Edges and Useful Features	38
2.6 From Edges to Surfaces	42
2.7 Generalization	49
2.8 Concluding Remarks	51
3 Looking at Images	53
3.1 Introduction	53
3.2 Looking at Individual Pixels	53
3.3 The More You Look, the More You See	55
3.4 The Eye of the Artist	57
3.5 Tree Shadows and Image Formation	58
3.6 Horizontal or Vertical	59
3.7 Motion Blur	60
3.8 Accidents Happen	62
3.9 Cues for Support	63
3.10 Looking at Raindrops	64
3.11 Plato's Cave	65

3.12	How Do You Know Something Is Wet?	65
3.13	Concluding Remarks	66
4	Computer Vision and Society	67
4.1	Introduction	67
4.2	Fairness	67
4.3	Ethics	72
4.4	Concluding Remarks	73
II	IMAGE FORMATION	75
5	Imaging	77
5.1	Introduction	77
5.2	Light Interacting with Surfaces	77
5.3	The Pinhole Camera and Image Formation	79
5.4	Concluding Remarks	88
6	Lenses	89
6.1	Introduction	89
6.2	Lensmaker's Formula	91
6.3	Imaging with Lenses	97
6.4	Concluding Remarks	106
7	Cameras as Linear Systems	107
7.1	Introduction	107
7.2	Flatland	107
7.3	Cameras as Linear Systems	108
7.4	More General Imagers	110
7.5	Concluding Remarks	116
8	Color	117
8.1	Introduction	117
8.2	Color Physics	117
8.3	Color Perception	125
8.4	Spatial Resolution and Color	131
8.5	Concluding Remarks	134
III	FOUNDATIONS OF LEARNING	135
9	Introduction to Learning	137
9.1	Introduction	137
9.2	Learning from Examples	138
9.3	Learning without Examples	140
9.4	Key Ingredients	140

9.5	Empirical Risk Minimization: A Formalization of Learning from Examples	141
9.6	Learning as Probabilistic Inference	142
9.7	Case Studies	142
9.8	Learning to Learn	149
9.9	Concluding Remarks	150
10	Gradient-Based Learning Algorithms	151
10.1	Introduction	151
10.2	Technical Setting	151
10.3	Basic Gradient Descent	152
10.4	Learning Rate Schedules	153
10.5	Momentum	153
10.6	What Kinds of Functions Can Be Minimized with Gradient Descent?	155
10.7	Stochastic Gradient Descent	159
10.8	Concluding Remarks	160
11	The Problem of Generalization	161
11.1	Introduction	161
11.2	Underfitting and Overfitting	161
11.3	Regularization	165
11.4	Rethinking Generalization	167
11.5	Three Tools in the Search for Truth: Data, Priors, and Hypotheses	167
11.6	Concluding Remarks	173
12	Neural Networks	175
12.1	Introduction	175
12.2	The Perceptron: A Simple Model of a Single Neuron	175
12.3	Multilayer Perceptrons	177
12.4	Activations Versus Parameters	179
12.5	Deep Nets	180
12.6	Deep Learning: Learning with Neural Nets	184
12.7	Catalog of Layers	186
12.8	Why Are Neural Networks a Good Architecture?	189
12.9	Concluding Remarks	190
13	Neural Networks as Distribution Transformers	191
13.1	Introduction	191
13.2	A Different Way of Plotting Functions	191
13.3	How Deep Nets Remap a Data Distribution	193
13.4	Binary Classifier Example	194
13.5	How High-Dimensional Datapoints Get Remapped by Deep Net	196
13.6	Concluding Remarks	198

14	Backpropagation	199
14.1	Introduction	199
14.2	The Trick of Backpropagation: Reuse of Computation	200
14.3	Backward for a Generic Layer	201
14.4	The Full Algorithm: Forward, Then Backward	203
14.5	Backpropagation Over Data Batches	204
14.6	Example: Backpropagation for an MLP	205
14.7	Backpropagation through DAGs: Branch and Merge	212
14.8	Parameter Sharing	214
14.9	Backpropagation to the Data	214
14.10	Concluding Remarks	216
IV	FOUNDATIONS OF IMAGE PROCESSING	217
15	Linear Image Filtering	219
15.1	Introduction	219
15.2	Signals and Images	219
15.3	Systems	223
15.4	Convolution	227
15.5	Cross-Correlation Versus Convolution	235
15.6	System Identification	238
15.7	Concluding Remarks	239
16	Fourier Analysis	241
16.1	Introduction	241
16.2	Image Transforms	241
16.3	Fourier Series	241
16.4	Continuous and Discrete Waves	243
16.5	The Discrete Fourier Transform	247
16.6	Useful Transforms	251
16.7	Discrete Fourier Transform Properties	255
16.8	A Family of Fourier Transforms	261
16.9	Fourier Analysis as an Image Representation	262
16.10	Fourier Analysis of Linear Filters	267
16.11	Concluding Remarks	272
V	LINEAR FILTERS	273
17	Blur Filters	275
17.1	Introduction	275
17.2	Box Filter	276
17.3	Gaussian Filter	279
17.4	Binomial Filters	283

17.5	Concluding Remarks	286
18	Image Derivatives	287
18.1	Introduction	287
18.2	Discretizing Image Derivatives	287
18.3	Gradient-Based Image Representation	291
18.4	Image Editing in the Gradient Domain	292
18.5	Gaussian Derivatives	293
18.6	High-Order Gaussian Derivatives	295
18.7	Derivatives of Binomial Filters	299
18.8	Image Gradient and Directional Derivatives	301
18.9	Image Laplacian	302
18.10	A Simple Model of the Early Visual System	305
18.11	Sharpening Filter	307
18.12	Retinex	309
18.13	Concluding Remarks	313
19	Temporal Filters	315
19.1	Introduction	315
19.2	Modeling Sequences	315
19.3	Modeling Sequences in the Fourier Domain	317
19.4	Temporal Filters	318
19.5	Concluding Remarks	324
VI	SAMPLING AND MULTISCALE IMAGE REPRESENTATIONS	325
20	Image Sampling and Aliasing	327
20.1	Introduction	327
20.2	Aliasing	327
20.3	Sampling Theorem	329
20.4	Reconstruction	334
20.5	Ideal Reconstruction	334
20.6	A Family of 2D Spatial Samplings	338
20.7	Anti-Aliasing Filter	340
20.8	Spatiotemporal Sampling	342
20.9	Concluding Remarks	342
21	Downsampling and Upsampling Images	345
21.1	Introduction	345
21.2	Example: Aliasing-Based Adversarial Attack	345
21.3	Downsampling	346
21.4	Upsampling	358
21.5	Concluding Remarks	363

22	Filter Banks	365
22.1	Introduction	365
22.2	Gabor Filters	365
22.3	Steerable Filters and Orientation Analysis	374
22.4	Motion Analysis	380
22.5	Concluding Remarks	383
23	Image Pyramids	385
23.1	Introduction	385
23.2	Image Pyramids and Multiscale Image Analysis	386
23.3	Linear Image Transforms	387
23.4	Gaussian Pyramid	388
23.5	Laplacian Pyramid	390
23.6	Steerable Pyramid	395
23.7	A Pictorial Summary	397
23.8	Concluding Remarks	399
VII	NEURAL ARCHITECTURES FOR VISION	401
24	Convolutional Neural Nets	403
24.1	Introduction	403
24.2	Convolutional Layers	404
24.3	Nonlinear Filtering Layers	414
24.4	A Simple CNN Classifier	415
24.5	A Worked Example	417
24.6	Feature Maps in CNNs	420
24.7	Receptive Fields	423
24.8	Spatial Outputs	424
24.9	CNN as a Sliding Filter	425
24.10	Why Process Images Patch by Patch?	426
24.11	Popular CNN Architectures	427
24.12	Concluding Remarks	430
25	Recurrent Neural Nets	431
25.1	Introduction	431
25.2	Recurrent Layer	433
25.3	Backpropagation through Time	433
25.4	Stacking Recurrent Layers	435
25.5	Long Short-Term Memory	436
25.6	Concluding Remarks	437
26	Transformers	439
26.1	Introduction	439

26.2	A Limitation of CNNs: Independence between Far Apart Patches	439
26.3	The Idea of Attention	440
26.4	A New Data Type: Tokens	440
26.5	Token Nets	444
26.6	The Attention Layer	445
26.7	The Full Transformer Architecture	453
26.8	Permutation Equivariance	455
26.9	CNNs in Disguise	456
26.10	Masked Attention	458
26.11	Positional Encodings	460
26.12	Comparing Fully Connected, Convolutional, and Self-Attention Layers	462
26.13	Concluding Remarks	463
VIII	PROBABILISTIC MODELS OF IMAGES	465
27	Statistical Image Models	467
27.1	Introduction	467
27.2	How Do We Tell Noise from Texture?	469
27.3	Independent Pixels	470
27.4	Dead Leaves Model	474
27.5	The Gaussian Model	477
27.6	The Wavelet Marginal Model	482
27.7	Nonparametric Markov Random Field Image Models	489
27.8	Concluding Remarks	490
28	Textures	493
28.1	Introduction	493
28.2	A Few Notes about Human Perception	494
28.3	Heeger-Bergen Texture Analysis and Synthesis	496
28.4	Efros-Leung Texture Analysis and Synthesis Model	501
28.5	Connection to Deep Generative Models	503
28.6	Concluding Remarks	504
29	Probabilistic Graphical Models	505
29.1	Introduction	505
29.2	Simple Examples	505
29.3	Directed Graphical Models	509
29.4	Inference in Graphical Models	510
29.5	Simple Example of Inference in a Graphical Model	511
29.6	Belief Propagation	512
29.7	Loopy Belief Propagation	520
29.8	Relationship of Probabilistic Graphical Models to Neural Networks	523
29.9	Concluding Remarks	523

IX	GENERATIVE IMAGE MODELS AND REPRESENTATION LEARNING	525
30	Representation Learning	527
30.1	Introduction	527
30.2	Problem Setting	527
30.3	What Makes for a Good Representation?	528
30.4	Autoencoders	530
30.5	Predictive Encodings	533
30.6	Self-Supervised Learning	535
30.7	Imputation	536
30.8	Abstract Pretext Tasks	537
30.9	Clustering	537
30.10	Contrastive Learning	542
30.11	Concluding Remarks	547
31	Perceptual Grouping	549
31.1	Introduction	549
31.2	Why Group?	550
31.3	Segments	551
31.4	Edges, Boundaries, and Contours	555
31.5	Layers	556
31.6	Emergent Groups	556
31.7	Concluding Remarks	557
32	Generative Models	559
32.1	Introduction	559
32.2	Unconditional Generative Models	561
32.3	Learning Generative Models	563
32.4	Density Models	565
32.5	Energy-Based Models	567
32.6	Gaussian Density Models	570
32.7	Autoregressive Density Models	572
32.8	Diffusion Models	576
32.9	Generative Adversarial Networks	579
32.10	Concluding Remarks	581
33	Generative Modeling Meets Representation Learning	583
33.1	Introduction	583
33.2	Latent Variables as Representations	584
33.3	Technical Setting	585
33.4	Variational Autoencoders	586
33.5	Do VAEs Learn Good Representations?	598
33.6	Generative Adversarial Networks Are Representation Learners Too	600

33.7	Concluding Remarks	601
34	Conditional Generative Models	603
34.1	Introduction	603
34.2	A Motivating Example: Image Colorization	603
34.3	Conditional Generative Models Solve Multimodal Structured Prediction	608
34.4	A Tour of Popular Conditional Models	609
34.5	Structured Prediction in Vision	613
34.6	Image-to-Image Translation	614
34.7	Concluding Remarks	620
X	CHALLENGES IN LEARNING-BASED VISION	621
35	Data Bias and Shift	623
35.1	Introduction	623
35.2	Out-of-Distribution Generalization	625
35.3	A Toy Example	627
35.4	Dataset Bias	630
35.5	Sources of Bias	632
35.6	Adversarial Shifts	636
35.7	Concluding Remarks	637
36	Training for Robustness and Generality	639
36.1	Introduction	639
36.2	Data Augmentation	639
36.3	Adversarial Training	643
36.4	Toward General-Purpose Vision Models	643
36.5	Concluding Remarks	644
37	Transfer Learning and Adaptation	645
37.1	Introduction	645
37.2	Problem Setting	645
37.3	Finetuning	646
37.4	Learning from a Teacher	649
37.5	Prompting	651
37.6	Domain Adaptation	653
37.7	Generative Data	654
37.8	Other Kinds of Knowledge that Can Be Transferred	655
37.9	A Combinatorial Catalog of Transfer Learning Methods	655
37.10	Sequence Models from the Lens of Adaptation	656
37.11	Concluding Remarks	656
XI	UNDERSTANDING GEOMETRY	657

38	Representing Images and Geometry	659
38.1	Introduction	659
38.2	Homogeneous and Heterogenous Coordinates	660
38.3	2D Image Transformations	661
38.4	Lines and Planes in Homogeneous Coordinates	668
38.5	Image Warping	670
38.6	Implicit Image Representations	671
38.7	Concluding Remarks	673
39	Camera Modeling and Calibration	675
39.1	Introduction	675
39.2	3D Camera Projections in Homogeneous Coordinates	676
39.3	Camera-Intrinsic Parameters	678
39.4	Camera-Extrinsic Parameters	683
39.5	Full Camera Model	685
39.6	A Few Concrete Examples	686
39.7	Camera Calibration	692
39.8	Concluding Remarks	699
40	Stereo Vision	701
40.1	Introduction	701
40.2	Stereo Cues	702
40.3	Model-Based Methods	706
40.4	Learning-Based Methods	717
40.5	Evaluation	719
40.6	Concluding Remarks	719
41	Homographies	721
41.1	Introduction	721
41.2	Homography	722
41.3	Creating Image Panoramas	727
41.4	Concluding Remarks	730
42	Single View Metrology	731
42.1	Introduction	731
42.2	A Few Notes about Perception of Depth by Humans	732
42.3	Linear Perspective	735
42.4	Measuring Heights Using Parallel Lines	741
42.5	3D Metrology from a Single View	749
42.6	Camera Calibration from Vanishing Points	753
42.7	Concluding Remarks	755
43	Learning to Estimate Depth from a Single Image	757

43.1	Introduction	757
43.2	Monocular Depth Cues	757
43.3	3D Representations	758
43.4	Supervised Methods for Depth from a Single Image	761
43.5	Unsupervised Methods for Depth from a Single Image	764
43.6	Concluding Remarks	767
44	Multiview Geometry and Structure from Motion	769
44.1	Introduction	769
44.2	Structure from Motion	769
44.3	Sparse SFM	771
44.4	Concluding Remarks	780
45	Radiance Fields	783
45.1	Introduction	783
45.2	What is a Radiance Field?	784
45.3	Representing Radiance Fields With Parameterized Functions	787
45.4	Rendering Radiance Fields	789
45.5	Fitting a Radiance Field to Explain a Scene	793
45.6	Beyond Radiance Fields: The Rendering Equation	797
45.7	Concluding Remarks	798
XII	UNDERSTANDING MOTION	799
46	Motion Estimation	801
46.1	Introduction	801
46.2	Motion Perception in the Human Visual System	802
46.3	Matching-Based Motion Estimation	804
46.4	Does the Human Visual System Use Matching to Estimate Motion?	808
46.5	Concluding Remarks	811
47	3D Motion and Its 2D Projection	813
47.1	Introduction	813
47.2	3D Motion and Its 2D Projection	813
47.3	Concluding Remarks	822
48	Optical Flow Estimation	823
48.1	Introduction	823
48.2	2D Motion Field and Optical Flow	823
48.3	Model-Based Approaches	826
48.4	Concluding Remarks	834
49	Learning to Estimate Motion	835
49.1	Introduction	835

49.2	Learning-Based Approaches	835
49.3	Concluding Remarks	839
XIII	UNDERSTANDING VISION WITH LANGUAGE	841
50	Object Recognition	843
50.1	Introduction	843
50.2	A Few Notes About Object Recognition in Humans	844
50.3	Image Classification	847
50.4	Object Localization	854
50.5	Class Segmentation	863
50.6	Instance Segmentation	865
50.7	Concluding Remarks	867
51	Vision and Language	869
51.1	Introduction	869
51.2	Background: Representing Text as Tokens	869
51.3	Learning Visual Representations from Language Supervision	871
51.4	Translating between Images and Text	877
51.5	Text as a Visual Representation	882
51.6	Visual Question Answering	883
51.7	Concluding Remarks	884
XIV	ON RESEARCH, WRITING AND SPEAKING	885
52	How to Do Research	887
52.1	Introduction	887
52.2	Research Advice	887
52.3	Concluding Remarks	891
53	How to Write Papers	893
53.1	Introduction	893
53.2	Organization	894
53.3	General Writing Tips	896
53.4	Concluding Remarks	901
54	How to Give Talks	903
54.1	Introduction	903
54.2	Very Short Talks (2 – 10 minutes)	903
54.3	Preparation	904
54.4	Nervousness	905
54.5	Your Distracted Audience	905
54.6	Ways to Engage the Audience	905
54.7	Show Yourself to the Audience	906

54.8	Concluding Remarks	907
XV	CLOSING REMARKS	909
55	A Simple Vision System—Revisited	911
55.1	Introduction	911
55.2	A Simple Neural Network	911
55.3	From 2D Images to 3D	914
55.4	Large Language Model-based Scene Understanding	916
55.5	Unsolved Solved Computer Vision Problems	918
55.6	Concluding Remarks	918
	Bibliography	921
	Index	943