Agglomerative Grouping of Observations by Bounding Entropy Variation

Christian Beder

Institute for Photogrammetry
Bonn University, Germany
beder@ipb.uni-bonn.de

Abstract. An information theoretic framework for grouping observations is proposed. The entropy change incurred by new observations is analyzed using the Kalman filter update equations. It is found, that the entropy variation is caused by a positive similarity term and a negative proximity term. Bounding the similarity term in the spirit of the minimum description length principle and the proximity term in the spirit of maximum entropy inference a robust and efficient grouping procedure is devised. Some of its properties are demonstrated for the exemplary task of edgel grouping.

1 Introduction

Grouping observations has been identified as an important issue in many computer vision tasks and has been studied by many researchers (cf. [9], [10], [1], [11]). In this context the Gestalt laws of psychology have received much attention and the criterion of Prägnanz is considered extremely useful (cf. [13]). Its close connection to the information theoretic minimum description length criterion (cf. [14]) has been pointed out by [10] and [13].

In [12] the grouping is established based on local measures specially tailored for the task of edgel grouping. A similar approach is made in [3], but there the probability distributions of the observations are explicitly modeled and used to guide the grouping. An information theoretic approach is made in [13] by phrasing the various Gestalt principles in terms of energy functions and minimizing the overall free energy. Also the tensor voting approach of [5], [6] or [11] uses a global consistency measure based on local measures of similarity and proximity.

The problem with the minimum description length criterion of [14] in the context of an agglomerative grouping procedure is, that locally minimizing entropy contradicts the principle of maximum entropy inference (cf. [7]), since greedily grouping distant observations leads to the greatest entropy reduction. This effect is also known from robust statistics as leverage points (cf. [8]). To cope with this problem, the Kalman filter update equations (cf. [4]) will be reviewed, and the entropy change incurred by grouping a new observation is analyzed. It is found, that this entropy variation is caused by a positive observation dependent term, that measures similarity and a negative design dependent term, that measures proximity. A grouping algorithm based on bounding both influences on the

entropy variation is proposed, so that entropy reduction is caused by the observations in the spirit of minimum description length but the reduction through decisions by the algorithm is bounded from below in the spirit of maximum entropy inference. The algorithm and some of its properties will be demonstrated for the exemplary task of edgel grouping.

2 The Kalman Filter

Having two sets of independent observations $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ of size N_1 and N_2 with known covariance matrices $\boldsymbol{C}_{11}$ and $\boldsymbol{C}_{22}$ and a model depending on the parameter vector $\boldsymbol{p}$ of size U given by the two functions

$$\boldsymbol{g}_1(\boldsymbol{p}) = \boldsymbol{l}_1 \qquad \text{and} \qquad \boldsymbol{g}_2(\boldsymbol{p}) = \boldsymbol{l}_2$$

with the Jacobians

$$\frac{\partial \boldsymbol{g}_1}{\partial \boldsymbol{p}} = \boldsymbol{A}_1 \qquad \text{and} \qquad \frac{\partial \boldsymbol{g}_2}{\partial \boldsymbol{p}} = \boldsymbol{A}_2$$

the best linear unbiased estimation of the parameters $\hat{\boldsymbol{p}}^{(-)}$ is found for the first set of observations using the expected covariance matrix (cf. [8])

$$\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} = (\boldsymbol{A}_1^T \boldsymbol{C}_{11}^{-1} \boldsymbol{A}_1)^{-1} \tag{1}$$

to be

$$\hat{\boldsymbol{p}}^{(-)} = \boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} \boldsymbol{A}_1^T \boldsymbol{C}_{11}^{-1} \boldsymbol{l}_1 \tag{2}$$

The redundancy of the estimation is given by

$$R_1 = N_1 - U \tag{3}$$

and in case $R_1 > 0$, using the residuals

$$\hat{\boldsymbol{v}}_1 = \boldsymbol{A}_1 \hat{\boldsymbol{p}}^{(-)} - \boldsymbol{l}_1$$

and their weighted squared sum

$$\Omega^{2(-)} = \hat{\boldsymbol{v}}_1^T \boldsymbol{C}_{11}^{-1} \hat{\boldsymbol{v}}_1 \tag{4}$$

the covariance matrix of the estimated parameters can be obtained as

$$\hat{\boldsymbol{C}}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^{(-)}} = \frac{\Omega^{2(-)}}{R_1} \boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)}$$

Thereafter it is possible to updated the estimation sequentially including the second set of observations. This is well known as Kalman filtering (cf. [4]) and using the prediction error and its covariance matrix

$$\hat{\boldsymbol{v}}_2 = \boldsymbol{A}_2 \hat{\boldsymbol{p}}^{(-)} - \boldsymbol{l}_2 \qquad\qquad \boldsymbol{C}_{\hat{\boldsymbol{v}}_2\hat{\boldsymbol{v}}_2} = \boldsymbol{C}_{22} + \boldsymbol{A}_2 \boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} \boldsymbol{A}_2^T$$

and the Kalman filter gain matrix

$$\boldsymbol{F} = \boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} \boldsymbol{A}_2^T \boldsymbol{C}_{\hat{\boldsymbol{v}}_2\hat{\boldsymbol{v}}_2}^{-1} \tag{5}$$

the Kalman filter update equations are obtained as

$$\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(+)} = \boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} - \boldsymbol{F}\boldsymbol{A}_2\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} \quad (6) \qquad \Omega^{2(+)} = \Omega^{2(-)} + \Delta\Omega^2 \quad (9)$$

$$\hat{\boldsymbol{p}}^{(+)} = \hat{\boldsymbol{p}}^{(-)} + \boldsymbol{F}\hat{\boldsymbol{v}}_2 \quad (7) \qquad \Delta R = N_2 \quad (10)$$

$$\Delta\Omega^2 = \hat{\boldsymbol{v}}_2^T \boldsymbol{C}_{\hat{\boldsymbol{v}}_2\hat{\boldsymbol{v}}_2}^{-1} \hat{\boldsymbol{v}}_2 \quad (8) \qquad R_2 = R_1 + \Delta R \quad (11)$$

Finally the estimated covariance matrix of the parameters may be recomputed from the residuals using

$$\hat{\boldsymbol{C}}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(+)} = \tfrac{\Omega^{2(+)}}{R_2} \boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(+)}$$

3 Estimated Entropy Variation

The Kalman filter was used to sequentially estimate the first two moments $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{C}}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}$ of the distribution of the parameters. Knowing only those two moments, the maximum possible entropy of the estimation is (cf. [2])

$$\begin{aligned}\hat{h}(\hat{\boldsymbol{p}}) &= \frac{1}{2}\log\left|2\pi e \hat{\boldsymbol{C}}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}\right| = \frac{1}{2}\log\left|2\pi e \frac{\Omega^2}{R}\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}\right| \\ &= \frac{U}{2}\log\left(2\pi e \frac{\Omega^2}{R}\right) + \frac{1}{2}\log\left|\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}\right|\end{aligned}$$

Note that only the first term is caused by the randomness of the observations and the second term depends only on the geometry of the design of the estimation. Applying the results from Kalman filtering, the estimated entropy change by including the second set of observations is in case, that $R_1 > 0$

$$\begin{aligned}\Delta h = \hat{h}(\hat{\boldsymbol{p}}^{(+)}) - \hat{h}(\hat{\boldsymbol{p}}^{(-)}) &= \frac{1}{2}\log\left|2\pi e \hat{\boldsymbol{C}}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(+)}\right| - \frac{1}{2}\log\left|2\pi e \hat{\boldsymbol{C}}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)}\right| \\ &= \frac{U}{2}\log\frac{1+\frac{\Delta\Omega^2}{\Omega^{2(-)}}}{1+\frac{\Delta R}{R_1}} + \frac{1}{2}\log\frac{\left|\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)} - \boldsymbol{F}\boldsymbol{A}_2\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)}\right|}{\left|\boldsymbol{C}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}^{(-)}\right|} \\ &= \underbrace{\frac{U}{2}\log\frac{1+\frac{\Delta\Omega^2}{\Omega^{2(-)}}}{1+\frac{\Delta R}{R_1}}}_{\Delta h_o} + \underbrace{\frac{1}{2}\log|\boldsymbol{I} - \boldsymbol{F}\boldsymbol{A}_2|}_{\Delta h_d} \quad (12)\end{aligned}$$

Again the entropy change is constituted from a positive term Δh_o, that refelcts the increase in randomness due to the new observation, and a second negative term Δh_d, that reflects the decrease in randomness due to the decision of including the new observation into the estimation.

The first term Δh_o is closely related to the well known, and in case of Normal distributed observations Fisher distributed, test statistic

$$T = \frac{\frac{\Delta\Omega^2}{\Omega^{2(-)}}}{\frac{\Delta R}{R_1}} \propto \mathcal{F}(\Delta R, R_1)$$

that is frequently used to decide, if the second observation fits the model defined by the first. Given a significance level α, a threshold T_α is derived from the inverse of the Fisher distribution and the decision is made by comparing it with the test statistic T. The test is not rejected, if $T < T_\alpha$ or equivalent

$$\Delta h_o < \tfrac{U}{2} \log \left(T_\alpha \left(1 - \tfrac{1}{1+\frac{\Delta R}{R_1}} \right) + \tfrac{1}{1+\frac{\Delta R}{R_1}} \right) =: L_o \tag{13}$$

If $R_1 = 0$, the variance factor cannot be estimated from the observations and must therefore assumed to be known. Thus the entropy change incurred by including the new observation into the observation in case of $R_1 = 0$ is

$$\begin{aligned} \Delta h = \hat{h}(\hat{\boldsymbol{p}}^{(+)}) - h(\hat{\boldsymbol{p}}^{(-)}) &= \frac{1}{2} \log \left| 2\pi e \hat{\boldsymbol{C}}^{(+)}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}} \right| - \frac{1}{2} \log \left| 2\pi e \boldsymbol{C}^{(-)}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}} \right| \\ &= \underbrace{\frac{U}{2} \log \left(2\pi e \frac{\Delta \Omega^2}{\Delta R} \right)}_{\Delta h_o} + \underbrace{\frac{1}{2} \log |\boldsymbol{I} - \boldsymbol{F}\boldsymbol{A}_2|}_{\Delta h_d} \end{aligned} \tag{14}$$

Again the first term Δh_o is related to a well known, and in case of Normal distributed observations χ^2-distributed, test statistic for the case, that the variance factor is known

$$T' = \tfrac{\Delta \Omega^2}{\Delta R} \propto \chi^2(\Delta R)$$

so that again a T'_α is derived from the inverse of the χ^2-distribution, and the hypothesis is not rejected, if

$$\Delta h_o < \tfrac{U}{2} \log \left(2\pi e T'_\alpha \right) =: L_o \tag{15}$$

This must be used to decide, if the new observation fits the model defined by the previous observation in case that $R_1 = 0$.

The above criterion measures the similarity between the new observation and the model estimated from the previous observations. The key idea here is, that also the entropy decrease Δh_d resulting from any decision made by the algorithm should be bounded, yielding a proximity criterion for the observations. To find this bound, the history of previous design matrices $\boldsymbol{A}_j$ and covariance matrices $\boldsymbol{C}_{jj}$ is analyzed, because the geometry of the previous observations defines the border, inside which the new observations may be encountered. Allowing new observations to be a bit outside the range of the previous observations by introducing a proximity factor $\lambda > 1$, the bound is found to be

$$\Delta h_d > \lambda \min_j \tfrac{1}{2} \log \left| \boldsymbol{I} - (\boldsymbol{C}^{(-)}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}} \boldsymbol{A}_j^T (\boldsymbol{C}_{jj} + \boldsymbol{A}_j \boldsymbol{C}^{(-)}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}} \boldsymbol{A}_j^T)^{-1} \boldsymbol{A}_j \right| =: L_d \tag{16}$$

4 The Grouping Algorithm

In the preceding section a similarity and a proximity criterion based on the information increase of including a new set of observations into an estimation were derived. Those two criteria could be used to decide, if a new set of observations could be grouped with an existing set of observations. Furthermore the Kalman

filter update equations yield an efficient method to aggregate observations sequentially, thus enabling a very efficient agglomerative grouping strategy.

The greedy method proposed here starts from an arbitrary observation and sequentially aggregates new observations. In order to decide, which new observation is to be aggregated next, first the threshold on the design dependent entropy loss L_d is computed from all observations already aggregated. Then for every possible observation the observation dependent entropy increase Δh_o and the design dependent entropy loss $-\Delta h_d$ are computed and compared to the two thresholds. Among the qualifying observations, the grouping decision, that destroys fewest information, is chosen in the spirit of maximum entropy inference, i.e. the candidate observation is aggregated, for which the design dependent entropy loss $-\Delta h_d$ is minimal. This aggregation process is continued, until no more observations qualify according to the two criteria. Note that the criteria are efficiently computable due to the Kalman filter update equations. Finally the aggregated observations are removed and the whole process is repeated until all groups are found.

The complete grouping procedure is summarized in algorithm 1.

Algorithm 1 Grouping Algorithm

let the observations be $\mathcal{Y} = \{(\boldsymbol{l}_i, \boldsymbol{C}_{ii})\}$
while $\mathcal{Y} \neq \emptyset$ **do**
  pick initial $\boldsymbol{l}_1 \in \mathcal{Y}$, compute the Jacobian $\boldsymbol{A}_1$ of $\boldsymbol{g}_1$
  start the group $\mathcal{G} = \{\boldsymbol{l}_1\}$
  compute initial $\boldsymbol{C}^{(-)}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}$, $\hat{\boldsymbol{p}}^{(-)}$, $\Omega^{2\,(-)}$ and R_1 according to (1), (2), (4) and (3)
  repeat
    compute the threshold L_d from $\mathcal{G}$ according to (16)
    determine the threshold L_o depending on R_1 according to (13) or (15)
    initialize the candidate set $\mathcal{C} = \emptyset$
    for all $\boldsymbol{l}_i \in \mathcal{Y} \backslash \mathcal{G}$ **do**
      compute the Jacobian $\boldsymbol{A}_i$ of $\boldsymbol{g}_i$ at $\hat{\boldsymbol{p}}^{(-)}$
      compute $\boldsymbol{F}^{(i)}$, $\Delta\Omega^{2\,(i)}$ and $\Delta R^{(i)}$ for $\boldsymbol{l}_i$ according to (5), (8) and (10)
      compute ${\Delta h_o}^{(i)}$ depending on R_1 according to (12) or (14)
      compute ${\Delta h_d}^{(i)}$ according to (12)
      if ${\Delta h_o}^{(i)} < L_o \wedge {\Delta h_d}^{(i)} > L_d$ **then**
        include the candidate $\mathcal{C} = \mathcal{C} \cup \{\boldsymbol{l}_i\}$
      end if
    end for
    pick $\boldsymbol{l}_c \in \mathcal{C}$ with maximum ${\Delta h_d}^{(c)}$
    include it into the group $\mathcal{G} = \mathcal{G} \cup \{\boldsymbol{l}_c\}$
    update $\boldsymbol{C}^{(-)}_{\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}}$, $\hat{\boldsymbol{p}}^{(-)}$, $\Omega^{2\,(-)}$ and R_1 for $\boldsymbol{l}_c$ according to (6), (7), (9) and (11)
  until $\mathcal{C} = \emptyset$
  output group $\mathcal{G}$
  $\mathcal{Y} = \mathcal{Y} \backslash \mathcal{G}$
end while

5 Example: Edgel Grouping

The simple problem of grouping edgels in images to straight lines has been studied extensively and will be used here to demonstrate some properties of the presented grouping algorithm.

Using the measured image coordinates (r_i, c_i) together with the image gradients (n_{r_i}, n_{c_i}) the simple linear line model

$$\begin{pmatrix} c_i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} r_i \\ \frac{n_{r_i}}{n_{c_i}} \end{pmatrix}$$

can be used. Note that the coordinate system can be rotated for each group, so that the model can easily deal with vertical lines.

The 1000×1000 pixel patch depicted on the left hand side of figure 1 was cut out of an aerial image. The edge pixels were extracted using the Canny edge detector and for each edge pixel the gradient was computed using the Sobel operator. The resulting set of edgels and their gradients are shown on the right hand side of figure 1.

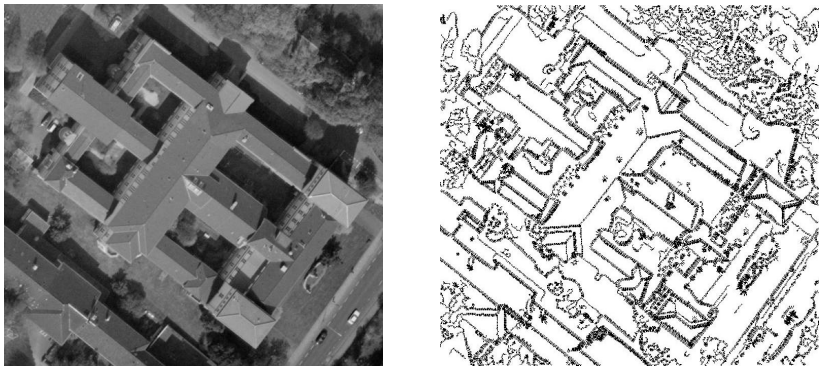

Fig. 1. 1000×1000 pixel patch cut out of an aerial image and the extracted edgels.

The edgels were aggregated using the proposed grouping algorithm and two exemplary groups are shown in figure 2. Note that the two groups were not linked together, although they are on the same line and would be joined, if only the observation dependent similarity criterion had been used. Since no intermediate observation points are present, the proximity criterion imposed by the design of the estimation prevented further growth of the group.

The proximity factor was chosen as $\lambda = 1.5$, i.e. the grouping algorithm was allowed to decrease the entropy by $\frac{3}{2}$ the maximum number of bits, that any edgel lying between the other edgels would do. The resulting candidate sets of new observations for two stages of aggregation of the same exemplary group are shown in figure 3. The black dots are the qualifying observations according to the

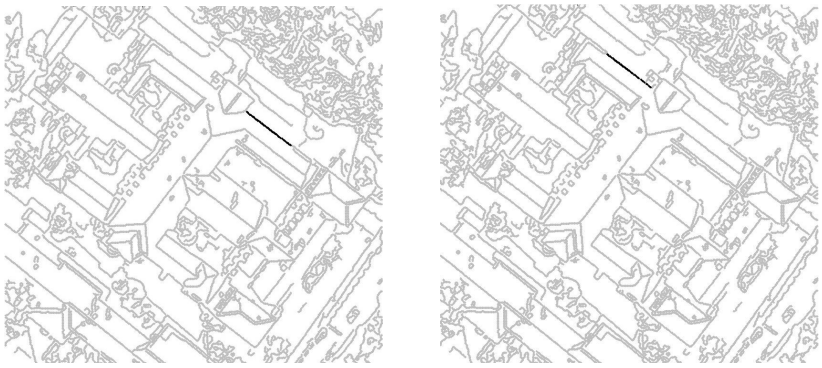

Fig. 2. Two exemplary groups of edgels.

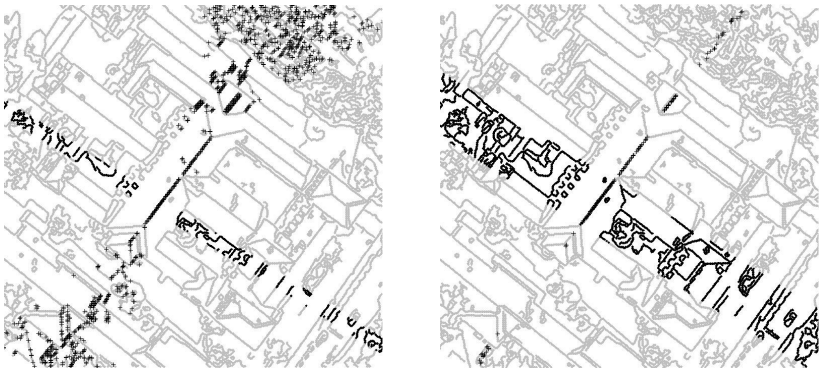

Fig. 3. Candidate sets of new observations at two stages of the same exemplary group. The black dots are the qualifying observations according to the proximity criterion, the black crosses according to the similarity criterion and the black asterisks are the intersection of both criteria.

proximity criterion, the black crosses are the qualifying observations according to the similarity criterion and the black asterisks at the intersection of both are the qualifying observations for the grouping. It can also be seen, that the two criteria are orthogonal.

On the left hand side of figure 3 an early stage of aggregation is shown. The model line is still very uncertain especially far away from the few defining edgels. On the other hand, the range of qualifying observations on the line imposed by the proximity criterion is very narrow so that this effect is compensated and does not affect the grouping.

On the right hand side of figure 3 a latter stage of aggregation is shown. Observe that the model line has now become very narrow, reflecting the fact, that now many aggregated observations contribute to the estimation. The range of qualifying observations along the line has become wider, so that new observations can be collected.

6 Summary and Conclusions

An information theoretic framework for the grouping of observations was proposed. By analyzing the entropy change incurred by including a new observation into an estimation a similarity criterion, that minimizes description length, and a proximity criterion, that enforces maximum entropy decisions, was derived. Based on those two criteria a grouping algorithm was proposed, that, using the efficient Kalman filter update equations, greedily reduces description length and at the same time ensures robustness through maximum entropy inference.

The applicability of the presented method goes far beyond the presented edgel grouping example. Whenever similarity is defined by a known parametric object model, the presented method may be applied. This is the case for many important geometric grouping problems, like for example aggregating 3D-surface patches obtained from dense stereo matching or laser scanning to planes or conics, and will be subject to further investigation.

References

1. Kim L. Boyer and Sudeep Sarkar. Perceptual organization in computer vision: status, challenges, and potential. *Comput. Vis. Image Underst.*, 76(1):1–5, 1999.
2. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.
3. Daniel Crevier. A probabilistic method for extracting chains of collinear segments. *CVIU*, 76(1):36–53, 1999.
4. Wolfgang Förstner and Bernhard Wrobel. Mathematical concepts in photogrammetry. In J.C.McGlone, E.M.Mikhail, and J.Bethel, editors, *Manual of Photogrammetry*. ASPRS, 2004.
5. Gideon Guy and Gérard Medioni. Inferring global perceptual contours from local features. *Int. J. Comput. Vision*, 20(1-2):113–133, 1996.
6. Gideon Guy and Gérard G. Medioni. Inference of surfaces, 3d curves, and junctions from sparse, noisy, 3d data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1265–1277, 1997.
7. E.T. Jaynes. On the rationale of maximum entropy methods. *Proc. IEEE*, 70:939–952, 1982.
8. K.-R. Koch. *Parameter estimation and hypothesis testing in linear models*. Springer, 1988.
9. David G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.
10. J.D. McCafferty. *Human And Machine Vision*. Ellis Horwood, 1990.
11. Gerard Medioni, Chi-Keung Tang, and Mi-Suen Lee. *A Computational Framework for Segmentation and Grouping*. Elsevier, 2000.
12. Rakesh Mohan and Ramakant Nevatia. Perceptual organization for scene segmentation and description. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(6):616–635, 1992.
13. Björn Ommer and Joachim M. Buhmann. A compositionality architecture for perceptual feature grouping. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, number 2683 in LNCS, pages 275–290. Springer, 2003.
14. Jorma Rissanen. Minimum-Description-Length Principle. In S.Kotz and N.L.Johnson, editors, *Encyclopedia of Statistical Science*, volume 5, pages 523–527. John Wiley & Sons, 1985.