Mathematisches Forschungsinstitut Oberwolfach

Report No. 49/2001

The Mathematical, Computational and Biological Study of Vision

November 4th – November 10th, 2001

David Mumford, Jean-Michel Morel, Christoph von der Malsburg, Organizers

In spite of mathematic's widespread disdain for applications, it has found in them a consistent source of inspiration. This is particularly true for geometry in its relation to visual perception and visual imagination. That marriage having been consumed a long time ago, it may now be time for another love affair between mathematics and vision. To get it going was the intent and purpose of this meeting. The study of vision is now mainly a matter of neuroscience and numerical experimentation. The former, in the form of psychophysics, neuroanatomy and neurophysiology, is presently in a period concentrated on fact-finding with little emphasis on conceptual development. The latter, in its embodiment as computer vision, has been dominated awhile by an overdose of conceptualization, being prisoner to the prejudice that direct modeling in terms of geometry and physics should help to interpret images. This approach has been a protracted failure, explicit analytical modeling evidently being incapable of living up to the enormous variability of natural visual scenes.

How can mathematics come to the rescue? One important theme must be the mathematical structure that is rich enough to represent vision in all its aspects. Others are a probability metric on that structure, mechanisms to match stored entities to visual input and to each other, and methods to build up structures from input, that is, to learn. The visual modality can be evaluated in terms of a number of sub-modalities (color, motion, stereo, form, texture, to name the most important), none of which can be made to deliver reliable information at all times, and a theme of great importance is integration of sub-systems to exploit complementarities between them. All of these issues are tightly interwoven, and for a great while it will be highly recommendable to pay close attention to what neuroscience has to tell us. All of these themes were represented in the talks and discussions of the meeting, if to a large part in somewhat immature form, reflecting the state of development of the field. Thus, although some key invitees from the States didn't turn up as a consequence of the events of September 11, the meeting was rich, inspiring and full of promise.

Abstracts

Vanishing Point Detection using Helmholtz Principle

Andrés Almansa
(joint work with Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel)

We address the problem of detecting meaningful vanishing points in an image, without any a priory information. For this purpose we apply a general methodology for detecting geometric structures, recently developed by Morel, Moisan and Desolneux, which is based on the Helmholtz principle and Gestalt laws. [1]

As a primitive for our detector we use a set of line segments (edges) that have been in turn detected on the image using the same methodology and with a very limited number of false alarms. However, an improvement was necessary with respect to the work in [2] in order to suppress multiple responses for the same segment, due to blurred edges. It turns out that a "Minimum Description Length"-based criterion is as effective as a Canny-based criterion to choose the best among many meaningful candidates for a single edge. This allows us to significantly constrain our search space, leading to a computationally more efficient segment detector, which was obtained as a byproduct of multiple response suppression.

Then, under the assumption that the detected segments are uniformly distributed on the image plane, we consider a set of concurrent lines to be an ε-meaningful vanishing point, if the expected number of false alarms for this event is smaller than ε. As in the case of alignments we have to address the following issues:

- How to correctly define the event "at least n concurrent lines" in such a way that it takes into account the variable angular precision in the detected segments, while still alowing a simple computation of the expected number of false alarms.
- How to avoid spurious responses.

With respect to the the first point we choose a fixed multi-precision tiling of the image plane to express all possible vanishing points. Then, the second point is solved by only keeping those vanishing points which satisfy two optimization criteria:

- *Local minimum:* We only keep a vanishing point if its expected number of false alarms is a local minimum both in space and precision dimensions
- *Minimum description length:* We recompute the expected number of false alarms after requiring each segment to belong to only one vanishing point. Then we only keep those vanishing points which are still ε-meaningful after this operation.

According to our experiments, the proposed method is able to detect the most salient vanishing points in a scene, with a very low number of false alarms. We found the MDL criterion especially useful for reducing this number, by avoiding vanishing points that result from an accidental mixture of two different directions in 3D.

In outdoor or indoor urban scenes we typically detect the three main orthogonal directions and only these three, but unlike the work in [3] we don't need to introduce this hypothesis a priori, which requires knowledge of the camera's calibration parameters. On the contrary, the results obtain from our method toghether with this orthogonality hypothesis can be used to find some of the calibration parameters of the camera.

References

[1] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. Maximal meaningful events and applications to image analysis. Technical report, preprint CMLA No 2000-22, http://www.cmla.ens-cachan.fr/Cmla/Publications/2000/Abstract2000-22.html, 2000.

[2] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. Meaningful alignments. *International Journal of Computer Vision,*, 40(1):7–23, 2000.

[3] Evelyne Lutton, Henri Maître, and Jaime Lopez-Krahe. Contribution to the determination of vanishing points using hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(4):430–438, 1994.

Optimality of Coarse-to-fine search for different models of image interpretation

Gilles Blanchard

(joint work with Donald Geman)

"Coarse-to-fine" analysis for image interpretation, introduced by Fleuret and Geman (2000) in the context of face detection, has proven to be an accurate and computationally very efficient method to identify a target object (e.g. a face) in an image.

The Coarse-to-fine "paradigm" consists in two main building blocks. The first one is a tree-structured hierarchical representation of the set of target objects via a nested decomposition of the "space of poses", which is recursively divided into smaller subsets called cells. For each cell, a test is built which checks if one of the targets in the cell is present in an image, under the constraint that it has (ideally) a false negative rate of zero. Generally, the finer the cell, the smaller is the false positive rate of the associated test, but more computation power is required.

The second step is the design of a testing strategy which should determine for a given image whether a target object is present. A detection is confirmed if and only if there exists a complete "branch" of cells (from the root of the hierarchy to one of the finest cells) whose associated tests have given a positive answer. We consider sequential testing strategies, the test at step k being chosen as a function of the answers to the tests performed up to step $k-1$. A testing strategy stops once a "branch of ones" is found or when the performed tests rule out this possibility ("zero-blocking"). Note that a strategy can therefore itself be represented as a decision tree whose nodes are labeled by cells in the hierarcchy of the pose decomposition.

The coarse-to-fine approach emphasizes the computational power as the criterion for optimality: among all possible strategies, one wants to choose those for which the expected computation time is minimized. An intuitively natural family of strategies for that purpose is the set of coarse-to-fine strategies, which have the property that if the test at a given step yields a positive answer, the next test performed should be associated with a finer cell.

The results of Fleuret and Geman show that these strategies behave well in practical applications. Moreover, they show that the "computational burden" is not spatially uniform in an image: it is much more intensive in regions where a target object is present than in the "background". This bears a striking similarity with what can be observed in biological vision and studies of visual attention. It can be argued that the coarse-to-fine framework offers a model for visual attention which is both top-down (model-, or target-, based) and bottom-up ("uninteresting" zones are rapidly excluded by the coarser, and simpler, tests).

We studied the optimality of CTF strategies in the following formal setting. We consider a dual representation of the target objects: more precisely, we consider a set of attributes

$\mathcal{A}$ which is a disjoint union of set of attributes at different levels of resolution $\mathcal{A}_1, \ldots, \mathcal{A}_L$, where $\mathcal{A}_1$ represents the set of "coarse" attributes and $\mathcal{A}_L$ the set of "fine" attributes. To each attribute is associated a dedicated test which checks for the presence of the attribute in an image. In this representation, an "object" can be regarded as an element of $\prod_{i=1}^{L} \mathcal{A}_i$, i.e. a list of attributes.

We have considered various possible settings:

Target to identify:

- The target is a fixed element of $\prod_{i=1}^{L} \mathcal{A}_i$.
- The target is a fixed subset of $\prod_{i=1}^{L} \mathcal{A}_i$ and can be represented as the set of branches of a coarse-to-fine tree structure.
- Either of the above, except the target representation is not fixed but rather drawn at random.

Model for the unknown image:

- The unknown image is a random element of $\prod_{i=1}^{L} \mathcal{A}_i$. ("Guess who" game).
- "Background image model": every attribute in $\mathcal{A}_i$ has probability p_i to be present in the image and all attributes are independent. This model is supposed to represent what happens under the "null hypothesis" that *no* target is present in the image, which is what happens most of the time.

Computational cost:

- Unit costs (all tests have the same cost).
- Discrimination-based cost: the cost of testing for attribute $x \in \mathcal{A}_i$ is given by $c(x) = \Phi(1 - p_i)$, where Φ is a positive, increasing and convex function.
- Usage-based cost: the allocation of the costs for attributes $x \in \mathcal{A}$ are free under the "resource constraint" $\sum_{x \in \mathcal{A}} \exp -c(x) \leq 1$ (This model is perhaps most relevant for multi-threaded computing).

Jung (2001) proved the optimality of depth-first CTF strategy for discrimination-based cost and the background image model.

We were able to prove that the CTF strategy is optimal in various other combinations of the above cases, if it is required that the testing strategy finds *all* the "path of ones" in the case of a hierarchical representation of the targets.

References

Coarse-to-fine visual selection , F. Fleuret and D. Geman, *International Journal of Computer Vision*, january 2001, volume 41, number 1/2, pages 85-107.

Algorithmes de classification et de focalisation automatiques pour l'analyse d'images, Frank Jung, PhD. thesis.

Robust Segmentation for Computer Vision

Joachim M. Buhmann

Image segmentation is often defined as a partition of the pixels or image blocks into homogeneous groups. These groups are characterized by a prototypical vector in feature space, e.g., the space of Gabor filter responses, by a prototypical histograms of features or by pairwise dissimilarities between image blocks. For all three data formats cost functions have been proposed to measure distortion and, thereby, to encode the quality of a partition.

Robust algorithms for image processing are designed according to the following three steps: First, an appropriate definition of structure in images has to be defined. For segmentation these structures are formalized as pixel or pixel block partitions. Second, an

efficient optimization procedure to find good structures has to be determined. I advocate stochastic optimization methods like simulated annealing or deterministic variants of it which maximize the entropy while maintaining the approximation accuracy of the structure measure. Other optimization algorithms like interior point methods or continuation methods are equally suitable. Third, a validation procedure has to test the noise sensitivity of the discovered image structures. Statistical learning theory allows us (at least in principle) to calculate how much the quality of an image interpretation deviates from an interpretation of a second image with the same image content. Furthermore, statistical learning theory provides means to couple the image resolution scale to the approximation quality of the segmentation solution and the complexity scale of the model order selection problem, e.g., how many segments should be selected.

Object Recognition in Man and Machines

Heinrich H. Bülthoff

Theories of visual object recognition must solve the problem of recognizing 3D objects given that perceivers only receive 2D patterns of light on their retinae. Recent findings from human psychophysics, neurophysiology and machine vision provide converging evidence for image-based models in which objects are represented as collections of viewpoint specific local features. This approach is contrasted with structural-description models in which objects are represented as configurations of 3D volumes or parts.

I will report on recognition experiments which show strong viewpoint effects and speak in favor of an image-based representation of objects in which the physical similarity can account for recognition with small viewpoint changes. Recently, together with Guy Wallis we started to look at the importance of temporal similarity on the representation and recognition of objects. Temporal similarity can link many views of one object to one object identify, because different views of objects are usually seen in close succession. To test this hypothesis observers were presented sequences of unfamiliar faces in which the identity of the face changed as the head rotated. The observers showed a tendency to treat the views as if they were of the same person. Our results counter the proposal that object views are recognized simply on the basis of objective, structural components. Instead, they suggest that we are continuously associating views of objects to support later recognition, and that we do so not only on the basis of their physical similarity, but also their correlated appearance in time.

Morse Description and Morphological Encoding of Continuous Data

Vicent Caselles and A. Solé

The use of a topographic description of images, surfaces or $3D$ data has been introduced and motivated in different areas of research: image processing, computer graphics and geographic information systems. The motivation for such a description is different depending on the field of application but in all cases it aims to a description of the basic shapes in the given data and their topological change when varying a parameter relevant in each case (height in data elevation models, intensity in images,etc.). Such a description can be viewed as a practical implementetion of Morse theory. Morse theory describes the topological change of the isocontours of an scalar data or height function as the height varies, and relates these topological changes to the criticalities of the function. Given the scalar data u defined in a domain Ω of $I\!R^N$, the contour map has been defined as the family

of isocontours $[u = \lambda] = \{x \in \Omega : u(x) = \lambda\}$, $\lambda \in I\!R$, or in terms of the boundaries of upper (or lower) level sets $[u \geq \lambda] = \{x \in \Omega : u(x) \geq \lambda\}$ ($[u \leq \lambda]$). The first description is more adapted to the case of smooth data while the second description can be adapted to more general continuous data where there are plateaus of constant elevation data, or even discontinuous data. Some recent approaches using this second description are [1, 2].

The aim of this work is to deeply analyze Morse theory for the case of continuous functions in terms of its upper (lower) sets. As a result of this analysis a new simple algorithm for computing the Morse structure of an image has been developed. Essentially this algorithm is based on computing the maximal monotone sections of the upper (lower) topographic map. The definition for a monotone section is the following:

Let $u : D \to I\!R$ be a function. For each $\lambda, \mu \in I\!R$, $\lambda \leq \mu$ we define

$$U_{\lambda,\mu} = \{x \in D : \lambda \leq u(x) \leq \mu\}$$

Definition 1. *Let $u : D \to I\!R$ be a continuous function. A monotone section of the topographic map of u is a set of the form*

$$X_{\lambda,\mu} = cc(U_{\lambda,\mu}), \tag{1}$$

for some $\lambda, \mu \in I\!R$ with $\lambda \leq \mu$, such that for any $\lambda', \mu' \in [\lambda, \mu]$, $\lambda' \leq \mu'$ the set

$$\{x \in X_{\lambda,\mu} : \lambda' \leq u(x) \leq \mu'\}$$

is a connected component of $U_{\lambda',\mu'}$.

One can proof that under some assumptions the number of maximal monotone sections is finite. In addition, it is also proven that monotone sections can only contain a zonal maximum or minimum and that topological changes hold only at levels where a maximal monotone section begins or ends.

We have studied the special case of compressing data elevation models (DEM) as a possible application. In this terrain models it is also very important the creasenes structure (drainage patterns). There exists many algorithms to compute this creasenes structure but we have developed a simple morphological approach which provides us the information that one cannot recover from the Morse structure only. This morphological approach does not correspond exactly to the drainage patterns, in fact it can be viewed as a morphological sampling which recovers in some sense a set of non differentiable points. It has been proven that this sets of points are organized as curves and in fact this curves describe mainly the creases and valleys presents on the terrain.

Merging the information provided by the Morse structure and the morphological sampling one obtain a set of curves and points which suffices to interpolate the rest accurately using an adequate interpolator as the AMLE (Absolut Minimal Lipschitz Extension) model for example which is an excellent cone interpolator. As said, this structural sampling of the image is composed mainly of curves and a few isolated points (local maxima and minima mainly). Finally, these curves (and points) can be organized in trees using chain code based techniques and finally these trees can be coded by means of an efficient entropy coder such as an arithmetic coder. The L_∞ norm of the error between the original image and the coded one can be controlled by coding the errors which are greater than a specified one. In order to improve the results a multiscale approach has been also applied.

References

[1] C. Ballester, V. Caselles, P. Monasse, *The Tree of Shapes of an Image*, Preprint CMLA, 2001.

[2] J.L. Lisani, *Comparaison Automatique d'Images par Leurs Formes*, Ph.D Thesis, Université de Paris-Dauphine, July 2001.

Some recent results on the minimizers of the Mumford-Shah functional

Gianni Dal Maso

The *Mumford-Shah functional* in dimension n is defined by

$$F_\lambda(u,K) := \int_{\Omega\setminus K} |\nabla u|^2 dx + \mathcal{H}^{n-1}(K) + \lambda \int_{\Omega\setminus K} |u-g|^2 dx\,,$$

where Ω is a bounded open set in $\mathbf{R}^n$, $\mathcal{H}^{n-1}$ is the $(n-1)$-dimensional Hausdorff measure, $g\colon \Omega \to \mathbf{R}$ is a bounded measurable function, and λ is a nonnegative constant. The functional acts on pairs (u,K), with K closed subset of Ω and $u \in C^1(\Omega\setminus K)$. A pair (u,K) is said to be a *local minimizer* of F_λ in Ω if $F_\lambda(u,K) \le F_\lambda(u',K')$ for every (u',K') with $u'=u$ on $\partial\Omega$. The *Euler conditions* for local minimizers were found by Mumford and Shah (1985).

Some new results on the minimizers of the Mumford-Shah functional have been recently obtained by using a calibration method introduced by Alberti, Bouchitté and myself (1999). The first result has been proved by Mora and Morini (2000) in dimension $n=2$ for the case $\lambda=0$: if (u,K) satisfies the Euler conditions for F_0 in Ω, and if K is the union of a finite number of disjoint analytic curves, which are either closed or have their end-points in $\partial\Omega$, then for every $x_0 \in \Omega$ there exists an open neighbourhood Ω_0 of x_0 such that (u,K) is a local minimizer of the Mumford-Shah functional F_0 on Ω_0. In other words, under these regularity assumptions on K, the Euler conditions imply the minimality on sufficiently small domains.

The second result has been proved by Morini (2001) for arbitrary $n \ge 2$: if g is smooth out of a closed hypersurface M, on which g is discontinuous, then there exists a threshold λ_0 such that for every $\lambda \ge \lambda_0$ the functional F_λ has a unique absolute minimizer (u_λ, K_λ), and we have $K_\lambda = M$. In other words, under these assumptions on g the discontinuity set M is reconstructed exactly by the solution of the Mumford-Shah functional F_λ when λ is large enough.

The Computational Neuroscience of Visual Attention

Gustavo Deco

Experimental observations in functional imaging and single-cell recording provides strong evidences that attention modulates visual processing by enhancing the responses of the neurons representing the features or location of the attended stimulus and reducing the suppressive interactions of neurons representing nearby distractors. In this talk, we formulate a neurodynamical system consisting of interconnected populations of cortical neurons distributed in different brain modules which can be related with the different areas of the dorsal and ventral path of the primate cortex . We show that object recognition and visual search can be explained in the theoretical framework of a biased competitive neurodynamics. The top-down bias can guide attention to concentrate at a given spatial location or at given features. The neural population dynamics are handled in the framework of the mean-field approximation, i.e. by the analytical description of the mean activity of a population of neurons.

Detection of Geometric Structures in an Image by Helmholtz Principle

Agnès Desolneux
(joint work with Lionel Moisan and Jean-Michel Morel)

According to gestalt theory, grouping is the law of visual perception: whenever points or previously formed objects have a geometric characteristic in common (alignment, same colour, parallelism, etc.) they get grouped and form a new, larger, visual object, called a "gestalt". In the present work, we try to give a mathematical framework to this grouping phenomenon. We use a genericity principle, also called Helmholtz principle, which roughly says that we can do our probabilistic estimates as if the points were independent and had uniformly distributed characteristics. The main definition is then the one of ε-meaningful event: an event is said ε-meaningful if the expectation of the number of occurrences of this event in an image is less than ε. We apply this definition to different types of geometric events:
- alignment in an image (at each pixel of the image, we compute an orientation, and then consider the segments which contain "a lot of" points having their orientation aligned with the one of the segment, according to a given precision),
- boundaries and edges (closed level lines, or pieces of level lines of the image which have a high minimal contrast),
- alignments of points (equivalent to finding meaningful peaks in the Hough Transform),
- grouping objects according to their size, or orientation or grey-level,
- grouping points which are close.

In each case, we also define a notion of maximality (related in some sense to the "masking phenomenon" described by Gestaltists): an event is said maximal meaningful if it does not contain or is not contained in a more meaningful event.

Learning the Statistical Model of a Perceptive System in a Natural Visual Environment

Thomas Feldman

Our project aims at designing a low-level perceptive system which task should be to learn the background statistical model of its environnement.It is based on to principles : the first one is that the internal representation in the system should reflect the most essential local information about the environment in the sense of the Information Theory while preserving the low complexity of the system; the second one is that the system should mimic its environment by learning short-range interactions between responses to the environment across the system.

The system is given a grayscale images database, over which are learned second order statistics, in order to decompose each 12x12 patch image of the database according to their principal components. Thus we design around 10 P.C.A filters which role is to retain essential local information about the perceived images.

The marginal histograms of the responses to each filter are then computed. All but the first one can be coarsely quantified in 3 values, discriminating typical from rare responses to the filters. The internal representation is then designed by a redundant grid of columns of P.C.A filters followed by the coarse quantification mentioned above. It has been experimentally shown that this light representation is sufficient for reconstructing the visual environment.

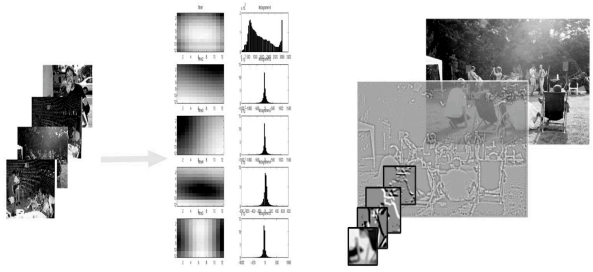

FIGURE 1. Marginal histograms of the responses to P.C.A. filters

FIGURE 2. Layers of the quantized representation

Until now, the system had only learned the marginal model of its response to its environment by selecting principal directions and quantification thresholds. We then learn short-range pairwise interactions across layers of filters. Each statistic between pairs of response cells located at neighbor spatial position and any vertical range is then computed. This exhaustive computation has been made possible by the coarse quantification that drastically reduce the number of pairwise statistics needed to infer the response model, according to Maximum Entropy Principle.

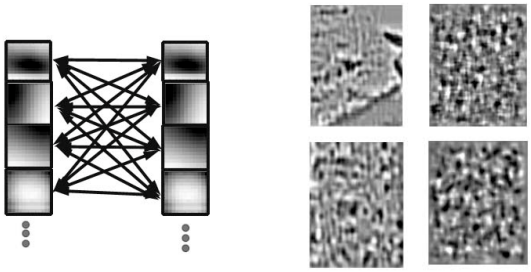

FIGURE 3. Local model of interactions

FIGURE 4. Right: original patches Left: sampled patches

The inferred model is a Gibbs Field over the grid of columns. This field is learned by maximizing the classical log-likelihood of the database by a stochastic gradient algorithm using Monte-Carlo Markov Chains. The model is then sampled in order to check the visual consistency of the system. Our experiments clearly show that it has learned a visually structured noise that present medium-range interactions despite its low complexity compared to classical visual models like Potts model.

This methodology allows us to naturally construct a visual noise model which accounts for a large part of the variability of images without containing any meaningful information. Considered as a background model, these noisy structures are to be removed from the image allowing oneself to focus on structures of interest. This procedure should be useful for robust indexing of image databases and automatic detection of objects in an image.

Inference for Vision

Bill Freeman
(joint work with Jonathan Yedidia, Yair Weiss, and Egon Pasztor)

Abstract: In order to interpret images, we need to propagate image interpretation information across space. A useful probabilistic model that allows this is a Markov network. We use synthetic examples to generate labeled training sets for two different problem domains: super-resolution (estimating a high-resolution imgae from a low-resolution one) and motion estimation (estimating projected optical flow from a pair of images).

We use belief propagation in a loopy Markov network to infer the scene estimates from the input image data. Yedidia, Freeman, and Weiss recently showed that fixed points of belief propagation correspond to local stationary points of the Bethe Free Energy, giving theoretical justification to this approach. I presented results for both the super-resolution and motion estimation problems using the same probabilistic machinery.

Web pointer for related papers: http://www.ai.mit.edu/people/wtf/publications.html

Neural Model for the recognition of complex biological movements

Martin A. Giese

The perception of biological movements plays an important role for the survival of many species. In spite of this fact, the underlying neural mechanisms are largely unknown. We have developed a biologically plausible neural model that accounts for a variety of experimental results from psychophysics, neurophysiology, and functional imaging. The model suggests that complex movements are neurally encoded in terms of prototypical examples of body configurations and optic flow field patterns in neurons in the *superior temporal sulcus*, and potentially the *infratemporal cortex*. The model shows that position- and scale-invariant recognition of such patterns can be accounted for by a hierarchical system of neural detectors with two pathways that analyze from and motion information, where invariance is achieved by nonlinear pooling of neural detector responses.

As one possible mechanism for the association of information over time the model postulates asymmetric lateral connections between high-level neural pattern detectors. The underlying neural dynamics can be described in an idealized form by the nonlinear integro-differential equation for the membrane potential $u(x,t)$:

$$\tau \dot{u}(x,t) + u(x,t) = \int_{-\infty}^{\infty} w(x-y) f(u(y,t)\,\mathrm{d}y + s(x,t) \tag{2}$$

with the asymmetric interaction kernel $w(x) \neq w(-x)$, and the translating stimulus $s(x,t) = S(x - vt)$, where v is a real constant. The threshold function f is monotonically increasing and bounded. Under appropriate conditions, we can show the existence of a form-stable traveling pulse solution of this equation, where the pulse propagates with the stimulus velocity v.

A Bayesian multiple-blob tracker

Michael Isard

This talk describes how a multiple-person tracker can be formulated as a problem in Bayesian sequential inference. A state-space is defined including the number of people in the scene as a discrete variable along with continuous variables describing the position

and shape of each person. A whole-image likelihood model is then described which relies on a static-camera assumption to build an accurate model of the background, and an efficient algorithm for evaluating this likelihood is demonstrated. Finally the talk shows that despite the highly non-linear nature of both dynamical model and likelihood model it is easy to implement the tracker using a particle filter to get good robust results.

Two Regularities in Visual Scenes: Statistical Interdependencies and Rigid Body Motion

Norbert Krüger

Vision faces the problem of an extremely high degree of vagueness and uncertainty in its low level processes such as edge detection, optic flow analysis and stereo estimation. However, the human visual systems acquires visual representations which allows actions with high precision and certainty within the 3D world under rather uncontrolled conditions. The human visual system can achieve the needed certainty and completeness by integrating visual information across modalities. This integration is manifested in the huge connectivity between brain areas in which the different visual modalities are processed as well as in the large number of feedback connections between higher and lower cortical areas.

The essential need for integrating visual information across modalities in addition to optimising single modalities has been recognised in the vision community after a long period of work on improving single modalities. The power of modality fusion arises from the huge intrinsic relations given by deterministic and statistic regularities across visual modalities, such as e.g., the coherent motion of objects or the high likelihood of the occurrence of collinear line segments in visual scenes. Two important regularities in visual data with distinct properties are (1) motion (most importantly rigid body motion, RBM) and (2) statistical interdependencies between features such as collinearity and symmetry. In contrast to RBM, the statistical interdependencies between features are much harder to describe analytically. Accordingly, developmental psychology shows strong evidence that visual experience plays an important role to achieve the ability to use these interdependencies in visual processing (e.g., the effect of illusionary contours appears after 5 month).

Collinearity and parallelism do not describe a deterministic relation between features but probabilistic relation, e.g., the occurrence of a line segment in visual data has a distinct impact on the likelihood of the occurrence of a line segments at a different position with different orientation (see, e.g. Krueger (1998). Collinearity and Parallelism are Statistically Significant Second Order Relations of Complex Cell Responses. Neural Processing Letters 8(2)). In my talk I address the statistics of natural scenes regarding additional modalities such as color or optic flow. As a main result it turns out that statistical interdependencies in visual scenes become significantly stronger when multiple modalities are taken into account. This result gives further evidence for the assumption, that despite the vagueness of low level processes stability can be achieved by integration information across modalities. Second, the attempt to model the application of Gestalt laws based on statistical measurements, as suggested recently by some researchers (e.g., Geisler, Elder, Krueger, Sigman) gets further support. Third, the results in this paper suggest to formulate the application of Gestalt principles in a multi-modal way. Finally, as a by-product of our simulations it turns out that edge–like structures are more dominant compared to line–like structures in intrinsically one-dimensional image patches.

Visual localization in the presence of saccades and motion

Markus Lappe

Estimating the location of an object in visual space becomes difficult when motion is involved. Because of latencies in the visual system the position of the object might have changed by the time its visual image is processed in the cortex. This happens irrespective of wether it is the object that moves, or the eye (e.g. by a saccade). My talk describes illusory mislocalisations of briefly presented objects in these situations and discusses their implication for the mechanisms of dynamic visual space representation in the brain.

The first part of the talk is concerned with errors in the localization of a moving object. In the so-called flash-lag-effect, a stroboscopically illuminated moving object appears to lag behind a continuously lit moving object when both are physically aligned. Originally this has been interpreted as a predictive component in the perception of the continuously moving object. More recent studies instead suggested a delayed processing of the flashed object. As a third alternative, I present a model of the flash-lag effect that is based simply on an extended temporal averaging of the position between the two objects, thus involving again a relative distance measure. Predictions of the model, among them a flash-lead effect for certain parameter combinations, are corroborated in experiments.

The second part of the talk is concerned with errors in the localization of visual stimuli that are flashed shortly before or during a saccadic eye movement. When this is done in darkness without other visual cues present, the flash appears shifted in the direction of the saccade. In contrast, when visuospatial references are available the flash is mislocalized towards the saccade target, implying a compression of the metric of space by the saccade. The first type of error ('shift') is attributed to a mismatched time course of the extraretinal signal that accompanies the saccade. It is an error in the absolute judgement of position in space. The second type of error ('compression') appears to involve a misperception of the distances between the flash and other visual objects, the judgement based on relative, retinal signals.

The role of feedback in visual perception

Tai Sing Lee

Simon Thorpe showed that when an image was flashed on the screen for only 20 msec, both human and monkey subjects can grasp the gist of the scene very rapidly and react with a mean reaction time of 220 msec. This leaves very little time for interactive computation to happen across the visual hierarchy, suggesting that recognition and categorization might happen primarily in a feedforward manner as the first volley of spikes thundering through the brain. Many existing face detection computer vision systems indeed can operate in this way, and are pretty successful in detecting faces using maximum likelihood test simply on the statistics of Gabor filter responses. I would suggest the recognition does not necessarily mean perception. When the subjects reacted to the images, they might be acting on a subliminal level, without conscious perception, as in blind sight. Visual processing might best be understood as a two-stage process: The first volley of spikes stimulates the memory areas to generate hypotheses about objects in a visual scene. These hypotheses are then fed back to the early visual areas to impose the contextual priors to guide perceptual processing, or in Andrew Blake's words, cleaning up the details. In contrast to Marr's model, this view suggests recognition precedes perception, rather than the other way around. The contextual priors are communicated top-down through the visual

hierarchy with the massive recurrent feedback connections. They made robust perceptual inference possible. I reviewed some of the neurophysiological literature on the neural basis of perception, in particular, the works of Logothetis, Newsome, Desimone, Motter and Lamme. I also reported some of my own data on evidence of feedback modulation on early visual processing.

Neural Implementation of Figure-Ground Segmentation

Christoph von der Malsburg

Opening remark 1: A solid experience of computer vision is that individual functional components, such as shape from shading etc., cannot be made to act reliably in natural visual environments. The goal must therefore be to integrate subsystems with each other. Thus, segmentation of scenes will only start to work if several cues, among them commong motion, texture, contour, color, stereo and known form, are coupled to help each other solve the problem. Important issues in systems integration are the determination of momentary relevance, confidence levels, and appropriate interfaces.

Opening remark 2: I gave a coarse overview of the visual system, especially its cortical organization in terms of representation areas, columns and hypercolumns and their fibre connection patterns.

Opening remark 3: The binding problem. The classical view has it that the brain represents things in terms of elementary symbols, corresponding to individual neurons. This raises the problem that simultaneously active sets of neurons have, according to that view, no means of keeping themselves separate from each other. I discussed the need to introduce dynamic links, giving the brain dynamic graphs as data structure. An elementary way to represent links is by signal synchronization and rapidly switching synapses. A more powerful implementation uses multicellular nodes which activate subsets of appropriately connected neurons in order to dynamically activate the desired links.

Figure-ground segmentation: The basic idea is that neurons that belong to the figure are to be synchronized with each other, and similarly for the neurons belonging to the ground. To achieve this end, feature-representing neurons in the visual cortex are connected positively if they are likely to be part of the same figure: $P(i,j) > \theta \Rightarrow T_{ij} > 0$, or $T_{ij} \leq 0$ otherwise. Here, i, j stand for neurons, $i \sim j$ means that i and j belong the the same figure, $P(i \sim j)$ gives the probability thereof, and T_{ij} is the strength of connection between neurons i and j. The gestalt laws can all be implemented in this way in terms of neural connection strenghts. When lumping together all neurons activated from one point in visual space, that is, all neurons belonging to the same hypercolumn, one can define W_{lm}, the combined strength of connection between points l and m. Thus, all segmentation cues are integrated into the quantity W_{lm}. Segmentation can now be formulated in terms of a set of differential equations describing neural signals, where strong connections conspire to create signal correlations within the figure and within the ground, and anticorrelation between them. An alternative description of signal dynamics is in terms of the "energy" function $E = -1/2 \sum_{lm} T_{lm}\sigma_l\sigma_m$, where $\sigma_i \in \{1, -1\}$ are labels ("spins") for figure and ground, resp., and the probability of a global label distribution $\{\sigma\}$ is described by a Gibbs distribution $P(\{\sigma\}) = (1/Z)\exp(-\beta E(\{\sigma\}))$.

Introduction to Gestalt theory

Jean-Michel Morel

In this brief introduction to a monumental group work (1923-75), I have outlined the aims, tools and results of the so called geometric Gestalt. Founded by Wertheimer, this phenomenological methodology is based on presentations to subjects (mostly humans, but also animals) of geometric figures. It tries to track the geometric "organizing laws" by which points of the retinal perceptum end perceptually grouped into organized entities. The organizing laws are of a geometric nature (alignement, closeness, closedness, parallelism, similarity of shape, constant width, convexity, symmetry...) and are at work in the perception of any image. The geometric laws mostly collaborate in the formation of Gestalts and this led me to define them as "partial gestalts", in opposition to the global gestalts. The perception of global gestalts is somewhat opaque, in that by the "Gliederung" law, "only parts of the whole are visible, which contribute to the overall perceived organization of the whole". For instance, the perceived parts of a square are its sides, its corners, and nothing else.

The "masking" phenomenon is a consequence of the Gliederung (articulation whole-parts) and is illustrated in spectacular and simple experiments, some of which were thoroughly discussed in an evening session. The gestaltist's method is extremely clever, in that he uses a wide variety of "ungünstigen Bedingungen" (unfavourable observation conditions) like darkness, distance, short exposition, lateral vision to enforce the prevalence of organizing laws against the influence of the presented image. The perceived image is driven by geometric laws towards a much more regular pattern than the presented one and the comparison illustrates the relative strength of the various "partial gestalts".

Of course, during my exposition, Computer Vision was the aim and the masked partner. The aims of Computer Vision are exactly parallel to the ones of Gestaltists : to define organizing laws for detecting patterns in an image. The method is equally parallel : in classical computer vision, one tries to define "features", which clearly are the computational counterpart of partial gestalts.

The understanding of Gliederung remained as widely open as the problem in Computer Vision of the global understanding of a digital image. The gestaltists ended in somewhat byzantine experiments on conflicts between partial gestalts. These experiments are too particular, with too many partial gestalts acting together. The conclusions drawn from such experiments remained uncertain. Clearly, computer vision is the logical continuation of gestaltism and permits to develop a new experimental device : the computation of partial gestalts and, hopefully, to launch the search for mathematical principles, probably of a variational type, giving an account of the Gliederung. I suggested as a partial realistic aim the scanning and automatic analysis of the gestalt figures. A lively discussion ensued.

Speculation on the modeling of cortex

David Mumford

This talk presented a set of issues involving what possible neural mechanisms may solve a series of different computational issues. I began by reviewing a set of new physiological results, from the last 5 years, which suggest that the substrate of neural computation, the neurons and their local circuits, may follow very different principles from standard neural nets. These were the results of Markram and Abbott on the complex dynamics of single synapses, the results of Larkum, Zhu and Sakmann on back propagating Ca++ spikes and

the results of Connors on gap junctions between inhibitory neurons. Then the spike decoding problem was addressed and the suggestion was made that a spike train is not a message to be decoded but a tiny fragment of a dance being carried out by an extended assembly. This was reinforced by recalling that cortex has no "fire-walls": 50pyramidal output is sent immediately to distant areas. The binding problem was discussed and a formulation for dealing with this, "Mixed Markov Models", due to A. Fridman was described. The rest of the talk dealt with the proposal that a fundamental problem for cortical computation is the absence of "registers" or "caches", places to tuck current percepts, ideas, plans, hypotheses while activity progresses. I call this the "2 idea problem": how can a column maintain ¿ 1 states at once, one being the immediate one but the other being a previous state not yet understood, or a hypothesis or one of several interpretations not yet disambiguated. It was suggested that the LTS inhibiting neurons of Connors might put assemblies of cells into "idle" mode, which could be later reactivated.

Affine-invariant shape recognition

Pablo Musé
(joint work with Frédéric Sur and Jean-Michel Morel)

1. Affine-invariant coding of a shape

- **Smoothing** : *affine curve shortening* is used : $\frac{\partial x}{\partial t} = |\text{curv}(x)|^{1/3}\,\overrightarrow{n}$, where x is a point on the border of the shape, $\overrightarrow{n}$ the normal vector to the curve at this point (pointing towards the concavity) and $\text{curv}(x)$ the curvature. See [1] for a fast algorithm.
- **Local codings** : bitangency and parallelism are invariant features, so frames are defined with bitangents and tangents at inflexion and "flat" points. In each frame, the curve is affine-invariant normalized, and a piece of it is described by a "word" (a regular subsample of N points). See [2].

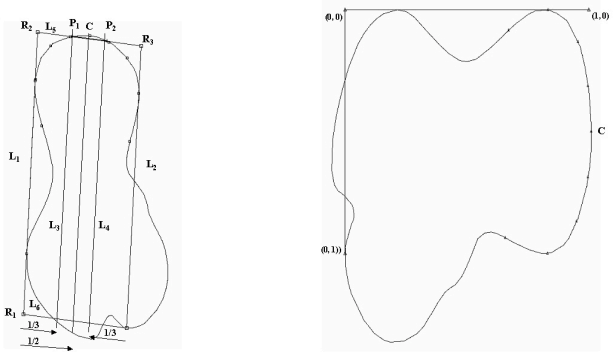

Figure 5. On the left : affine-invariant frame on the original curve. On the right : normalized curve. The N points lie on both sides of the point C.

- **Registration** : a dictionary is built for each curve belonging to the database.
- **Query** : only words encoding the database which are similar to the query are kept and considered as pre-matchings. Then the mappings between pre-matchings are estimated and pre-matchings are extended. Real matchings are chosen to be long enough extensions.

2. Experimental results

- **logo recognition** : the considered shapes are *maximal meaningful level lines*. See [3] and [4] for definitions and algorithms.

Figure 6. Matching lines between left and right images are colored. This mapping is not affine. Nevertheless a projective mapping can be locally understood as an affine one.

- **further developments** :
  - image recognition in a huge database.
  - meaningful shapes : where is the information ?

References

[1] L. Moisan. Affine plane curve evolution: a fully consistent scheme. *IEEE Trans. on Image Processing*, 7(3):411–420, 1998.

[2] J.-L. Lisani. *Comparaison automatique d'images par leurs formes*. PhD thesis, Université Paris IX Dauphine, France, 2001.

[3] A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by Helmholtz principle. *IJCV*, 2000.

[4] J.-L. Lisani, L. Moisan, P. Monasse, and J.-M. Morel. Shape recognition algorithm robust under partial occlusions and affine deformations. In *proceedings ISMM 2000*, pages 91–98, Palo-Alto, 2000.

Variable Lifting and Mathematical Constraint Modeling in Computational Vision

Christoph Schnörr

(joint work with Daniel Cremers, Jens Keuchel, and Christian Schellewald)

Introduction. Mathematical approaches to the design of computer vision systems vary considerably from level to level in the processing hierarchy. At the signal level, it is convenient to work in vector spaces. At higher levels, on the other hand, there is no natural order of visual primitives extracted at the signal level. This gives rise to intricate combinatorial constraints related to partitioning, grouping, and matching of these primitives. A natural question therefore is: How can these constraints mathematically be represented (i) such that the model is more compatible to those applied at the signal level, and (ii) such that computationally efficient implementations can be derived?

Another import issue concerns the representation of knowledge at low– and mid–level processing stages. How can intricate constraints be learned (from visual data) representing knowledge which is relevant for various visual tasks?

Several research projects in our group are concerned with the study of these problems in various specific contexts. These projects are sketched in the following two sections. The primary underlying mathematical theme which is well known in pattern recognition since decades [1], may be summarized as follows:

: *Intricate sets of feasible solutions have simpler descriptions in higher dimensional spaces.*

This fact is of crucial importance for applying both statistical learning theory [2] and advanced optimization strategies [3] to computational vision.

Statistical shape learning and variational segmentation. The objective of this project is to model statistical learning of outlines of 3D-objects in an unsupervised way for the purpose of variational segmentation. To this end, vector representations $x_{\mathcal{C}}$ of contours $\mathcal{C}$ of the sample set are transformed by a mapping ϕ into a high–dimensional feature space $\mathcal{F}$ using kernels K which satisfy the Mercer condition [2]: $(\phi(x_{\mathcal{C}}), \phi(x_{\mathcal{C}'}))_{\mathcal{F}} = K(x_{\mathcal{C}}, x_{\mathcal{C}'})$. By statistical decorrelation and compression in the feature space a nonlinear potential in the original shape space is obtained which represents familiar shapes. This representation is used in combination with a modified version of the Mumford and Shah approach for variational image segmentation (Fig. 7).

Numerical experiments show that this model is able to model real 2D–shape variations of projected views of 3D–objects as well as to discriminate the views of different objects in an unsupervised way. The nonlinear shape statistics makes the approach robust against clutter. The variational approach makes the approach robust against initialization, local minima and noise.

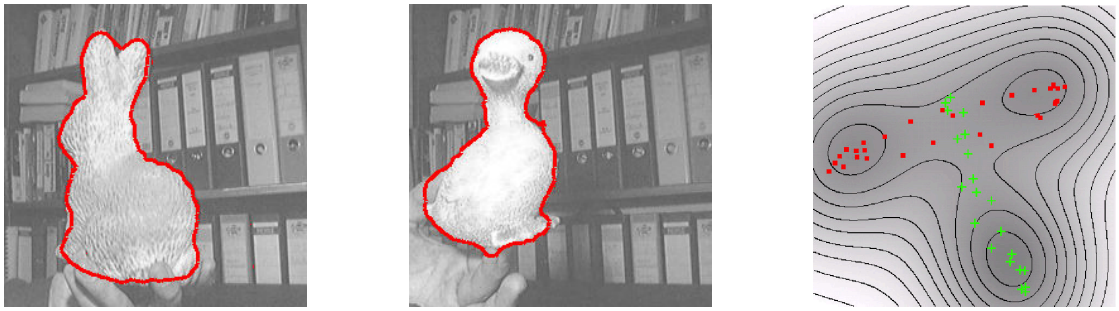

FIGURE 7. Segmentation (left, mid) and unsupervised representation of visual shapes (right).

Convex relaxation of problems of mid–level vision. The objective of this project is constraint modeling and mathematical relaxation of difficult combinatorial problems of mid–level computational vision like image partitioning, perceptual grouping and graph matching. Mathematically, these problems can be represented as instances of the following optimization problem: $\inf_{x \in \Omega \cap S} J(x)$, where Ω models a set of indicator variables and $S \subset \mathbb{R}^n$ represents further constraints depending on the problem instance.

Computationally tractable approaches are obtained by relaxing the Langrangian dual of these optimization problems and solving a convex optimization problem in a higher–dimensional matrix space (Fig. 8). By this, configurations of visual primitives are embedded into a high–dimensional vector space along with tight approximations of the combinatorial constraints which however, are much more convenient from the optimization point-of-view. Numerical results for established benchmark problems [4] show the remarkable performance of this approach.

FIGURE 8. Image partitioning (left) and perceptual grouping (right) by convex programming.

REFERENCES

[1] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electronic Computers*, 14:326–334, 1965.

[2] V.N. Vapnik. *Statistical learning theory.* Wiley, New York, 1998.

[3] L. Lovász and A. Schrijver. Cones of matrices and set–functions and 0–1 optimization. *SIAM J. Optimization*, 1(2):166–190, 1991.

[4] R.E. Burkard, S. Karisch, and F. Rendl. Qaplib – a quadratic assignment problem library. *J. Global Optimization*, 10:391–403, 1997.

Algebraic Embedding of the Perception-Action Cycle

GERALD SOMMER

Starting with a DFG project in 1996, our research group established a new direction of research which has been attracting more and more attention in the community ever since.

The Perception-Action Cycle

Vision is too hard a task to be performed successfully by inactive systems. A perception-action cycle is the representational framework for perception and action within the behaviour based paradigm for designing competent systems. A PAC cannot be designed by separating the problem into perception and action tasks. Instead, the interaction between perception and action has to be an integrate part in the design of either task.

The Geometric Algebra (Clifford Algebra) as Embedding Frame

We believe that the choice of representations is crucial to obtain life-like system behaviours. The different scientific disciplines which address the different aspects of a PAC, largely work with completely different and at times unsatisfactory representation schemes. The preferred representation schemes are based on vector algebra. Typically, important transformations cannot be expressed by linear operations on vectors. Instead, non-linear transformations are often approximated by an iterative application of appropriate linear transformations. A further drawback of vector algebra is that higher order geometric entities like lines, planes, circles, spheres, have no compact, linear representation. Consequently, linear transformation operators of such higher order entities do not exist, as well. We propose the use of geometric algebra (GA) as introduced by D. Hestenes, instead of vector algebra. Geometric algebra belongs to a certain class of Clifford algebras. In GA higher order geometric entities can be defined in a compact form. Furthermore, linear operators are available for transformations that in vector algebra are non-linear. These operators can also be applied to any type of geometric entity expressible in GA and not just to vectors. One advantage of GA that follows from these properties is that real-time capability is more likely to be achieved. Another advantage is that GA enables us to express geometric entities we are interested in, directly as algebraic objects. This improves the geometric

insight into a problem dramatically and leads more easily to the sought for solution. In the following some applications of GA to computer vision problems that were developed in our group are presented.

Problem: Missing Linear Theory of Multi-dimensional Signals

Theoretically, the methodology of contemporary image analysis is based on one-dimensional signals. The topologically new quality of multi-dimensional signals cannot be adequately modelled in complex algebra. Furthermore, intrinsically multi-dimensional structures can only be detected by non-linear operations. These facts are related to the missing definition of a multi-dimensional phase as the geometrically relevant feature because of limited possibilities of representing symmetries in the complex domain. In principle, in Clifford harmonic analysis the way to overcome these problems is outlined. We could show how to extent the concepts of the Fourier transform, the Hilbert transform and the analytic signal by adequate embedding in a geometric algebra. For 2D signals the Riesz transform generalizes the Hilbert transform and the monogenic signal generalizes the analytic signal. This is derived from the 3D Laplace equation. In that frame, images are vector fields and operations are spinors. Interestingly, the third coordinate corresponds to a scale parameter. It could be shown that the corresponding linear scale-space is a real alternative to the Gaussian scale-space. Quadrature filters can be constructed for intrinsically 1D signals from orders 0 and 1 of the spherical harmonics and for intrinsically 2D signals from orders 2 and 3. In total, a set of seven spinor-valued filters can be applied by convolution. Thus, instead of the commonly known local features (energy and phase) in 1D signals, a multivector of seven orthogonal features exists for 2D signals.

Problem: Limited Bias of the Real Perceptron

The universal approximation property of MLP nets is reduced in its importance if the neurons only accept real data. The computed scalar product results in a bad balance of bias/variance in the case of noisy data. In other words, in real vector space there are too few constraints for a successful separation of any intrinsic variance of data from noise contributions. A Clifford neuron, on the other hand, operates in Clifford (or geometric) algebra. By the nature of the chosen product the resulting linear space is presenting a rich subspace structure which constrains learning. Another interpretation of the advantageous behaviour of Clifford neurons is related to Clifford groups which are induced by the chosen algebra and which result in usefull constraints. We could show that a single Clifford neuron learns geometric transformations which only can be learned by several real neurons (or even not at all, as in the case of the Moebius transform). Not only the computational ressources are reduced. Clifford neurons act as linear operators in the algebraically deformed space. Because of their group-based constraint they give a far better generalization of noisy data than real neurons.

Problem: Stratification of Geometry in Computer Vision

O. Faugeras proposed a stratification hierarchy of projective, affine and metric spaces an observer should access depending on his/her situation and embodiment. Up to now all these strata have only been used in a few cases. Pose estimation, for example, can and has been formulated and solved as a projective, metric and kinematic task. Changing from one stratum to another is not well understood so far. Especially if instead of points higher order geometric entities (e.g. lines) are used.

By embedding the 2D/3D pose estimation problem into conformal geometric algebra, we could develop an algorithm which uses all three strata simultaneously. This was possible because within conformal geometric algebra we can express kinematic transformations, projections and metric measurements.

Note that the basic observable entities whose kinematic parameters can be linearly estimated are not only points, but also spheres, circles, planes, and lines in 3D space. As a direct application we developed an algorithm that linearly estimates the pose of articulated objects as kinematic chains. This algorithm runs at video rate real-time. The pose estimation can be performed based on different geometric entities, also taking into account the relative trustworthyness of different measurements. The algorithm works by estimating the effect of motors (multiplicative spinor representations of rotation and translation) on geometric entities such that a distance measure of a constraint in Euclidean space is minimized. But instead of using the Lie group representation of motors as spinors, we use their twist representation and perform the pose estimation in a Lie algebra.

Ultra-Rapid Visual Processing – Computing with one spike per neuron

Simon Thorpe

Monkeys and humans are very fast and accurate at deciding whether previously unseen natural images contains a target category (ex. animal). Indeed, processing is so fast that it appears to rule out many popular models of visual processing and coding in the nervous system. In particular, I will argue that such tasks must be possible under conditions where neurons in any particular processing layer may only get to emit one spike before the neurons in the next layer have to make a decision. I propose that one solution to this dilemma is to use the relative ordering of spikes across a population of neurons to encode information, rather than a conventional firing rate code. Computer simulations show that this approach is not only viable, but that systems using a single wave of asynchronous spikes can out-perform many of the conventional image processing techniques used in computer vision.

Variational principles for second order functionals

Franco Tomarelli
(joint work with Michele Carriero and Antonio Leaci)

We focus the Blake & Zisserman functional in image segmentation.

$$F(K_0, K_1, u) := \int_{\Omega\setminus(K_0\cup K_1)} \left(|D^2u|^2+\mu|u-g|^q\right) d\mathbf{x}+\alpha\mathcal{H}^{n-1}(K_0\cap\Omega)+\beta\mathcal{H}^{n-1}((K_1\setminus K_0)\cap\Omega) ,$$

where $\Omega \subset \mathbf{R}^n$ is an open set, $n \geq 2$, $\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure and

$$q > 1 \ , \ \mu > 0 \ , \ 0 < \beta \leq \alpha \leq 2\beta \ , \ g \in L^q(\Omega)$$

are given; while $K_0,\ K_1 \subset \mathbf{R}^n$ are Borel sets (a priori unknown) with $K_0 \cup K_1$ closed, u is approximately continuous on $\Omega \setminus K_0$ and $u \in C^2(\Omega \setminus (K_0 \cup K_1))$.

If the triplet (K_0, K_1, u) is a minimizer and $n = 2, 3$ then $K_0 \cup K_1$ may be interpreted as an optimal segmentation of a monochromatic image of given intensity g.

We review sufficient conditions for existence of minimizing triplet, quantitative and qualitative properties of such triplets and some results about numerical approximation. Moreover we show necessary conditions fulfilled by minimizers, obtained by many types of variations, and we explicit Euler conditions of integral and geometric type fulfilled by the optimal segmentation.

Existence of minimizers is proved by regularizing weak solutions when $n = 2$ and $g \in L^{2q}_{loc}(\Omega)$.

The Euler equation in the distribution sense, outside optimal segmentation $K_0 \cup K_1$, is

$$\Delta^2 u \;=\; -\frac{q}{2}\mu|u-g|^{q-2}(u-g) \quad \text{in } \Omega \setminus K_0 \cup K_1 \ ,$$

coupled with homogeneous conditions for natural boundary operators associated with the decomposition to the bi-harmonic operator.

First variations of the energy with respect to compactly supported deformations of the optimal segmentation provide an additional global Euler equation and a link between curvature of the segmentation and the jump of the traces of hessian matrix.

We exhibit a non-trivial triplet satisfying all the necessary conditions proved for the main part of the energy and a variational principle of equi-partition of volume and surface energy.

We conjecture that such triplet is a local minimizer, unique up to sign change, rigid deformations of co-ordinates and/or addition of affine functions.

Learning similarity metrics between shapes

Alain Trouve

The design of good features and good similarity measures between features play a central role in any retrieval system for searching a database. The use of *metric* similarities (ie coming from a real distance) is also very important to allow a fast retrieval on large databases. Moreover, these similarity functions should be flexible enough to be tuned to fit some users model. These two constraints, *flexibility* and *metricity* are generaly difficult to fulfill. Our contribution is two folds: We show that the kernel approach introduced by Vapnik, can be used to generate metric similarities, especially for the difficult case of planar shapes seen in a rotation invariant way. Moreover, we show that much more flexibility can be added by non rigid deformation of the induced feature space. Defining an adequate Bayesian users model, we describe an estimation procedure based on the minimization of the underlying log-likehood function.

Inference in Markov Random Fields using Belief Propagation

Yair Weiss
(joint work with Bill Freeman and Jonathan Yedidia)

Inference in Markov Random fields is typically exponential in the number of nodes. For singly connected graphs, the calculations can be done efficiently using a simple, parallel algorithm called "belief propagation". This same algorithm can also be applied to multiply connected graphs. Such "loopy belief propagation" was thought to be a bad idea until the dramatic empirical successes of Turbo codes and other applications. Recently, we have been able to shed light on this by a number of analytical results.

Unsupervised Learning of Invariances in a Simple Model of the Visual System

Laurenz Wiskott

A new algorithm for unsupervised learning of invariances is presented. The basic idea is to learn a nonlinear input-output function which extracts slowly varying aspects from the input signal by minimizing the temporal variation of the output signal. This is a known approach. The algorithm, however, differs from existing learning rules. Firstly, it computes

the solution in a closed form (like PCA) and is guaranteed to find the optimum within the considered function class. Secondly, not only one but many uncorrelated output signal components can be generated easily, which is important for hierarchical networks.

The algorithm is applied to a simple model of the primate visual system with a one-dimensional retina. Depending on what stimuli are used for training, the network can learn translation-, scale-, rotation- (cyclic shift), contrast-, or illumination-invariance. Relatively few stimulus patterns are needed for training to achieve good generalization to new patterns. The representation generated is suitable for pattern recognition. Overall the model suggests that it may be plausible that our visual system learns invariances based on fairly limited visual experience.

Representing Images by Gabor Wavelet Transform Magnitudes

Ingo Wundrich

Several object detection and recognition approaches rely on Gabor responses as their representation in order to obtain point-to-point correspondences between the input image and the object model. Similarity functions constructed from the *magnitudes* of the Gabor wavelet provide a much smoother similarity function which can be optimized in a reduced resolution. This approach turned out to be quite powerful in several object detection and recognition frameworks. Despite its success in such application domains the more profound question arises wheter these magnitudes retain all the image information without introducing ambiguities.

The transition from the $\langle I, \psi_{\vec{n}_0,m,l}\rangle$ to the $|\langle I, \psi_{\vec{n}_0,m,l}\rangle|$ image representation is supported by a collection of theorems about magnitudes of the Fourier transform $|\mathrm{FT}I|$ stated by Hayes. The major result concerning the magnitudes is that almost all images defined on an $N_1 \times N_2$ support can be represented by $|\mathrm{FT}I|$ sampled at $(2N_1 - 1)(2N_2 - 1)$ points uniquely up to the sign and a point reflection within this support. The exceptions from this statement are of measure zero. For an application to the subband images of the discrete Gabor wavelet transform the first step to take is to drop the assumption of real-valued input to the transform from which we are about to take the magnitudes.

Theorem 1. *Let $\mathcal{B}(N_1, N_2)$ be the space of all bandlimited functions on the finite support $\{0, \ldots, N_1 - 1\} \times \{0, \ldots, N_2 - 1\}$ such that $\mathrm{DFT}I(\vec{\rho}) = 0$ for $|\rho_1| \geq \frac{N_1}{4}, |\rho_2| >\geq \frac{N_2}{4}$, and let the wavelet family $\psi_{\vec{n}_0,m,l}$ constitute a frame in $\mathcal{B}(N_1, N_2)$. For all $I_1, I_2 \in \mathcal{B}(N_1, N_2)$ such that $\langle I_1, \psi_{\vec{n}_0,m,l}\rangle$ and $\langle I_2, \psi_{\vec{n}_0,m,l}\rangle$ are only trivially reducible polynomials and $|\langle I_1, \psi_{\vec{n}_0,m,l}\rangle| = |\langle I_2, \psi_{\vec{n}_0,m,l}\rangle|$ $\forall \vec{n}_0, m, l$ it follows that $I_1(\vec{n}) = \pm I_2(\vec{n})$.*

In fact we do not lose more than the global sign information if we ignore the phases of complex Gabor responses in every subband. The price one has to pay for it is the presumed oversampling of the input image.

Motivated by existing Fourier phase retrieval algorithms an iterative procedure is constructed having the magnitude subband images and one *arbitrary* image as inputs. The latter provides its Gabor phases to be combined with the magnitudes. After less than 1200 iterations one gets images retaining all the structural information of necessity for robust object detection/ recognition.

Gabor phase space molecules for image understanding

Rolf P. Würtz

Exemplar-based object recognition algorithms usually consist of the following steps.

1. Extraction of "atomic" visual primitives.
2. Combination of these to higher order "molecular" structures.
3. Estimation of correspondence maps ("matching") between the actual visual scene and stored object prototypes.
4. Organization of the memory of object prototypes such that matching and comparison can be efficient.
5. Filtering of the video stream from the camera

Gabor wavelet transforms provide a rich and convenient description of an image. They are atomic in the sense that they subdivide the 4-dimensional phase space spanned by all possible combinations of these parameters into cells of minimal volume. They provide a good model of simple cells in V1. The precise form of the function is not crucial, but the matched filter property (positivity in frequency space) turns out to be very helpful. It makes it very natural to describe the atoms in terms of a *local amplitude* (model for complex cells) and *local phase*. The former has very favorable properties concerning matching robustness, the latter is required for precise localization of correspondences.

Successful examples include

Jets: used for face recognition by elastic graph matching;
Minijets: used for face recognition by Gabor pyramid matching;
Graphs and pyramids: as structures coding for whole object aspects;
Correspondence structures: the intermediate and final results of matching;
Endstopped cells: a model for a special kind of cells in the visual areas V1 and V2.
Corner detectors: a multiscale combination of endstopped cells
Line elements: connection structures which support the Gestalt rule of *collinearity*;
Texture operators: used for the classification of natural textures.

I have described in detail the methods of Elastic Graph Matching, its extension to bunch graph matching,and its embedding into a real-time recognition system. Conceptual problems with the background have been solved by the method of Gabor pyramid matching, which has a detailed neuronal implementation. Finally, I gave an outlook on further applications of the concept.

Early Vision, Cortical Columns, and the Tangent Bundle

Steven W. Zucker

The visual cortex in primates is organized around orientation, with "columns" of cells exhibiting receptive fields selective for different edge and line orientations at each retinotopic position. We identify the orientation column with the unit tangent bundle, $R^2 \times S^1$, and consider the question of how to structure early vision in it. A differential-geometric position is adopted, which requires specifying the connection forms. We develop these analagously for curve detection, for stereo corresondence, and for texture-flow and shading analysis. In curve detection we interpret noisy edge "detector" responses as putative tangents to curves, and transport along the osculating circle to enforce‘ consistency. For texture-flow, frame transport requires two curvatures (the connection form evaluated in the tangential and the normal directions) and a helicoid in $R^2 \times S^1$ is the osculating object.

For stereo tangents to a space curve are transported along an osculating helix, and the Frenet 3-frame is projected to obtain two monocular problems. The result, for stereo, is a set of compatibilities in $(R^2 \times S^1) \times (R^2 \times S^1)$ that compute both spatial and orientation disparities. Mechanisms for estimating the relevant curvatures were described, and relaxation labeling (equivalent to a class of polymatrix games) supports all computations. Examples of all computations were shown.

Edited by Ingo Wundrich

Participants

Prof. Dr. Andres Almansa
almansa@cmla.ens-cachan.fr
CMLA
ENS Cachan
61, Avenue du President Wilson
F-94235 Cachan Cedex

Prof. Dr. Andrew Blake
ablake@microsoft.com
Microsoft Research
1 Guildhall St.
GB-Cambridge CB3 3NH

Prof. Dr. Gilles Blanchard
Gilles.Blanchard@ens.fr
Department of Mathematics
Univ. Paris-Sud
Bat. 425
F-91405 Orsay Cedex

Prof. Dr. Joachim Buhmann
jb@cs.uni-bonn.de
Institut für Informatik III
Universität Bonn
Römerstr. 164
53117 Bonn

Prof. Dr. Heinrich H. Bülthoff
heinrich.buelthoff@tuebingen.mpg.de
Max-Planck-Institut für
Biologische Kybernetik
Spemannstraße 38
72076 Tübingen

Prof. Dr. Hans Burkhardt
burkhardt@informatik.uni-freiburg.de
Institut für Informatik
Universität Freiburg
Universitätsgelände, Geb. 052
79085 Freiburg

Prof. Dr. Vicent Caselles
vicent.caselles@tecn.upf.es
Departamento de Tecnologia
Universidad Pompeu Fabra
Passeig Circumvallacio, 8
E-08003 Barcelona

Prof. Dr. Gianni Dal Maso
dalmaso@sissa.it
S.I.S.S.A.
Via Beirut 2 - 4
I-34014 Trieste

Dr. Gustavo Deco
gustavo.deco@mchp.siemens.de
Hauptstr. 76
85579 Neubiberg

Prof. Dr. Agnes Desolneux
agnes.desolneux@cmla.ens-cachan.fr
CMLA
ENS Cachan
61, Avenue du President Wilson
F-94235 Cachan Cedex

Thomas Feldman
feldman@cmla.ens-cachan.fr
CMLA
ENS Cachan
61, Avenue du President Wilson
F-94235 Cachan Cedex

Prof. Dr. Andrew Fitzgibbon
awf@robots.ox.ac.uk
Dept. Engineering Science
University of Oxford
Parks Road
GB-Oxford OX1 3PJ

Prof. Dr. William Freeman
freeman@merl.com
Mitsubishi Electric Research Lab.
201 Broadway
Cambridge, MA 02139
USA

Dr. Martin A. Giese
martin.giese@tuebingen.mpg.de
AG für Handlungsrepräsentation und Lernen
MPI für Biologische Kybernetik
Spemannstr. 34
72076 Tübingen

Prof. Dr. Yann Gousseau
yann.gousseau@cmla.ens-cachan.fr
CMLA
ENS Cachan
61, Avenue du President Wilson
F-94235 Cachan Cedex

Prof. Dr. Michael Isard
michael.isard@compaq.com
COMPAQ Computer Cooperation
System Research Center
130 Lytton Avenue
Palo Alto, CA 94301
USA

Dr. Norbert Krüger
norbert@cn.stir.ac.uk
Department of Psychology
University of Stirling
Stirling FK9 4LA
Scotland

Dr. Markus Lappe
lappe@neurobiologie.ruhr-uni-bochum.de
Allgemeine Zoologie und Neurobiologie
Ruhr-Universität Bochum
44780 Bochum

Prof. Dr. Tai Sing Lee
tai@cnbc.cmu.edu
Center for Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213
USA

Prof. Dr. Christoph von der Malsburg
malsburg@usc.edu
malsburg@neuroinformatik.ruhr-uni-bochum.de
University of Southern California
University Park
HNB 301A
Los Angeles CA 90089-2520
USA

Prof. Dr. Jean-Michel Morel
morel@cmla.ens-cachan.fr
Centre de Recherche de Mathematiques de la Decision
CEREMADE
Universite Paris IX
F-75016 Paris

Prof. Dr. David Mumford
david_mumford@brown.edu
Division of Applied Mathematics
Brown University
Box F
Providence, RI 02912
USA

Pablo Muse
muse@dptmaths.ens-cachan.fr
CMLA
ENS Cachan
61, Avenue du President Wilson
F-94235 Cachan Cedex

Prof. Dr. Jean Petitot
petitot@poly.polytechnique.fr
CREA
Ecole Polytechnique
1 rue Descartes
F-75005 Paris

Prof. Dr. Christoph Schnörr
schnoerr@ti.uni-mannheim.de
Universität Mannheim
FB Math./Inform.: Computer Vision, Graphics & Pattern Recognition Grp.
68131 Mannheim

Prof. Dr. Andres Sole
andreu.sole@tecn.upf.es
Departamento de Tecnologia
Universidad Pompeu Fabra
Passeig Circumvallacio, 8
E-08003 Barcelona

Prof. Dr. Gerald Sommer
gs@ks.informatik.uni-kiel.de
Institut für Informatik und Praktische Mathematik
Universität Kiel
Preußerstr. 1-9
24105 Kiel

Prof. Dr. Simon Thorpe
thorpe@cerco.ups-tlse.fr
Centre de Recherche Cerveau et Cognition
133, route de Narbonne
F-31062 Toulouse

Prof. Dr. Franco Tomarelli
fratom@mate.polimi.it
Dipartimento di Matematica
Politecnico di Milano
Via Bonardi 9
I-20133 Milano

Prof. Dr. Alain Trouve
trouve@zeus.math.univ-paris13.fr
Departement de Mathematiques et d'Informatique
Ecole Normale Superieure
45, rue d'Ulm
F-75005 Paris Cedex

Prof. Dr. Joachim Weickert
Joachim.Weickert@uni-mannheim.de
Fakultät für Mathematik und Informatik
Universität Mannheim
68131 Mannheim

Prof. Dr. Yair Weiss
yweiss@cs.berkeley.edu
School of Computer Science and Engineering
The Hebrew University
Givat-Ram
91904 Jerusalem
ISRAEL

Dr. Laurenz Wiskott
L.wiskott@biologie.hu-berlin.de
Innovationskolleg Theoret. Biologie
Humboldt-Universität Berlin
Invalidenstr. 43
10115 Berlin

Ingo Wundrich
ingo.wundrich@neuroinformatik.ruhr-uni-bochum.de
Institut für Neuroinformatik
Ruhr-Universität Bochum
44780 Bochum

Dr. Rolf P. Würtz
rolf.wuertz@neuroinformatik.ruhr-uni-bochum.de
Institut für Neuroinformatik
Ruhr-Universität Bochum
44780 Bochum

Prof. Dr. Steven W. Zucker
steven.zucker@yale.edu
Department of Computer Science
Yale University
P.O. Box 2158, Yale Station
New Haven, CT 06520-2158
USA