[Journal Name], [Volumn Number], 1–14 ([Volumn Year])

COMVIS: A Communication Framework for Computer Vision

ALEX COZZI
IBM Almaden Research Center, San Jose, CA, USA
cozzi@us.ibm.com

FLORENTIN WÖRGÖTTER
Institute of Physiology, Department of Neurophysiology, Ruhr-University, Bochum, Germany
worgott@neurop.ruhr-uni-bochum.de

Keywords: Computer vision, data fusion, communication, integration of multiple visual modules

;

Abstract. We describe a general approach to integrate the information produced by different visual modules with the goal of generating a quantitative 3D reconstruction of the observed scene and to estimate the reconstruction errors.

The integration is achieved in two steps. Firstly, several different visual modules analyze the scene in terms of a common data representation: planar patches are used by different visual modules to communicate and represent the 3D structure of the scene. We show how it is possible to use this simple data structure to share and integrate information from different visual modalities, and how it can support the necessities of the great majority of different visual modules known in literature. Secondly, we devise a communication scheme able to merge and improve the description of the scene in terms of planar patches. The applications of state-of-the-art algorithms allows to fuse information affected by an unknown grade of correlation and still guarantee conservative error estimates.

Tests on real and synthetic scene show that our system produces a consistent and marked improvement over the results of single visual modules, with error reduction up to a factor of ten and with typical reduction of a factor 2–4.

1. Introduction

Computer vision systems are commonly composed by one or more visual modules specialized to detect the presence of particular features in images or image sequences and, on that basis, to infer the presence of some significant structure in the scene.

The specialization of visual modules is necessary to keep visual processing at a manageable level of complexity. On the other hand, isolated visual modules have problems in dealing with the variability of real scenes, producing erroneous or incomplete results.

The limitations of the approaches based on specialized visual modules made researchers realize the necessity of combining information from different image cues [1]. Intuitively, it is to be expected that the problems and limitations of many early vision modules can be overcome by combining the information extracted by many different cues contained in the images.

The solution that we propose consists of the development of an advanced computer vision system based on the integration of the information produced by heterogeneous visual modules, with the goal to generate an improved, consistent reconstruction of the viewed scene.

2. Motivating the Approach: Biological Background

Biological vision systems can easily cope with the complexities of the real world. At least part of this capability is due to the integration step that takes place in biological vision systems: in order to arrive at a consistent interpretation of a scene, different features, that apparently do not belong together, have to be linked and interpreted as belonging to the same physical object.

It has been suggested that synchronization processes can underly such "feature binding" in the nervous system [4]. Nerve cells respond with temporal activity patterns to a stimulus and it is known that certain oscillatory patterns prevail during optimal stimulation. It has been shown that groups of nerve cells synchronize their oscillations during stimulation with a common object, even if this object is partly obscured. Such a synchronization mechanism provides a common communication scheme, which in principle is understood throughout the whole nervous system. In this way modules that deal with quite different aspects of the visual scene can be bound together as soon as their activity is synchronous.

The synchronization process used by neural oscillators has been successfully integrated in computer models [20, 22, 16]. Still, their application to computer vision system implemented on serial computer is very inefficient and difficult to control.

3. The Visual Modules

In our study we concentrated on vision modalities able to produce a quantitative reconstruction of the three-dimensional structure of a scene. This information is most easily generated by analyzing how the projection of the scene changes when the viewpoint changes, as it is done in the two main visual modules of our system, an optical flow module and a stereo module, both based on differential matching [11]. Various other visual modules have been implemented and used in the COMVIS system: a phase-based stereo algorithm [6], a correlation-based stereo module, and an optical flow module based on the Lucas and Kanade algorithm [14]. The actual type of algorithm is not important, and the interested reader is referred to the original literature.

4. The Data Representation

The representation chosen as basis of the communication mechanism of COMVIS has to fulfill several requirements: It has to be global in the sense that all modules must be able to use it. It has to be optimal in the sense that it should encode the information with little redundancy. It has to be unambiguous and it has to be efficient, not requiring complicated pre-processing to be attained or complicated post-processing to be decoded.

A structure that fulfills all these requirement is the *planar patch*: a delimited planar surface. A planar patch consists of the parameters of its 3-D plane equation ($aX + bY + cZ = 1$) and the region that defines the points of the plane that belong to the planar patch (Fig. 1), encoded as the bitmap of its projection on a reference image.

The planar patch is completed by a covariance matrix that describes the error of the plane parameters, denoted by $\mathbf{\Sigma}$. The covariance matrix is a measure of the quality of the estimated plane's parameter, and it is essential to devise a rigorous data fusion mechanism.

Most of the visual modules known in literature can be readily modified to produce 3-D plane equations and planar regions. The covariance matrix is more difficult to derive. The optical flow module and the stereo module based on differential matching are able to compute it directly: the plane's parameter are fitted by minimizing the mismatch error between two images. The fitting procedure produces not only the plane's parameters, but also their covariance matrix [19, page 671].

Because we claim that our approach should be generally applicable, we must show how generic visual modules can be modified to produce estimates of the covariance matrix. There are two possibilities: mathematical modeling and Monte Carlo estimation. The first approach is described in [11, 5]. It is based on the fact that the actions of many visual modules can be modeled as computing the value that minimizes an appropriate error function. In this case it is possible to express the covariance matrix of the produced estimates in terms of the minimized function and the covariance matrix of the input images.

The second possibility is applicable even when it is impossible or too complicated to build a mathematical model of the visual module. It is based on the technique of Monte Carlo simulation [19, page 689]. We used this approach to describe the performances of a phase-based stereo module, as reported in [3]. The idea is to generate a synthetic input for which the result that the visual module should produce is known, then to perturbate the input by adding random noise, and generating in this way a set of test inputs. Applying the test inputs to the vision algorithm we obtain a set of outputs that can be used to estimate the probability distribution of the error thus of the covariance. In this way the error distribution for a certain class of images can be estimated off-line and then used to predict the errors of the visual module at run-time.

5. The Communication Mechanism

The goal of the communication scheme that we propose is to merge the information generated by different visual modules or by the same visual module on different inputs and produce a unified, minimal, consistent, and complete representation of the scene in terms of planar patches.

In this perspective, two features are to be communicated: (1) the parameters of the plane equations and (2) the image regions.

The communication process is divided into four steps:

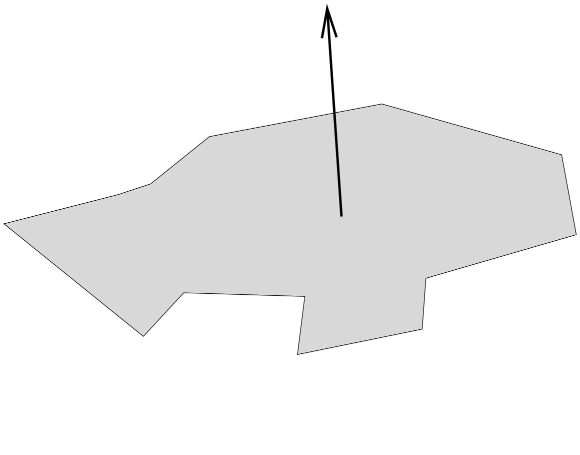

Fig. 1. A planar patch, characterized by a region and the plane's equation, here represented by the normal vector of the plane.

1. Collect all the planar patches generated by the different visual modules and transform the planes to the same coordinate system.
2. Identify the planar patches to be merged together.
3. Merge the plane equations.
4. Merge the image regions.

5.1. *Transform the planar patches to the same coordinate system*

The first step is performed applying the 3-D geometry of coordinates transforms. Using projective coordinates it is possible to express a general coordinate transform by matrix multiplication [5], a linear transformation. If $\mathbf{A}_p$ is the 4×4 matrix of the translation of a point expressed in projective coordinates, the corresponding transformation of the plane parameters is given by $\mathbf{A} = (\mathbf{A}_p^{-1})^{\mathrm{T}}$. We can then write the transformation of the plane's parameters as:

$$\mathbf{x}' = \mathbf{A}\mathbf{x} \tag{1}$$

where $\mathbf{x}'$ is the 4-dimensional vector of the plane's parameters in the new coordinate system, and $\mathbf{x}$ the plane's parameters in the old coordinate system. As a consequence the transformation of the covariance matrix can then be expressed as:

$$\mathbf{\Sigma}' = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}} \tag{2}$$

5.2. *Identify the planar patches to be merged together*

To perform the second step—identify the planar patches to be fused—we need a way to measure how "near" two planar patches are. This is done by an appropriate metric. The metric used in COMVIS takes in account only the parameters of the plane equation and the covariance matrix of a planar patch, ignoring the form or the size of its region. A distance measure that takes in account the covariances is the Mahalanobis distance [5]:

$$(\mathbf{a}-\mathbf{b})^{\mathrm{T}} \left(\mathbf{\Sigma}_{\mathbf{a}}+\mathbf{\Sigma}_{\mathbf{b}}\right)^{\mathbf{-1}} (\mathbf{a}-\mathbf{b})$$

where $\mathbf{a}$ and $\mathbf{b}$ are the vectors of the plane parameters and $\Sigma_{\mathbf{a}}$ and $\Sigma_{\mathbf{b}}$ the covariance matrices, respectively. The Mahalanobis distance has a chi-square distribution with as many degrees of freedom as the vectors $\mathbf{a}$ and $\mathbf{b}$ [5, page 318], i.e., 3 in this case. This allows to choose in a rigorous way what should it be the maximum distance to accept two planes as different samples extracted from the same multi-dimensional Gaussian distribution. This maximum distance, the threshold, is chosen in our experiments according to the probability of 99% that two vectors are different measurements of the same quantity, corresponding to a threshold value of 11.3.

The form and size of the plane region is actually not completely ignored: overlapping patches are merged in the the following phase.

5.3. *Merge the plane equations*

The third step is the fusion of two different estimates of the same plane, $\mathbf{a}$ and $\mathbf{b}$, to yield a new estimate $\mathbf{c}$, and a new covariance matrix, $\Sigma_{\mathbf{c}}$.

The most common data fusion algorithm, the Kalman filter [15], computes a new estimate by a linear combination of $\mathbf{a}$ and $\mathbf{b}$, assuming the two vectors to be independent estimates and choosing the weight matrices to minimize the trace of $\Sigma_{\mathbf{c}}$.

A problem with this approach is that in real situations the measures $\mathbf{a}$ and $\mathbf{b}$ are often correlated and their correlation grade can not be measured. In the COMVIS system we adopt a recently proposed novel extension of the Kalman filter paradigm: the *Covariance Intersection* (CI) algorithm [21].

CI performs a combination of the means and covariances in inverse covariance space. The CI estimate is computed as:

$$\begin{aligned} \Sigma_{\mathbf{c}} &= (\omega\Sigma_{\mathbf{a}}^{-1} + (1-\omega)\Sigma_{\mathbf{b}}^{-1})^{-1} \\ \mathbf{c} &= \Sigma_{\mathbf{c}}(\omega\Sigma_{\mathbf{a}}^{-1}\mathbf{a} + (1-\omega)\Sigma_{\mathbf{b}}^{-1}\mathbf{b}) \end{aligned} \tag{3}$$

where $\omega \in [0,1]$ controls how the two initial vectors, $\mathbf{a}$ and $\mathbf{b}$, influence the final estimate. The value of ω is determined by numerical minimization each time that two estimates are fused, by finding the value of ω that minimizes the determinant of the fused covariance matrix, $\Sigma_{\mathbf{c}}$. The estimate described by Eq. 3 can be proved to be consistent for all possible cross-correlation grades and choices of ω [21].

5.4. *Merge the image regions*

The goal of the last step, the segmentation of the regions, is to assign each pixel of the reference image to one and only one planar patch.

The problem of segmenting the planar patch regions is solved by selecting for each pixel the planar patch that minimize the matching error of that point in all the camera.

A pixel in the reference camera identifies a ray that starts from the camera's optical center and intersects the image plane at the pixel's location. Consider now a planar patch. When the ray does not intersect the plane of the planar patch, the plane is not visible in this particular pixel. In this case the cost of assigning the pixel to the planar patch should be very high. This situation is modeled by an error of $+\infty$. Otherwise, when the ray intersects the plane of the planar patch we can project the intersection point into all the other images of the scene obtained from the other cameras available to the system, thus identifying a corresponding pixel in each one of the other images. The squared difference of the pixel in the reference image with the corresponding pixels in the other images determines the error of assigning the pixel to the planar patch. Formally, the error (or cost) of assigning the pixel f_i to the planar patch P_k is

$$E(f_i = k) = \begin{cases} +\infty & \text{no intersection} \\ \sum_{c\in C}(\frac{v_i - v_{c,i}}{\sigma_n})^2 & \text{intersection} \end{cases} \tag{4}$$

where f_i is the label of the pixel i, v_i the luminance value in the reference image, $v_{c,i}$ the luminance value of the corresponding pixel in the image from camera c, C is the set of the cameras and σ_n is the standard deviation of the noise in the images.

This error function produces good results if the objects in the scene have enough texture to distinguish between true correspondences and casual ones. In many real scenes this is not the case, and as a consequence two kinds of problems occur: either different assignments have the same potential, or their differences are below the noise threshold—i.e., they are probably due to the noise—and thus should not be considered to be significant. This can be overcome by using the error of a small region (i.e., 5×5) instead of a single pixel.

5.5. *Computational Cost*

The first step of the communication procedure, transformation of the planar patches into the same coordinates system, is small and directly proportional to the number of patches generated by the visual modules. The second and third steps are efficiently performed together: the already processed patches are hold in a list. If a new patch matches one of the patches already in the list, in the sense that their Mahalanobis distance is under the threshold or their regions are overlapping, they are merged, otherwise the new patch is just added at the end of the list. The cost of this stage is potentially quadratic in the number of patches. The last step is the most computationally intensive, since it requires the matching error to be computed for every pixel in the reference image and for every possible patch. This computational cost can be greatly reduced by grouping the pixels in squared tiles, 16×16 pixels, for example. This reduces the resolution of the regions map, but makes the merging phase much faster.

6. The System

The above described communication scheme is employed by our advanced computer vision system, COMVIS. The computation performed by the COMVIS system is divided into several stages. The details of the data flow through the different stages is depicted in Fig. 2.

For most of the experiments we used a combination of several stereo and motion modules based on the same differential matching algorithm [11], because this is the technique that produces the most accurate estimates of the planar patches' plane equations and their covariance matrices. The same algorithm is applied to different input images, and each instance of the algorithm is treated by COMVIS like an independent visual module.

The planar patches generated by all the modules are collected in a common repository, called *blackboard*, where the fusion procedure takes place. The Mahalanobis distance of each pair of planar patches is computed, and each pair whose distance is less than 11.3 (the 99% quantile of the chi-square distribution with 3 degrees of freedom) is fused by the Covariance Intersection algorithm.

In the next step, the regions of all the fused planar patches are segmented and refined by minimizing the matching error, so that every pixel of a reference image is assigned to a single planar patch. This step concludes the data fusion procedure, and it generates a 3D model of the scene in terms of planar patches.

An additional step can further improve the 3D model of the scene by re-estimating the plane's equations after the segmentation step. At this stage the accumulated knowledge about the patches' regions allows a better estimate of the planes' normals: a new estimate is computed by minimizing the mismatch between the reference image and the corresponding points in all the other images of the scene for all pixels that belong to the region of the planar patch.

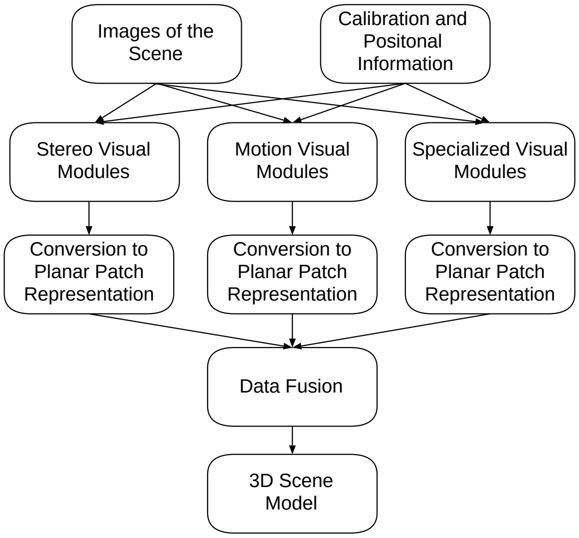
Images of the Scene
Calibration and Positonal Information
Stereo Visual Modules
Motion Visual Modules
Specialized Visual Modules
Conversion to Planar Patch Representation
Conversion to Planar Patch Representation
Conversion to Planar Patch Representation
Data Fusion
3D Scene Model

Fig. 2. The data flow through the several processing stages of the COMVIS system.

7. Experiments

The reconstruction of a synthetic scene is presented in Fig 3. This scene is composed of four planes in the four quadrants of the image. A stereo module analyzes the stereo pair and generates a set of planar patches. In order to clearly show the difference between Kalman fusion and CI, only planes that have neighboring regions are considered for fusion and the pixels of the regions are grouped in tiles of size 16×16 . From the regions images it is apparent how the CI algorithm generates a better segmentation. The reason being that CI generates more conservative, i.e. larger, covariance matrices for the fused planes. This is important in the fusion phase because if the covariance of a fused planes is too small—i.e., the fusion procedure assumes that the plane is more precise that what it really is—its Mahalanobis distance from other measurements of the same plane will probably exceed the threshold used to decide when to fuse two planes, thus forbidding fusion.

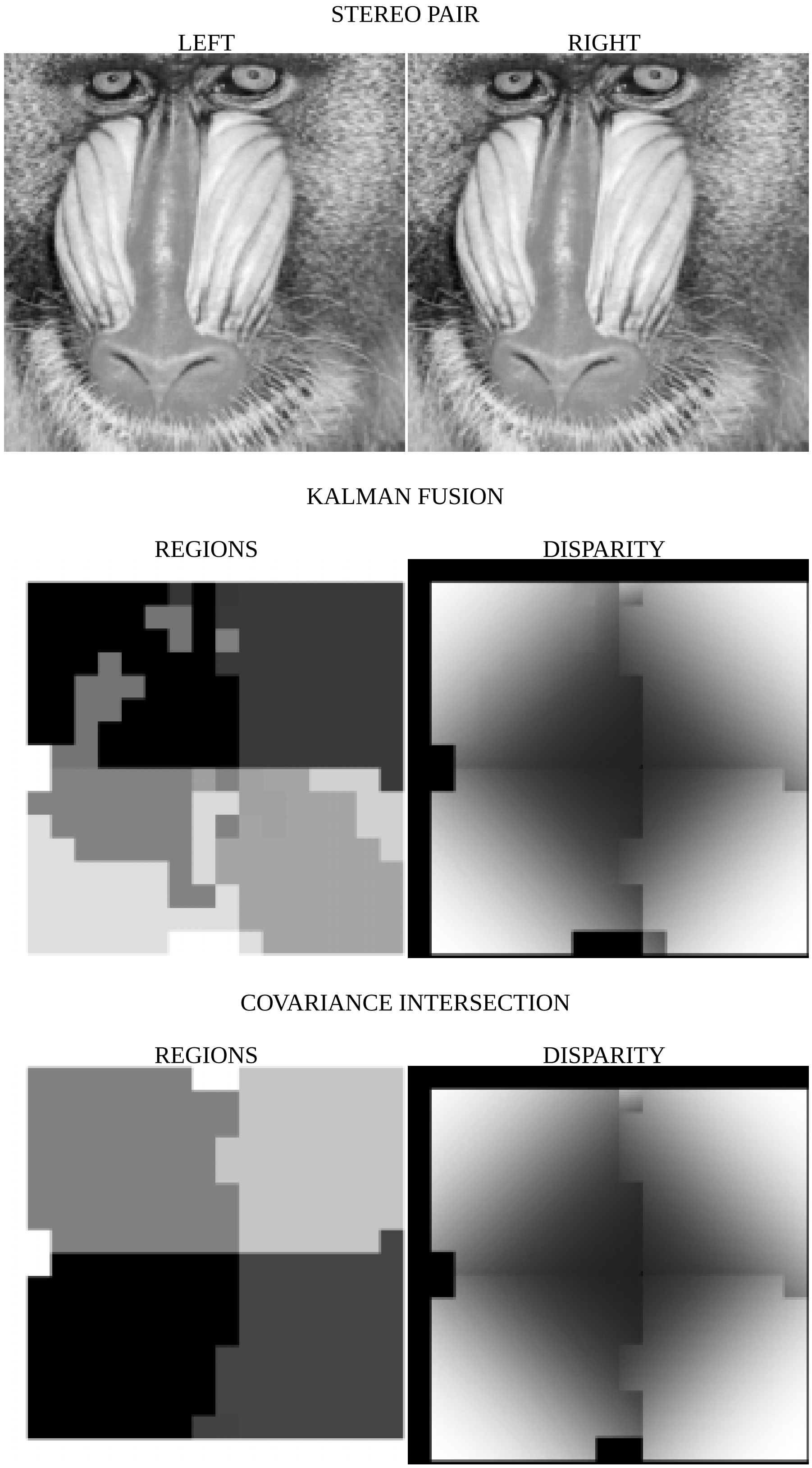

Fig. 3. The reconstruction of a synthetic scene. At the top is the stereo pair, in the middle the regions and the depth map reconstructed with Kalman fusion, at the bottom the same but with CI. Each shaded square represents a planar patch. The shading represents the distance of the object from the viewer: light gray means far and dark gray means near.

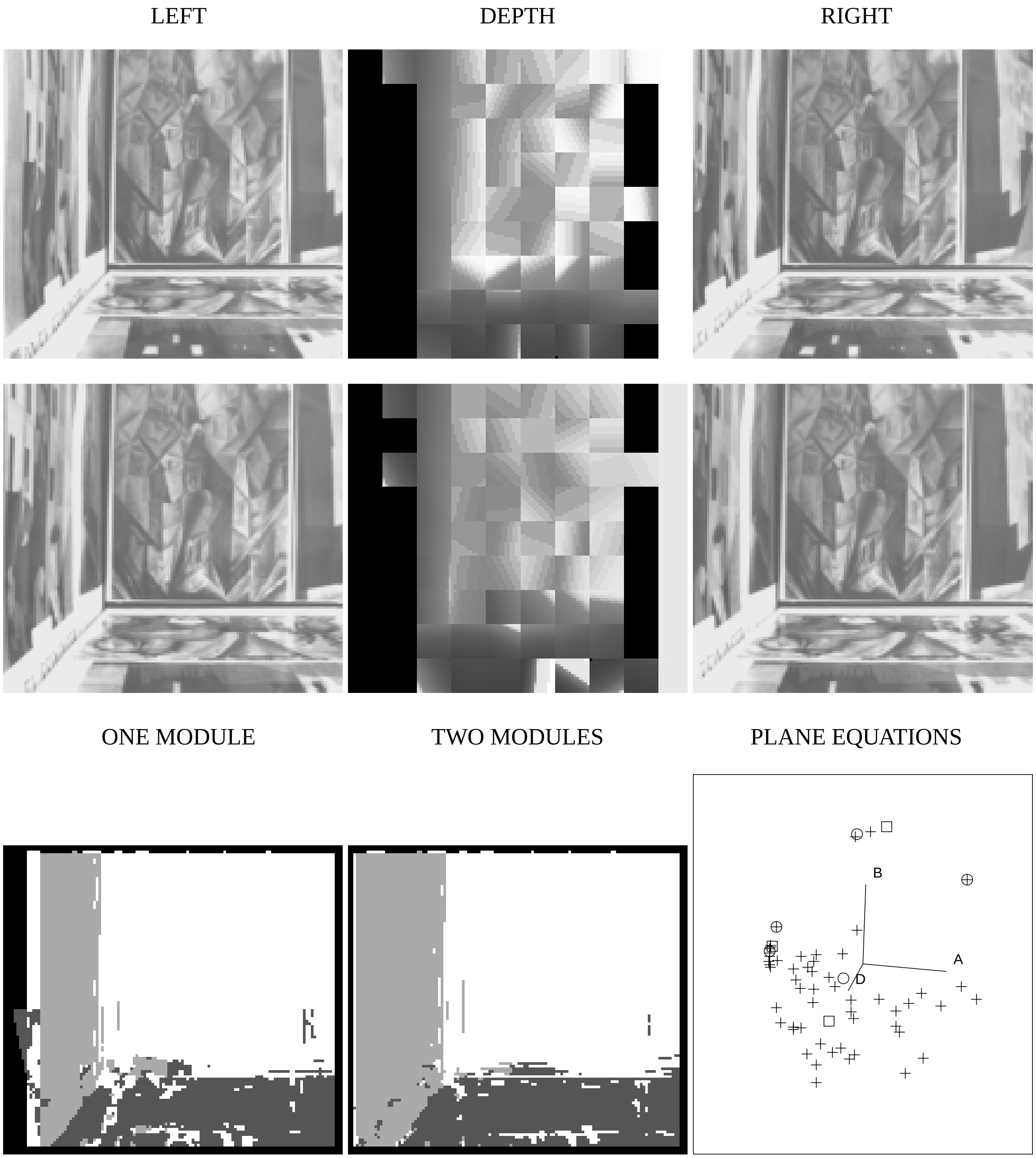

Fig. 4. Two stereo pairs of the same scene from slightly different viewpoints: the upper row is taken 5cm nearer to the wall. The central picture is the depth map of the scene as extracted by a stereo module. At the bottom are the reconstructed regions, on the left using only one stereo module and two viewpoint, then using both stereo modules and four viewpoints. At the bottom right the plot of the recovered parameters of the planes in the four viewpoints case using CI. The crosses are the planes generated by the stereo modules, the circles the planes at the end of the fusion procedure and the squares show the real values of the parameters.

To test the COMVIS system a real a scene with known characteristics was build in our laboratory (Fig. 4). This scene consists of three orthogonal planes: a floor plane, a left wall and a back wall, all with a textured surface.

In the experiment two differential stereo modules were used, each one acting processing a different stereo pair. The stereo pairs were taken from two different positions, shifted by 5cm respect each other along the Z axis. The COMVIS system segmented the scene into three planar patches. The patches' regions are shown in Fig. 4. In this figure the results for two configurations of the system are reported. In the first row are the results obtained using only one stereo module and two views of the scene (the upper stereo pair in Fig. 4). In the second row are reported the results of adding another stereo module and two more views in the computation of the error function (the lower stereo pair in Fig. 4). In the bottom left corner we show the plot of the planes parameters. After the fusion procedure four planes are left.

7.1. The CIL Sequences

In 1994, the Carnegie-Mellon University released on the Internet the images of three static scenes complete with accurate information about the camera calibration and the true positions of several points visible in each scene [2]. These sequences represent a fair test case for the COMVIS system, since they can be used to test the actual accuracy of the fusion procedure in estimating the position of points whose real location is known.

The first scene, CIL-1, depicts a castle. The challenges of this scene are the great number of different planes and discontinuities that it contains. For each scene, eleven pictures taken from different positions are provided. For our experiments we used the three images taken at the positions 2, 3, and 4 of the CIL sequences, where the camera position N.3 is the origin of the frame of reference, camera N.2 is on the left of the origin and camera N.4 is on the right of the origin. Figure 5 reports the points whose real positions are known, indicated in the picture by a white point and a number.

The original images have a size of 576×384 pixels. The viewer-centered coordinate system used to represent the image regions is coincident with the camera N.3 (the camera at the origin of the coordinate system).

For these experiments we used three independent differential stereo modules. Of the eleven camera positions provided by the CIL sequences, three stereo pairs had enough overlap and worked well with our stereo module. The stereo modules used the stereo pairs $(3, 2)$, $(4, 3)$, $(4, 2)$ as input images and each one produced a list of planar patches describing the scene. The tables 1 and 2 report the estimation errors Z coordinate of the test point. This position is computed as the intersection between the generated planar patches and the projection of the test point thorough the optical center of the reference camera.

The first stage produces from zero to three measurements for every pixel. The number of measurements depends on how many modules have produced a planar patch covering the test point.

The data fusion stage takes the planar patches produced by the stereo modules and fuses together similar planar patches. As explained in Sect. 5.3, the planes are fused mostly on the basis of their orientation. This means that even with a single module it is possible to improve the model of the scene by merging planar patches originating from different locations of the same image.

The regions of the patches are then segmented so as to minimize the matching error, which is then computed as described in Sect. 5.4 using the images 2, 3, 4, 6 and 10 of the sequence, since they provide highly overlapping views of the scene. The segmentation stage can extend the planar patches to regions where the stereo modules were unable to produce sensible results, this explains why the tables report an estimated Z after the fusion procedure for points that had no estimate provided by the stereo modules.

The final stage modifies the plane equation by minimizing the matching error and re-estimating the covariance matrix from the Hessian of the error surface at its minimum. In this stage COMVIS takes advantage of the whole region of the planar patch to re-estimate the plane parameter with greater precision.

In evaluating the results in table 1 and 2 it should be kept in mind that the COMVIS system is based on planar patches. On the other side, the test refers to the error of a *point*. It can happen that, in order to get a better fit for a plane, the error of a point gets larger. Another observation is that the fusion procedure do not work well at the borders of objects (unfortunately many of the test points are at the end of the objects in the scene) because in this points the planarity assumption is violated.

From the tables 1 and 2 it can be seen how the fusion procedure is able to reduce the errors of the estimates produced by the stereo modules and to fill in the regions where the stereo modules did not produced measurements. In general the differences between Kalman filtering and CI are not great, and mainly consist of the larger standard deviation estimates for CI, as it is to be expected.

8. Discussion

The characteristic that distinguishes COMVIS from other approaches is its theoretical foundation. We choose to restrict our system to a single data structure and to fuse the information in a manner that rigorously takes cross-correlations into account.

A good overview of other approaches is offered by first monograph about communication of different visual modules [1]. There is proposed a two-fold approach for the integration of visual modules. The first is defined as “bottom-up” and consists of the development of specialized techniques for combining two or more cues to recover intrinsic object parameters. The second is the “top-down” approach. It consists of a general theory of combining different visual cues in order to obtain a unique and robust solution to the visual reconstruction problem. This work provides an interesting starting point, and it is rather accurate in the description of the “bottom-up” approach. Nevertheless, the top-down approach is only drafted, and not developed into an usable form. Our work represents an instance of a top-down, general communication and fusion approach for visual modules based on a quantitative and geometric model of the scene.

An alternative approach is to perform the communication process at a lower level of abstraction. A good example is provided by the MIT Vision Machine [9, 18]. Its communication scheme is designed around the detection and representation of discontinuities inside the feature maps. In this scheme the intensity edges are used to relax the smoothing constraints imposed among neighboring features. The rationale being that changes in surface properties usually produce strong intensity edges, and thus discontinuities in the feature maps often appear in coincidence with the intensity edges. This approach is a generalization of the regularization approach, extended so to be able to deal with discontinuities and to integrate different visual modules. The communication mechanism uses coupled Markov random fields (MRF) [10] to implement a discontinuity-preserving regularization procedure. A MRF is associated to each feature map (each feature map is produced independently by a visual module) and another MRF is associated with its discontinuities (this second MRF is called a “line process” [10]). The communication is performed by coupling the MRFs to each other in order to model their interdependencies—for example, between surface depth and surface normal. These interdependencies have the form of local constraints, and impose smoothness conditions everywhere *except* at discontinuities, where the line processes allow for discontinuities in the recovered feature maps.

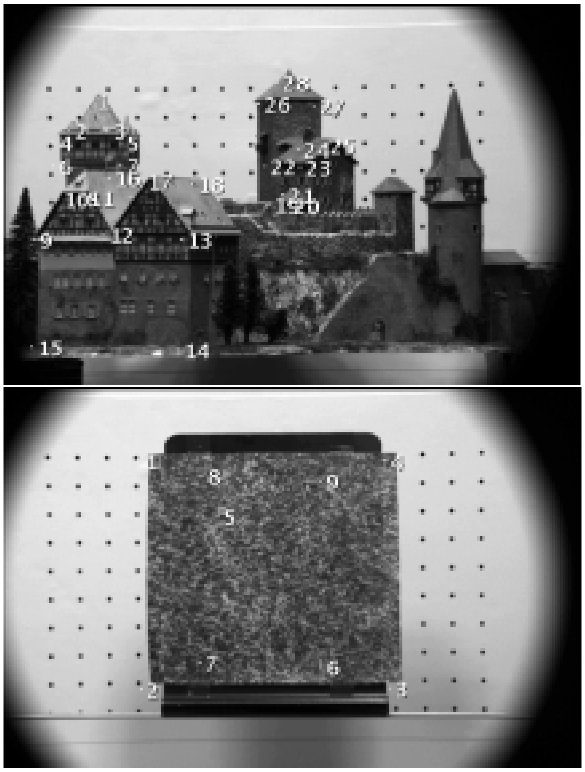

Fig. 5. The points of the CIL-1 and CIL-2 scenes provided by the Calibrated Imaging Laboratory of Carnegie-Mellon University whose real distances were measured.

A similar approach, based on the theory of Bayesian estimation, is proposed by Jain and Pankanti [17]. In their approach the feature maps produced by the different modules are improved by finding their maximum a posteriori estimate (MAP estimate). Different vision modalities are integrated by imposing consistency con-

Table 1. The errors in the estimation of the Z coordinate of the known points in the CIL-1 sequence. The coordinates are expressed in cm.

Point	Stereo Err. **cm**	Kalman Err. **cm**	Kalman Std.Dev. **cm**	CI Err. **cm**	CI Std.Dev. **cm**
1	N/A	6.06	0.98	5.70	1.39
2	N/A	19.15	2.06	20.79	1.08
3	3.52	8.13	0.72	7.75	1.03
4	N/A	7.06	4.60	7.12	4.61
5	N/A	8.09	0.66	4.32	0.49
6	N/A	12.79	0.43	15.85	0.74
7	N/A	19.55	0.65	19.57	0.65
8	26.41	25.13	0.79	25.19	0.79
9	N/A	3.99	0.78	3.55	0.94
10	N/A	7.05	0.96	7.12	0.95
11	6.73	4.00	0.78	4.05	0.78
12	3.44	3.01	0.67	3.07	0.67
13	1.35	5.82	0.09	6.22	0.13
14	4.09	0.79	1.17	0.87	1.17
15	N/A	N/A	0.00	N/A	0.00
16	0.05	6.53	0.65	6.57	0.65
17	N/A	0.13	2.68	0.21	2.95
18	N/A	4.52	0.07	1.42	0.27
19	1.72	9.15	0.11	0.71	1.46
20	1.30	0.31	0.25	3.65	1.41
21	0.95	0.13	0.23	0.39	1.03
22	N/A	5.37	0.51	5.34	0.51
23	2.47	1.49	1.53	1.44	2.17
24	4.06	0.80	1.47	0.70	1.47
25	1.98	5.79	0.11	6.11	0.16
26	N/A	2.94	1.05	8.42	2.35
27	9.53	1.37	0.82	1.53	0.82
28	2.63	1.63	2.58	11.01	0.17

Table 2. The errors in the estimation of the Z coordinate of the known points in the CIL-2 sequence. The coordinates are expressed in cm.

Point	Stereo Err. **cm**	Kalman Err. **cm**	Kalman Std.Dev. **cm**	CI Err. **cm**	CI Std.Dev. **cm**
1	4.37	21.45	1.05	21.95	1.33
2	5.60	28.83	2.21	29.48	2.74
3	N/A	6.69	3.49	6.74	3.44
4	5.33	0.17	0.19	2.99	0.37
5	2.20	0.26	0.09	0.20	0.16
6	2.57	1.45	4.83	1.82	6.59
7	1.48	2.12	3.85	2.13	4.95
8	2.67	2.37	3.26	2.63	3.21
9	9.35	2.01	0.15	1.92	0.26

ditions across different feature maps, thus coupling the global estimation process. A coherence principle is used to estimate the reliability of a feature. This is done by relating the estimate of a variable with the estimates of the same variable derived from the neighboring estimates—which amounts to a regularization procedure.

The Bayesian approach and the Vision Machine share the same limitations: in order to obtain a computable solution it is necessary to assume the independence of the different measurements (i.e., they disregard cross-correlations) and because of the huge number of involved statistical variables it is necessary to reduce the model of the probability distributions at the minimum, thus severely limiting its modeling capabilities. Another necessary simplification is to assume the independence of the estimated values contained in the maps and the interaction between the different feature maps must be limited temporally and spatially—the interactions between the maps are restricted to be sequential and only local interactions between the immediate neighbors are permitted. The effect of using such limiting assumptions is a simplified and solvable model, but also a model that does not realistically represent the complexity of the real distributions and is thus unable to give clear indications about the limits of the estimated solution.

Another problem with all these regularization-based approaches is that the communication mechanism is defined ad-hoc for each feature map and mostly heuristically derived.

The alternative that we propose with our system limits the size of the feature maps by the use of a single kind of higher level feature, the planar patch. The resulting smaller, though more complex, feature space allows more sophisticated and better grounded data fusion techniques and limits the amount of processing. The planar patch has been previously used as a primitive for representing 3-D scene models [5]. New in our work is its application to the problem of communication between visual modules and the pursuit of the possibilities offered by its simplicity for the development of rigorous solutions to the problems of data fusion and error estimation in a computer vision system.

A similar approach is exposed in [8]. The main difference is that a mesh of hexagonal planar patches is used to reconstruct a model of the scene instead of independent planar patches and that our system requires and uses the covariance matrices of the estimated patches. Our system does not use global constraints and tries to merge the planar patches on the basis of their plane equation, thus improving the results mostly by fusion, while the system of Fua and Leclerc fuses the patches on the basis of their region and adds global smoothness constraints on the reconstructed surface.

9. Conclusions

This study extends the standard approaches used in computer vision in an evolutionary manner, building on the classical works of the field and proposing new solutions that still fit within the current context of computer vision research. The approach advocated here integrates several recent research developments (e.g. Covariance Intersection, Haralick's theory of error propagation) in a coherent unity. A goal of this work is to show the potential and the flexibility of the communication approach, providing a solid basis for the further development of more complex communication schemes.

An important feature of this study is the rigorous mathematical treatment underlining most of the techniques that define the communication schema and the minimal recourse to heuristics. Thus, it was often possible not only to propose rigorous solutions to the problems, but also to prove their optimality—i.e., to show that no better solutions are available in this context.

The communication schema proposed in this work is very general because only a few assumptions are made about the nature of the visual modules and the scene being analyzed. Any visual module that encodes its results in terms of planar patches can be easily integrated into the system, and most computer vision modules fit this description, eventually requiring some additional work to estimate the covariance matrix of the results.

The actual applicability of the proposed communication mechanism depends critically on the solution to the problem of fusing information affected by an unknown degree of correlation. The technique of Covariance Intersection is a recent development of the Kalman filter theory which addresses this concern. Its use by the COMVIS system represents, to our knowledge, its first application in computer vision.

Our research is not limited to a theoretical inquiry into the problem of the integration of visual modules.

A very significant effort was put into implementing our proposed solution in a library of Java classes, freely available on the Internet at http://www.neurop2.ruhr-uni-bochum.de/java.

10. Problems and Limitations

The communication and data fusion scheme that we propose makes extensive use of detailed information about the geometrical configuration of the cameras, thus limiting its applicability to situations where this information is available.

A critical assumption is that the visual modules that generate the estimates of the planar patches are able to produce an estimate of their errors and that the distribution of such errors can be approximated with a multidimensional Gaussian distribution. If the visual modules underestimate their errors or the estimated patches are affected by significant bias the fusion procedure will not be able to improve the results.

Another limitation is the way that the region of the planar patches are represented. In this work, a viewer-centered representation has been used, as it was simpler to attain and to handle. Nevertheless, a more advanced, object-centered representation would make the system more flexible and powerful. A lot of experience about object-centered region representations is readily available from the field of computer graphics, and should be easily integrable into our communication scheme.

The current communication procedure has the limitation of not enforcing any kind of geometrical relationship among different planes: each plane is independent and no constraints are posed on neighboring or adjacent planes. Such constraints could improve the performance of the system and lead to more robust solutions (see, for example, the adjacency condition of optical flow [13, page 256] and [8] for more complex global smoothness conditions).

The performances of the current implementation of COMVIS are limited by the use of planar patches. For scenes of consisting of man-made objects this is often a good enough approximation, but when this assumption is violated, for example in the case of natural scenes cluttered with trees, rocks or other highly irregular objects, the reconstructed scene is divided in many tiny little patches, and the fusion procedure is not very effective.

Another problem of the representation we choose is that the similarity measure of the plane equations does not take in account the part of the plane that was used to generate the estimates of the plane parameters. This works well only when the images of the scene are generated by using a telephoto lens, thus using a quasi-orthographic projection. Otherwise the differences in perspective tend to magnify small estimation errors defeating the fusion procedure.

11. Extensions

One way of extending the capabilities of the COMVIS system is intrinsic in its modular architecture. It consists of adding more visual modules. Particularly interesting would be the integration of modules that exploit visual cues other than stereo and motion, for example, modules that use color information or shading. Even more intriguing would be the integration of modules that use non-visual information, e.g. radar or sonar systems.

Another way of generalizing the communication mechanism proposed here is to extend it so that it can handle more complex surface representation primitives, e.g. generic surfaces, superquadrics or generalized cylinders. A first step in this direction is to characterize the precision of the parameters that define the primitive and to describe how the parameters and their precisions are transformed by changes in the coordinate system. Another interesting idea is to keep the planar patches-based representation but to augment it with a “bump map” of the displacements of the points relatively to the plane, like the plane-parallax representation of Irani and Anadan [12].

The self-consistency principle of Leclerc et al. [7] allows to evaluate the performances of computer vision systems without requiring the knowledge of ground truth, and could prove very useful to tune the performances of COMVIS to different scenes.

References

1. J. Aloimonos and D. Shulman. *Integration of Visual Modules – An Extension of the Marr Paradigm*. Academic Press, London, 1989.
2. Calibrated Imaging Laboratory home page. http://www.cs.cmu.edu/∼cil/cil-ster.html
3. A. Cozzi, B. Crespi, F. Valentinotti, and F. Wörgötter. Performance of phase-based algorithms for disparity estimation. *Machine Vision and Applications*, 9(5-6):334–340, 1997.
4. R. Eckhorn, R. Bauer, W. Jordan, B. M., W. Kruse, M. Munk, and H. Reitböck. Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biological Cybernetics*, 60:121–130, 1988.

5. O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, 1993.
6. D. Fleet, A. Jepson, and M. Jenkin. Phase-based disparity measurement. *Computer Vision, Graphic and Image Processing*, 53(2):198–210, Mar. 1991.
7. P. Fua, Y. Leclerc, and Loung. Characterizing the performance of multiple-image point-correspondence algorithms usign self-consistency. In *Proceedings of: the Vision Algorithms: Theory and Practice Workshop (ICCV99)*, Corfu, Greece, 1999.
8. P. Fua and Y. G. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *IJCV*, 16:35–56, 1995.
9. E. Gamble and T. Poggio. Visual integration and detection of discontinuities: the key role of intensity edges. A.I. Memo 970, MIT, Artificial Intelligent Laboratory, Oct. 1987.
10. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Intelligence*, 6(6):721–741, Nov. 1984.
11. R. M. Haralick. Propagating covariance in computer vision. In H. I. Christensen, W. Förstner, and C. B. Madsen, editors, *Workshop on Performance Characteristics of Vision Algorithms*, pages 1–12, Cambridge, U.K., Apr. 1996. http://www.vision.auc.dk/∼hic/performance-ws.html
12. M. Irani and P. Anandan. Parallax geometry of pairs of points for 3d scene analysis. In *Proc. 4th European Conf. on Computer Vision, Cambridge*, volume 1, pages 17–30, Cambridge, UK, 1996.
13. K. Kanatani. *Group-Theoretical Methods in Image Understanding*, volume 20 of *Springer series in information sciences*. Springer-Verlag, Berlin, Heidelberg, 1990.
14. B. Lucas and T. Kanade. Optical navigation by the method of differences. In *DARPA84*, pages 272–281, 1984.
15. P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1 of *Mathematics in Science and Engineering*. Academic Press, New York, 1979.
16. R. Opara and F. Wörgötter. A fast and robust cluster update algorithm for image segmentation in spin-lattice models without annealing—visual latencies revisited. *Neural Computation*, 10:1547–1566, 1998.
17. S. Pankanti and A. K. Jain. A uniform bayesian framework for integration. Technical report, Michigan State University, 1995.
18. T. A. Poggio, E. Gamble, and J. J. Little. Parallel integration of vision modules. *Science*, 242:436–439, Oct. 1988.
19. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition edition, 1992.
20. W. Singer and C. M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.*, 18:555–586, 1995.
21. J. K. Uhlmann. A culminating advance in the theory and practice of data fusion, filtering, and decentralized estimation. Technical report, Covariance Intersection Working Group, 1997. http://www.ait.nrl.navy.mil/people/uhlmann/CovInt.html.
22. D. Wang and D. Terman. Image segmentation based on oscillatory correlation. *Neural Computation*, 9:1623–1626, 1997.