NeuroImage **13,** 459–471 (2001)
doi:10.1006/nimg.2000.0702, available online at http://www.idealibrary.com on IDEAL®

Reproducibility of fMRI: Effect of the Use of Contextual Information

Eero Salli,*,† Antti Korvenoja,†,‡ Ari Visa,§ Toivo Katila,*,‡ and Hannu J. Aronen†,‡,¶

**Laboratory of Biomedical Engineering, Helsinki University of Technology; †Department of Radiology and ‡BioMag Laboratory, Helsinki University Central Hospital; §Signal Processing Laboratory, Tampere University of Technology; and ¶Department of Clinical Radiology, Kuopio University, Finland*

Received May 2, 2000; published online January 24, 2001

We studied the effect of use of contextual information on the reproducibility of the results in analysis of fMRI data. We used data from a repeated simple motor fMRI experiment. In the first approach, statistical parametric maps were computed from a spatially unsmoothed data and thresholded using a Bonferroni corrected threshold. In the second approach, the maps were computed from a spatially unsmoothed data but were segmented into nonactive and active regions using a spatial contextual clustering method. In the third approach, the statistical parametric maps were computed from spatially smoothed data and thresholded, using, optionally, a spatial extent threshold. The variation in the classification was largest in the Bonferroni thresholded statistical parametric maps. There were no significant differences in variation between statistical parametric maps generated with all the other methods. In addition to reproducibility, the detection rates of weak simulated activations in the presence of measured scanner and physiological noise were investigated. Contextual clustering method was the most sensitive method, while the least sensitive method was the Bonferroni corrected thresholding. Using simulated data, we demonstrated that the contextual clustering method preserves the shapes of activation regions better than the method using spatial smoothing of the data. © 2001 Academic Press

INTRODUCTION

The primary goal of image analysis in functional magnetic resonance imaging (fMRI) activation studies is usually to detect and delineate the image areas that have a signal intensity time course, which can be related to the experimental parameters. The task is challenging because the images are noisy and often corrupted by motion. The problem is typically solved using a statistical testing procedure.

First, a statistical parametric map (SPM), also called a statistic image, is created. Thereafter, either a nonactive or active state is assigned to each voxel. In this article, we will call this the segmentation phase, as the goal of this step is to divide the statistical parametric map into nonactive and active regions. Prior to the computation of statistical parametric map, some preprocessing steps like motion correction and temporal and spatial filtering are usually performed.

The statistical parametric map is most often computed using the general linear model that subsume for example the simple *t* test (Friston *et al.*, 1995b). Also nonparametric methods, like the Kolmogorov–Smirnov (KS) test have been employed as well. The nonparametric methods do not assume that data are normal. However, independence of the samples is assumed. Hence, the temporal autocorrelations increase false-positive rates in the KS test over tabular values (Aguirre *et al.*, 1998). Since fMRI data is only slightly nonnormal (Aguirre *et al.*, 1998), and the temporal autocorrelations can be dealt within the general linear model (Worsley and Friston, 1995), the parametric general linear model is usually preferred.

A commonly used segmentation method is intensity thresholding. The voxels whose statistical value is larger than a predefined threshold, directly related to the significance level of a statistical test, are classified as active. The advantage of the thresholding is its simplicity. However, the restriction of the thresholding is widely known: the histograms of activation and nonactivation classes overlap and the classes may not be partitioned by using a single threshold. As a consequence, when high specificity, i.e., low false-positive rate is required, weak activations are not detected.

One way to improve the segmentation is to utilize an assumption that the probability of the activation is related to the existence of activation in the spatial neighborhood of a voxel. This *a priori* information may be incorporated into the classification procedure by using so called contextual methods. A simple way to use the contextual information is to ignore the original statistic values after thresholding and leave only voxels that have a sufficient number of active neighbors, i.e., use so called neighborhood filters (Constable *et al.*, 1993; Skudlarski *et al.*, 1999). The statistical parametric map can also be filtered before thresholding (Skud-

1053-8119/01 $35.00

larski *et al.*, 1999). Another extension to the intensity thresholding is the use of spatial extent, or cluster-size, thresholds (Friston *et al.*, 1994; Forman *et al.*, 1995). In this approach, only activation clusters larger than a defined cluster size are considered as statistically significant. The use of spatial extent thresholds allows detection of weak activations when their spatial extent is relatively large while preserving high overall specificity. Information from the spatial neighborhood can also be incorporated into the analysis by either spatially filtering the original fMRI data with a Gaussian shaped filter (Friston *et al.*, 1995a; Lowe and Sorenson, 1997) or by doing a Markov Random Field (MRF) based data restoration (Descombes *et al.*, 1998; Kruggel *et al.*, 1999).

Previously a computationally efficient contextual clustering algorithm for the segmentation of statistical parametric maps was introduced (Salli *et al.*, 1999). In the contextual clustering algorithm, both statistical parametric values and classification information from the neighborhood of each voxel are iteratively used to make a decision whether a voxel is active or not. Hence, the contextual clustering differs significantly from the neighborhood filters.

The main goal of the present study was to analyze the reproducibility of segmentation. Noll *et al.* (1997) have made reproducibility studies using motor and cognitive activation paradigms. Casey *et al.* (1998) studied reproducibility across four institutions using a spatial working memory task. However, neither of these studies analyzed the effect of segmentation methods on the reproducibility. In the present study our approach was to minimize the sources of variation and study reproducibility achieved with various segmentation methods. False-positive rates associated with the chosen segmentation parameters were studied using both simulated statistical parametric maps and empirical data acquired from a human volunteer.

Additionally, we studied segmentation accuracy and sensitivity using a simulated activation pattern embedded in a time-series acquired from a resting human volunteer. Comparison between the results obtained by using voxel-by-voxel intensity thresholding, spatial extent method combined with spatial smoothing and contextual clustering was made.

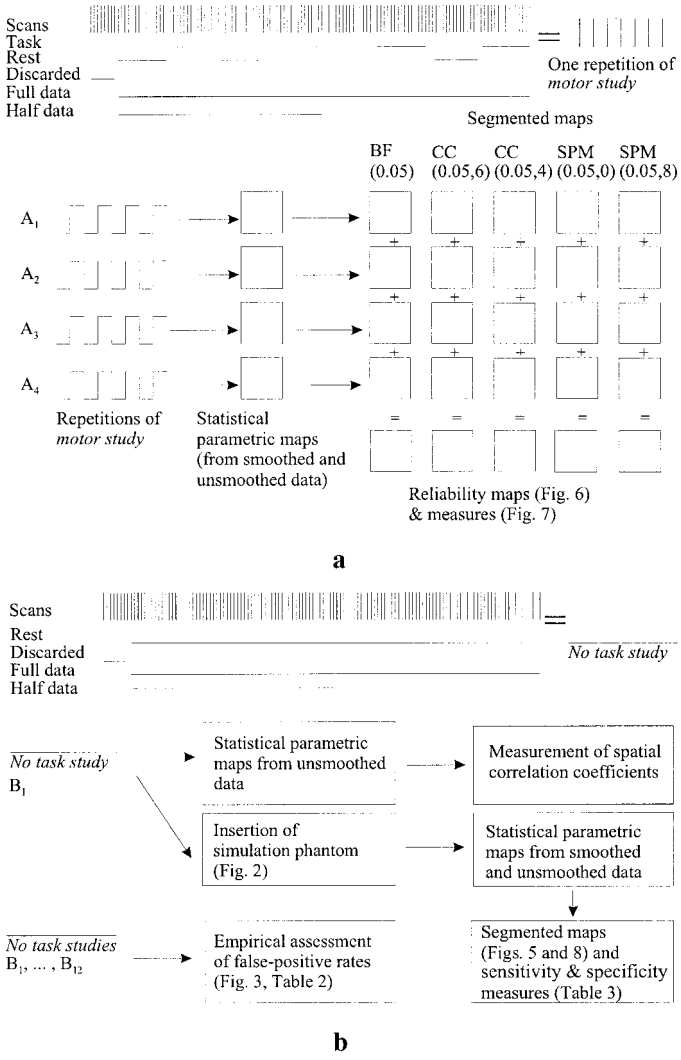

FIG. 1. Acquisition and use of data. (a) Motor task. (b) No task.

MATERIALS AND METHODS

Subjects and Data Acquisition

Magnetic resonance imaging was performed with a Siemens Vision 1.5 T MRI scanner (Siemens AG, Erlangen, Germany) at the Department of Radiology, Helsinki University Central Hospital. A set of 1-mm-thick sagittal T1-weighted images covering the whole head (field of view 256 mm; matrix 256 × 256, 180 slices) were acquired with a 3D-MPRAGE sequence (TR 9.7 ms; echo time TE 4 ms; TI 20 ms; flip angle 10°). These images were used as anatomical reference images in the visualization of the functional data. Functional MR images were acquired using a gradient-echo echoplanar imaging sequence (EPI) (TE 70 ms, TR 2.083 s, flip angle 90°, field of view 256 mm, matrix 64 × 64, 16 slices, slice thickness 3.0 mm, gap 1.0 mm). Two different types of studies were performed on a right-handed healthy volunteer: four repetitions of data acquisition during a simple motor experiment (studies A_1, A_2, A_3, A_4) and 12 without any task (studies B_1, B_2, . . . , B_{12}). The acquisition and use of the functional data is shown in Fig. 1. Between the studies A_1, A_2, A_3, A_4, and B_1 there was a pause of approximately two minutes. Studies B_2, . . . , B_{12} were acquired on another day with a 2-min pause between the studies. To minimize head movement, a head-supporting vacuum cast was used. The magnetic field was globally shimmed prior to imaging.

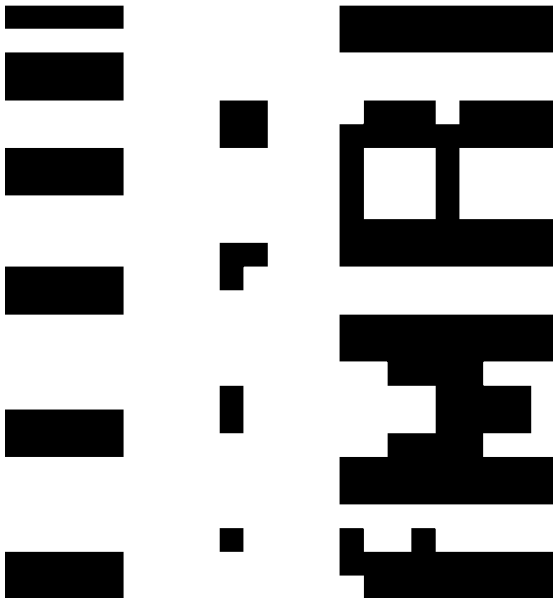

FIG. 2. One plane of the simulation phantom. The phantom consisted of three alike planes.

In the *motor studies* ($A_1, \ldots, A_4$) the volunteer was instructed to flex his right wrist during the presence of the character + on a screen and rest during the presence of the character *x*. The paradigm consisted of four rest and four motor execution blocks, or epochs, lasting 15 scans each. In addition, prior to the first rest epoch of each study, 8 scans were acquired to allow the MRI to reach a steady state of longitudinal magnetization. Thus, the total length of each study was $(8 + 15 \times 4 \times 2) \times 2.083$ s = 267 s. The data acquisition during the *no task* studies $B_1, \ldots, B_{12}$ was otherwise similar except the motor execution blocks were replaced with rest. The analysis of the studies was done using the all 120 scans (*Full data*). Part of the analysis was also done using only the first 60 scans (*Half data*).

To assess the segmentation accuracy and sensitivity of the segmentation, *simulated data* was created by adding an artificial activation pattern (Fig. 2) to the motion corrected (see the next Subsection) *no task* data B_1. The intensity values of the voxels that belong both to the activation epoch (defined as in motor execution studies) and to the activation pattern were multiplied by a value of 1.025 or 1.050, corresponding to signal increase of 2.5 or 5.0% during activation, respectively.

Computation of Statistical Parametric Maps

Statistical parametric maps were computed using the batch mode of the SPM99 software (Wellcome Department of Cognitive Neurology, London, UK, http://www.fil.ion.ucl.ac.uk/spm/) following the guidelines for a basic statistical analysis. First, the images were realigned in order to remove movement-related variance components. Sinc interpolation was used in the transformation. Next, smoothed data were generated using a Gaussian filter with a full width at half maximum FWHM = 8 mm in all three orthogonal directions. The smoothing is required when the theory of Gaussian fields is applied to set the intensity and spatial extent thresholds. On the other hand, unsmoothed data were used in the Bonferroni corrected thresholding and in the contextual clustering. The statistical parametric maps (*t* maps) were computed both from unsmoothed and smoothed data by using the general linear model. The linear model for the signal was specified to be the fixed box-car function convolved with a model hemodynamic response function (hrf). Global effects were dealt by scaling the volumes. Serial correlations were dealt with temporal filtering of the data. Cut-off period of the high-pass filter was set to 125 s. Temporal low-pass filtering was done to allow for a proper assessment of degrees of freedom.

Segmentation of Statistical Parametric Maps

The *t* maps were transformed ("Gaussianized") at each voxel *i* to *z* maps, using

$$z_i = q_{\text{norm}}[p_t(t_i, r)], \tag{1}$$

where z_i is the value in the *z* map, $q_{\text{norm}}(x)$ is the normal inverse distribution function at x, $p_t(t_i, r)$ is the cumulative distribution function for *t* distribution with *r* degrees of freedom at t_i and t_i is the value in *t* map. In the thresholding, a voxel is classified as active if

$$z_i > T_i, \tag{2}$$

where T_i is a predefined intensity threshold. Otherwise, the voxel is treated as nonactivated. We follow here the notation in which the activations have a positive mean. When spatial extent threshold T_e is used, it is required that the number of connected voxels in an activation cluster is at least T_e. In this paper, the abbreviation $SPM(p, T_e)$ will refer to the segmentation of the *z* map (computed from smoothed data) using an extent threshold T_e combined to such intensity threshold T_i that the probability of observing at least one false-positive voxel in the whole brain volume is *p*. In practice, the thresholding was done directly to *t* maps using corresponding thresholds in *t* statistics. The abbreviation $BF(p)$ will refer to the intensity thresholding made to the unsmoothed data using

$$T_i = -q_{\text{norm}}\left(\frac{p}{N_b}\right), \tag{3}$$

where N_b is number of voxels in the search volume (in the brain region). This conservative threshold, known as Bonferroni corrected threshold, gives an approxi-

mate false-positive rate p for the whole volume, assuming that there are no spatial autocorrelations.

In contextual clustering, the classification information from the neighborhood of voxels is utilized. The contextual clustering rule used in this paper is as follows (see Salli *et al.* (1999) and the Appendix): First, as an initialization step the voxels with

$$z_i > T_{cc} \tag{4}$$

are labeled as *active* and other voxels are labeled as *nonactive*, where T_{cc} is a predefined decision parameter. After the initialization, the voxels are reclassified. In the reclassification, a voxel i is considered as active if

$$z_i + \frac{\beta}{T_{cc}}(u_i - N_n/2) > T_{cc} \tag{5}$$

and otherwise as nonactive. The constant N_n is the number of neighbors in the neighborhood system. By defining a 3-D neighborhood consisting 26 closest voxels, $N_n = 26$. Variable u_i is the number of currently active neighborhood voxels for the voxel i. User specified parameter β determines the weighting of neighborhood information, and when positive, encourages neighbors to be of like class. The parameter β can be used to adjust the trade-off between sensitivity and segmentation accuracy. One way to set the β is to write

$$\beta = \frac{T_{cc}^2}{s}. \tag{6}$$

Then s is a user-specified parameter that can have any real positive value ($0 < s < \infty$). As $s \to 0$ the method approaches a recursive majority-vote classification. As $s \to \infty$ the method approaches the voxel-by-voxel thresholding. Intuitively, s can be understood, for example, as a required excess of activated voxels ($u_i - N_n/2$) in the neighborhood of voxel i to transform a value of zero to the level of decision parameter value T_{cc}. Classification rule (5) is repeated until convergence or oscillation between two states occurs. In each cycle the classification and u_i are updated. Voxels outside the brain are forced to nonactive by assigning a very small z value (e.g., −1000) to them. Correspondingly, a nonactive state is assigned to the voxels outside the image matrix. The abbreviation $CC(p, s)$ will refer to the segmentation made by contextual clustering using $N_n = 26$ and the combination of s and T_{cc} chosen so that the probability to detect at least one false-positive voxel in the search volume is approximately p.

The algorithm is similar with the iterated conditional modes (ICM) algorithm introduced by Besag (1986) with a few differences. Besag applied the algorithm for 2-D images while we are working with 3-D images. More importantly, in the original ICM algorithm it is assumed that the class densities and the number of object classes is known or can be estimated. In our approach the algorithm is used as a hypothesis testing technique. This means that a voxel is classified to an active voxel only if the statistical parametric value of a voxel adjusted with neighborhood information is significantly larger than zero. In the original form of ICM, T_{cc} would be related to the mean of the activation class. However, in our approach the T_{cc} does not model the real activation class but relates the algorithm to the desired significance level by controlling the false-positive rate. Essentially, the algorithm is capable of detecting activation regions whose median value is larger than T_{cc} but also activations containing only one voxel if the z_i value of a voxel is high enough.

Segmentation Parameters

Segmentation parameters were chosen so that the theoretical probability of false activation in the brain area would be approximately 5%. For $SPM(p, T_e)$ methods the parameters were chosen using the SPM99 software. First, the spatial extent threshold was set either to zero (no spatial extent threshold) or eight. Thereafter, using the study B_1 such an intensity threshold was searched that the corrected P value reported by SPM99 software was 0.05. For $BF(p)$ method the threshold T_i was calculated using Eq. (3).

For contextual clustering [$CC(p, s)$] method we do not currently have an analytic way to compute the parameters. Therefore simulations are needed to calculate the parameters for the desired significance level. One approach would be to fill the brain mask (search volume) with random numbers from standard normal distribution and segment the image. By repeating this procedure several times, we could iteratively find the correct parameters, but we chose to use a more approximate and general approach. Instead of using a real mask in the estimation, we ran simulations in a cubic volume of $16 \times 16 \times 16$ voxels. By repeating the simulations several times, false-positive rates for different combinations of T_{cc} and s (see Eqs. (5) and (6)) were found at voxel level. It is assumed that the false-positive rates at voxel level do not significantly depend on the size or shape of the mask and that the false-positive voxels exist at random locations independently of each other. Then the required false-positive rates at voxel level can be calculated using Eq. (3). To check the validity of the assumptions made, simulations with different search volumes and parameters T_{cc} and s were performed. The random numbers in the mask were spatially filtered to account for spatial autocorrelations of fMRI data.

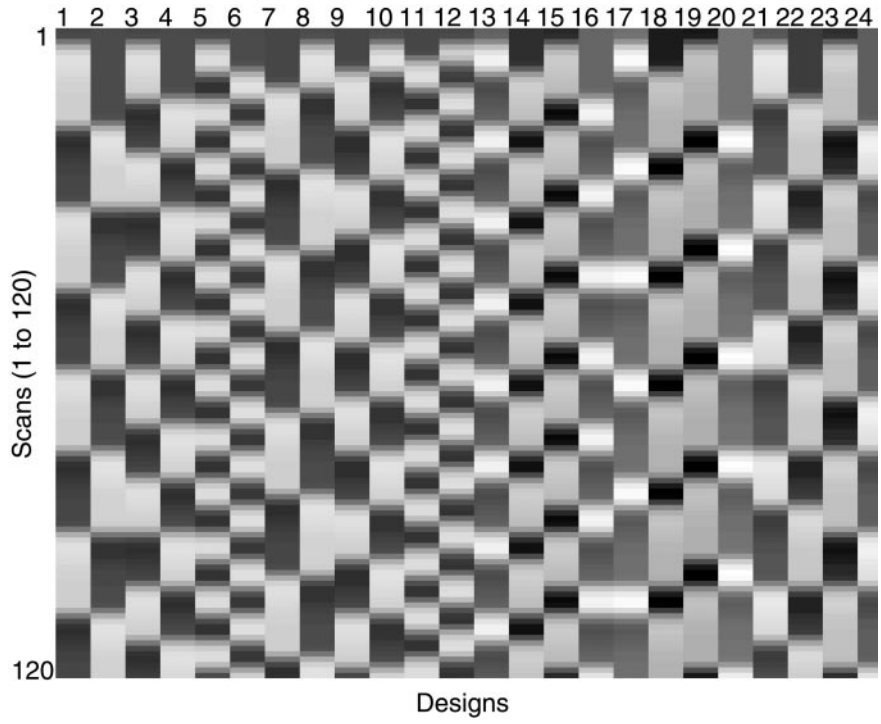

FIG. 3. Design matrices used in the empirical estimation of the false-positive rates. Designs from 1 to 6 are block designs with equal control and activation epoch lengths 15, 10, or 5 scans. These design matrices have been shifted with half epoch length in designs 7 to 12. In the designs 13 to 16 epoch lengths of 5 and 10 scans were used, either starting with the epoch of 5 scans (designs 13 and 14) or with the epoch of 10 scans (designs 15 and 16). In the designs 17 to 20 epoch lengths of 5 and 15 scans were used. In the designs 21 to 24 epoch lengths of 10 and 15 scans were used.

Empirical False-Positive Rates

Empirical whole volume false-positive rates were counted from the *Full data* versions of the fMRI rest studies $B_1, \ldots, B_{12}$. In the analysis, 24 different *epoch type* design matrices were used (see Fig. 3). The cut-off period of the high pass filter was always set to the value of two times the length of one on-off period. Otherwise, the analysis was done as explained in the Subsection "Computation of statistical parametric maps."

Reproducibility

Reproducibility of the segmentation was studied by comparing the analyzed images of four motor activation studies $A_1, \ldots, A_4$. Segmented (i.e., binary) images of all four studies were summed up voxel-by-voxel. A voxel value in the sum image represents the number of studies (from zero to four) in which the voxel is classified as active. This kind of a sum image was named as a reliability map by Genovese *et al.* (1997) and we will adopt the term here. The reliability map gives an idea about the reproducibility of the method. If all the nonzero voxels have maximum value (in this case the value four) the results have been perfectly reproducible. If all the nonzero voxels have a value of one, the results have not been reproducible at all.

The reliability maps were transformed to the coordinates of the anatomical MR image set for the visualization purposes by using slice-positioning information in the computation of the required rotations and translations.

Sensitivity and Segmentation Accuracy Analysis

Sensitivity and segmentation accuracy analysis cannot be done using real data as the true pattern of activation is unknown. Instead, the simulated phantom data and measured scanner and physiological noise from the *no task* study B_1 were used. The percentage of the detected phantom voxels was calculated. As the false-positive rates of all methods at distant locations from activations were tuned to very small and approximately to the same level, emphasis was put on the voxels near the phantom, i.e., to the accuracy of the segmentation. The numbers of false-positive voxels were counted to give a quantitative measure of the segmentation accuracy.

RESULTS

Segmentation Parameters

The counted number of brain voxels was $N_b = 12{,}000 \pm 600$. The voxelwise false-positive rates of

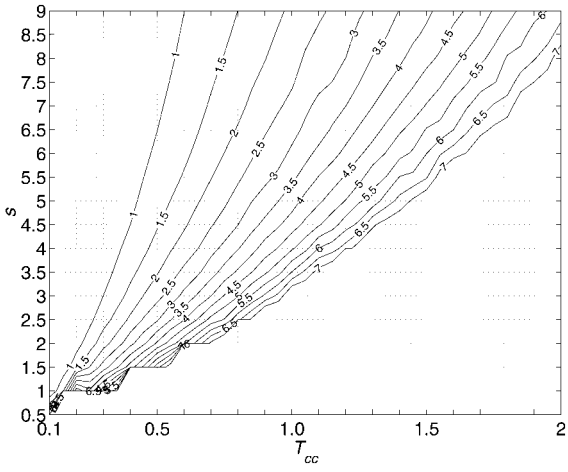

FIG. 4. False-positive rates at voxel level for contextual clustering. T_{cc} and s are the parameters of the algorithm (Eqs. (5) and (6)). Labels on the curves are $-\log P$ values, where P is the false-positive rate at voxel level. The step-wise behavior of the contours in the lower left corner is not real. It is due to the limited grid spacing and countour plotting algorithm. The results are obtained from simulated statistical parametric maps of $16 \times 16 \times 16$ voxels in size.

$CC(p, s)$ were estimated using $T_{cc} = 0.1, 0.15, \ldots, 1.95, 2.0$ and $s = 1, 1.5, \ldots, 8.5, 9$. Results are shown as a contour graph in Fig. 4. A desired false-positive rate can be selected for a chosen decision parameter T_{cc} with the aid of the contour graph. It should be noted that the false-positive rates are estimated using simulated random maps without any true activations. The likelihood of misclassifying a nonactive voxel adjacent to a truly activated voxel is higher than the corresponding estimated false-positive rate. In Fig. 5 we have illustrated the use of the contextual clustering algorithm and the dependence between segmentation parameters, sensitivity, and segmentation accuracy.

According to Bonferroni correction the required false-positive rate at the voxel level is $0.05/12{,}000 = 4.17 \times 10^{-6}$. Correspondingly, the threshold for Bonferroni corrected thresholding $BF(0.05)$ is $T_i = 4.46$ (Eq. 3). In the case of contextual clustering it is seen from Fig. 4 that parameters $(T_{cc}, s) = (1.44, 6)$ and $(T_{cc}, s) = (1.07, 4)$ lead to the false-positive rate slightly smaller than 0.05 for the whole volume. The corresponding segmentation methods are called $CC(0.05, 6)$ and $CC(0.05, 4)$, respectively. Verification of these val-

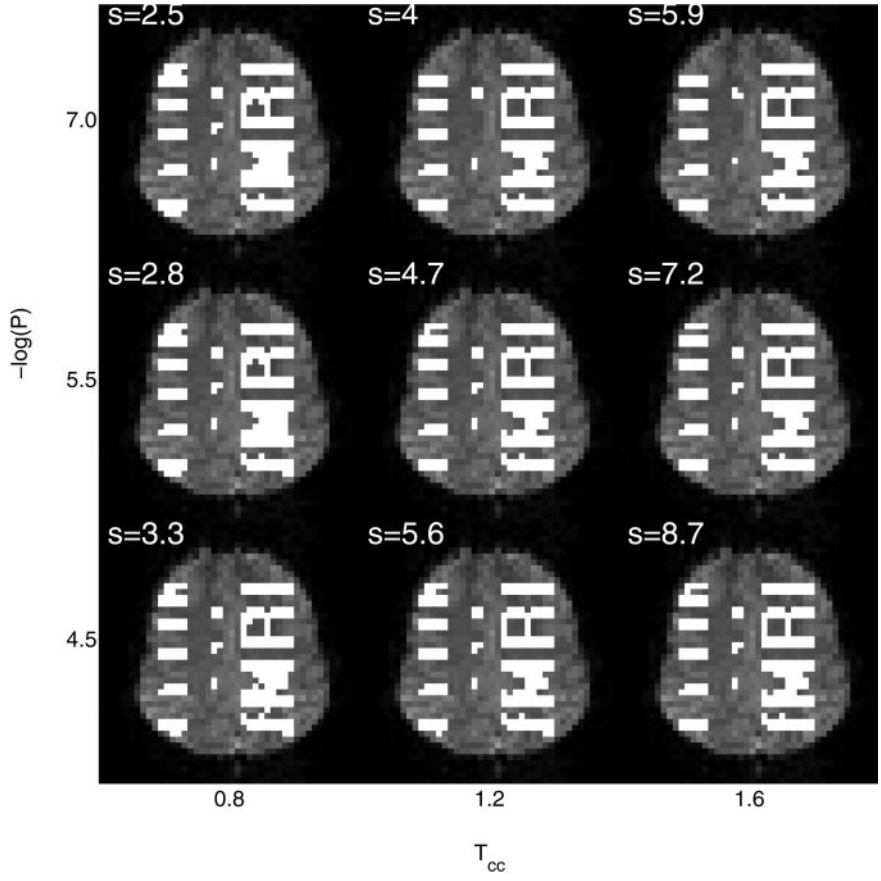

FIG. 5. Illustration of contextual clustering with different parameter values. Sensitivity decreases from left to right. False-positive rate at voxel level ($-\log P$) increases from bottom to top. The parameter s of the algorithm is obtained from Fig. 4. Unlike in the voxel-by-voxel thresholding, both sensitivity and level of significance can be defined by user when contextual clustering is used. Note that if the algorithm is tuned to detect weak activations (images on the left), the segmentation accuracy is slightly decreased.

TABLE 1

Parameters for Different Segmentation Methods to Achieve False-Positive Rate $P = 0.05$ for the Search Volumes

CC(0.05, 6)	$T_{cc} = 1.44$	$\beta = 0.3456$
CC(0.05, 4)	$T_{cc} = 1.07$	$\beta = 0.2862$
SPM(0.05, 0)	$T_1 = 4.53$	$T_e = 0$
SPM(0.05, 8)	$T_1 = 3.42$	$T_e = 8$
BF(0.05)	$T_1 = 4.46$	

Note. *CC*(*p*, *s*) refers to the contextual clustering for spatially unsmoothed data, *SPM*(*p*, T_e) to the thresholding with spatially smoothed data, and *BF*(*p*) to the Bonferroni thresholding with unsmoothed data. Parameters are for *z* maps.

ues was done using additional simulations. Pearson correlation coefficients (Milton and Arnold, 1995) between neighboring voxels in the statistical parametric map of the *no task* study B_1 were estimated. Pearson correlation coefficients in the *x*, *y*, and *z* directions were 0.18 ± 0.01. An image of $64 \times 64 \times 16$ voxels in size was filled with random values drawn from the standard normal distribution. The image was filtered using a Gaussian filter with a standard deviation of 0.5 voxels and normalized to a unit variance. The filter corresponds to a correlation coefficient of 0.24 according to our simulations. Therefore, the filter introduced slightly larger correlations than what we measured from the data. Then the filtered images were segmented. In addition to the filtered brain mask, the simulations were performed using unfiltered brain mask and using unfiltered matrices with sizes of $16 \times 16 \times 16$ and $64 \times 64 \times 64$ voxels. Each experiment was repeated 15,000 times using *CC*(0.05, 6) and *CC*(0.05, 4). The false-positive rates at voxel level were within $(0.27 \pm 0.03) \times 10^{-5}$ in all cases. The false-positive rates for the whole volume in the case of brain mask were counted, too. The false-positives rates were within 0.030 ± 0.003 whether unfiltered or filtered data were used. Hence, the segmentation methods *CC*(0.05, 4) and *CC*(0.05, 6) give a theoretical false-positive rate slightly less than 0.05 assuming that the real maps follow standard normal distribution under the null hypothesis.

The segmentation parameters for the *SPM*(0.05, T_e) method were obtained from the *Full data* version of the *no task* study B_1, using the SPM99 software as explained earlier. Due to small variations in the search volumes and smoothness estimates the corrected *P* values reported by SPM99 varied from 0.038 to 0.082 in the studies $A_1, \ldots, A_4$. The segmentation parameters of all methods are shown in the Table 1. The difference between the *SPM*(0.05, 0) and *BF*(0.05) is that in the former, spatially smoothed data and the theory of Gaussian random fields are used to derive the thresholds.

Empirical False-Positive Rates

The false-positive rates were investigated using the segmentation parameter values of Table 1. The obtained false-positive rates with different design matrices (Fig. 3) are shown in the Table 2. The false-positive rates increased as a function of average epoch length. In most cases, the measured false-positive rates were larger than the expected 5%. On average, *SPM*(0.05, 0) had the smallest false-positive rate and *SPM*(0.05, 8) the largest rate. According to visual inspection of the results many activations were located near the brain edges which indicated movement related problems despite of the movement correction. Hence, we run the analyses using the realignment parameters as covariates in the general linear model. This resulted in false-positive rates that were closer to the 5% theoretical

TABLE 2

Empirical False-Positive Rates (%)

Designs	HPF	*BF*(0.05)	*CC*(0.05, 6)	*CC*(0.05, 4)	*SPM*(0.05, 0)	*SPM*(0.05, 8)
1, 2	125 s	21	25	29	13	33
3, 4	83 s	13	13	21	13	33
5, 6	42 s	0	0	4	0	0
7, 8	125 s	33	38	42	25	46
9, 10	83 s	17	17	17	8	25
11, 12	42 s	4	4	0	0	4
13, 14	62 s	0	0	8	8	17
15, 16	62 s	8	4	8	0	8
17, 18	83 s	8	8	17	0	25
19, 20	83 s	8	8	8	4	13
21, 22	104 s	21	17	21	17	38
23, 24	104 s	17	17	17	8	29
On average		13	13	16	8	23

Note. The value in each cell is the percentage of analyses in which at least one active voxel was detected. The designs refer to Fig. 3. HPF is the cutoff period of the high pass filter used.

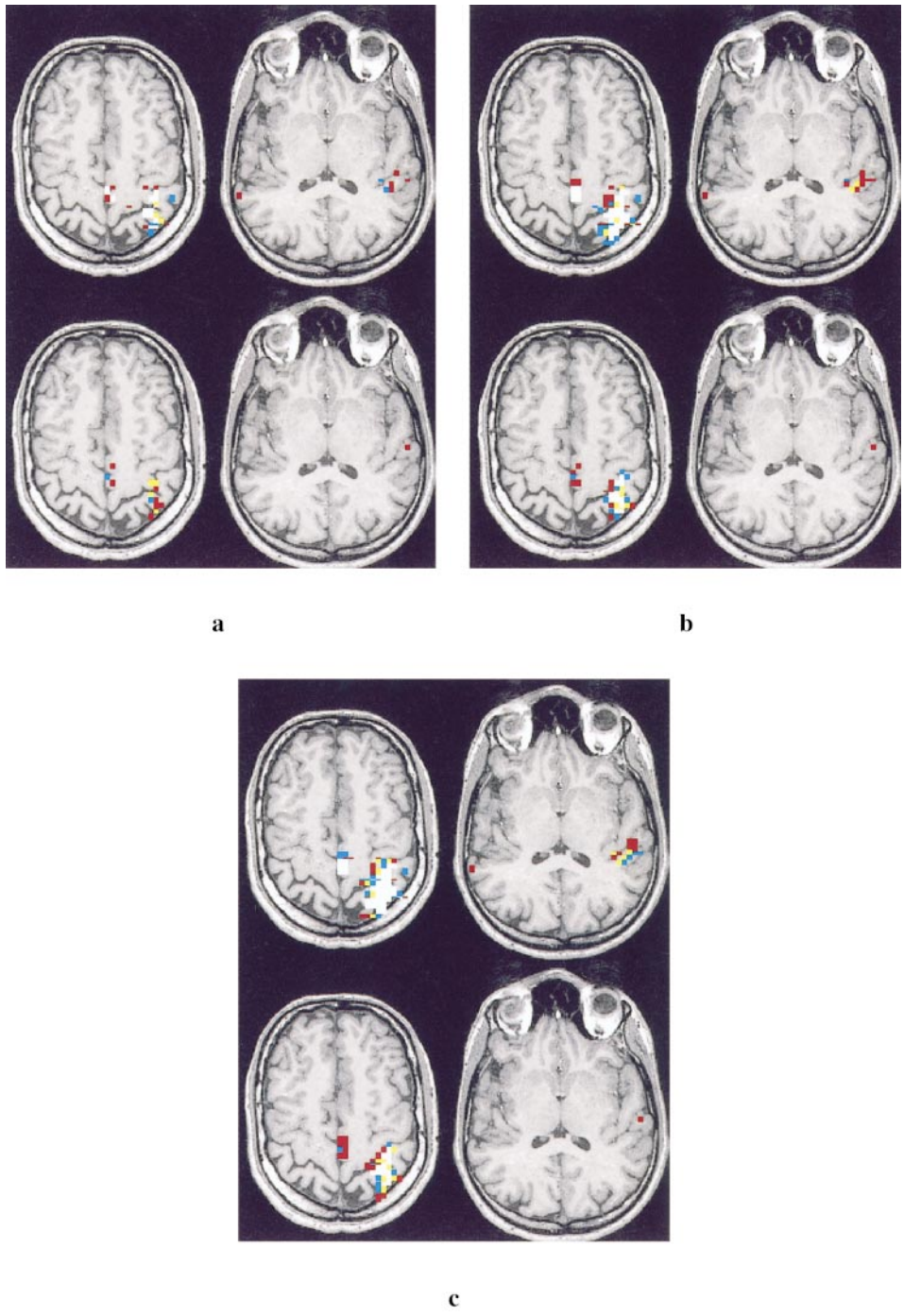

FIG. 6. Reliability maps. The voxel values in the maps represent the number of studies (R) in which the voxel is classified as active. (a) *BF*(0.05), (b) *CC*(0.05, 6), (c) *CC*(0.05, 4), (d) *SPM*(0.05, 0), (e) *SPM*(0.05, 8). Upper images in (a–e) are from the *Full data* and lower images from the *Half data.*

rate. For example, for *CC*(0.05, 6) the average error rate was 8% and for *CC*(0.05, 4) 10%.

Reproducibility

Figure 6 shows the nonzero voxels of the reliability maps superimposed on the top of anatomical MR images. The value R in the reliability maps is the number of studies (from 0 to 4) in which the voxel was classified as active. The core of the activation areas was detected in all studies and with all methods. Figure 7 shows the proportions of different values in the reliability maps. Worst performance was with the Bonferroni corrected voxel-by-voxel thresholding [*BF*(0.05)] as the proportion of $R = 3$ and $R = 4$ voxels was the lowest. Differ-

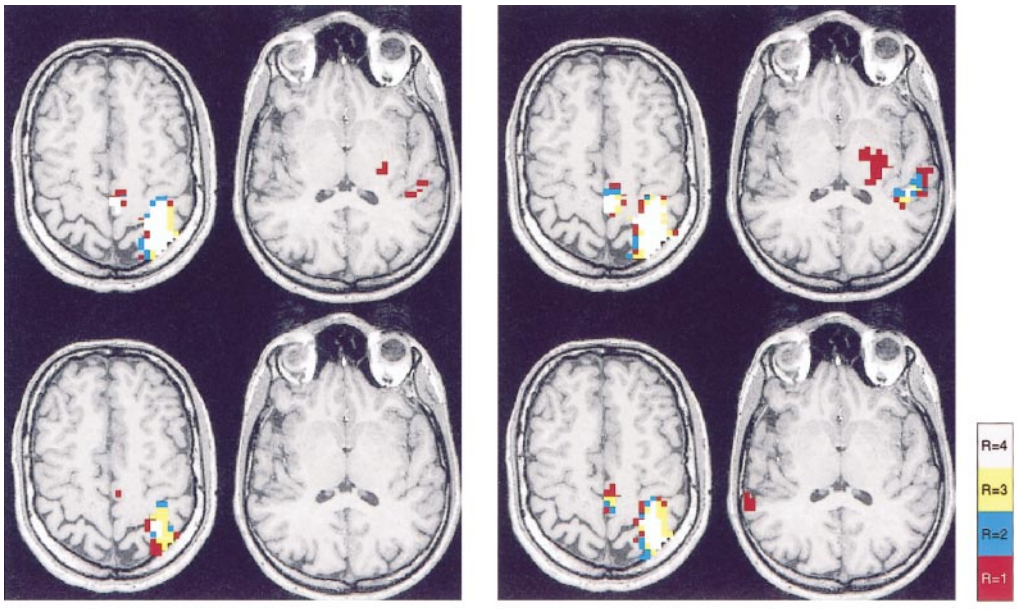

FIG. 6—*Continued*

ences between the other methods were not particularly large, i.e., the results do not indicate that *CC*(0.05, *s*) would have significantly better or worse reproducibility than *SPM*(0.05, T_e). It should be noted that large absolute number of active voxels does not necessarily mean good sensitivity. It may be due to spreading of the active regions as spatial information is used. Hence, primarily attention in the Fig. 7 should be paid on the proportions of numbers of voxels belonging to respective reliability value (*R*) groups.

Low reproducibility indicates low robustness against variations in the imaging process, e.g., against the thermal noise. The low reproducibility of the voxel-by-voxel thresholding can be explained by the fact that the thermal noise exist independently in neighboring voxels and its effect can be reduced by using information from voxels neighborhood.

Sensitivity and Segmentation Accuracy

Table 3 shows the percentages of the simulated phantom voxels detected with the different methods. The numbers of false-positives are shown for completeness, too. As expected, the sensitivity of *CC*(0.05, *s*) is

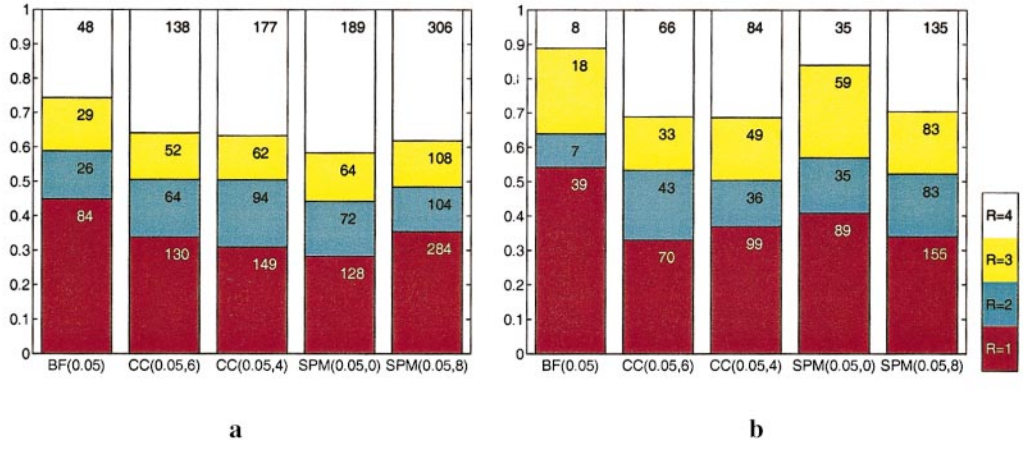

FIG. 7. Proportions of nonzero voxels in reliability maps (see Fig. 6). Beginning from the bottom, the bars correspond to values $R = 1$, 2, 3 and 4, respectively. Absolute number of voxels having a value of 1, 2, 3, or 4 are shown in bars. Results using (a) *Full data*, (b) *Half data* are shown. Note the small proportion of $R = 4$ voxels in the *BF*(0.05) method.

TABLE 3

Percentages of the Phantom Voxels Detected and the Number of False-Positive Voxels Detected with Different Segmentation Methods

	True-positives (%)				Number of false-positives			
Method	Full data		Half data		Full data		Half data	
Signal rise (%)	2.5	5.0	2.5	5.0	2.5	5.0	2.5	5.0
CC(0.05, 6)	95.9	99.1	46.4	93.1	2	2	4	8
CC(0.05, 4)	96.6	99.1	72.2	93.8	6	3	33	20
SPM(0.05, 0)	62.0	92.4	1.4	0.9	35	262	2	11
SPM(0.05, 8)	90.9	99.0	29.9	72.5	242	723	49	191
BF(0.05)	25.4	85.4	1.9	3.1	0	0	0	0

Note. Results are based on simulated data but real noise assuming 2.5% and 5.0% signal rise from baseline during activation. The false-positives are located into the neighborhood of true-positives so the ratio of false-positive and true-positive voxels can be used as a measure of segmentation accuracy.

increased as the parameter s is decreased, and the sensitivity of $SPM(0.05, T_e)$ is increased as spatial extent threshold T_e is increased. The results indicate that the most sensitive methods are *CC*(0.05, 6) and *CC*(0.05, 4). However, the difference to *SPM*(0.05, 8) is not large, and by further increasing the spatial extent threshold T_e, sensitivity might increase. With *Full data* the spatial smoothing increased sensitivity [*SPM*(0.05, 0) vs *BF*(0.05)] but with *Half data* the sensitivity was decreased. Segmentation accuracy can be assessed visually from Fig. 8. The false-positives tabulated in Table 3 may be used as a quantitative comparison of the segmentation accuracy. The results were as expected. *BF*(0.05) has the best segmentation accuracy but has low sensitivity. Lowest segmentation accuracy is achieved with spatially smoothed data.

DISCUSSION

In this article we studied the use of contextual clustering, thresholding and data smoothing in the analysis of motor fMRI experiments. Especially, the reproducibility of the results was investigated. In addition, using simulated phantom but real noise segmentation accuracy and sensitivity were studied.

Figure 9 shows the sources of variation that we have graded as the most important limiting factors in obtaining reproducible segmentation. The thermal noise follows the normal distribution well and exists independently in neighboring voxels. Therefore the effects of this noise source to the reproducibility can be successfully decreased with statistical and contextual methods. In addition to the thermal noise, the movements of head and cardiac or respiratory pulsations are significant sources of variation. In this study, we used head supporting vacuum cast and realignment algorithm in order to reduce the effect of head movements. However, it is obvious that all movement artifacts cannot be eliminated. In addition, movement correction may cause new artifacts. Physiological effects may cause drifts and temporal or spatial correlations to the data. Therefore we filtered the data in temporal do-

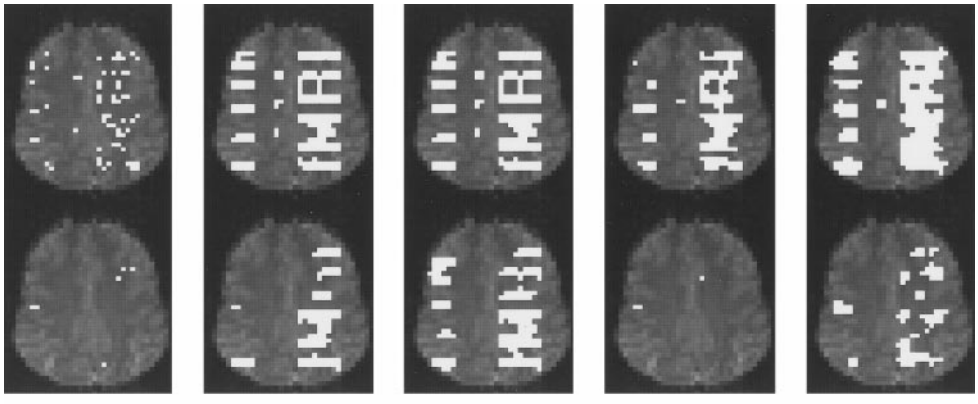

FIG. 8. Simulated phantom (2.5% signal rise) detected with various methods. In the upper row *Full data* are used and in the lower row *Half data* are used. (a) *BF*(0.05), (b) *CC*(0.05, 6), (c) *CC*(0.05, 4), (d) *SPM*(0.05, 0), (e) *SPM*(0.05, 8).

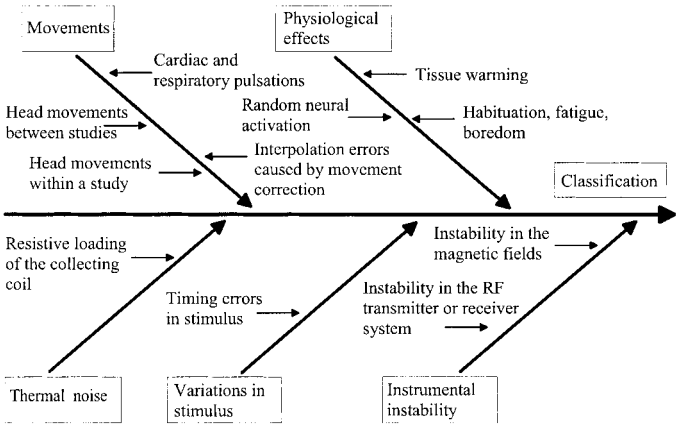

FIG. 9. Factors affecting the reproducibility visualized with an Ishikawa diagram. The Ishikawa diagram is a graphic tool used to display sources of variation in a process.

main and used reduced degrees of freedoms. The alternation of paradigm conditions at a frequency larger than the expected low-frequency drifts eliminates the contribution of the drifts to some extent. The average spatial autocorrelations were studied but found to be relatively small. However, it should be kept in mind that the correlation structure may vary significantly within the image. Some variation may arise from the timing errors between the stimulus signal and imaging sequence. In addition, the time between the stimulus-on signal and actual movement execution of a hand may slightly vary. However, these timing errors are small compared to the repetition time (TR) and to the length of a paradigm block. The variation caused by these timing errors is believed to be negligible. More serious variations may arise from system instability. A recent study by Smith *et al.* (1999) suggests that the major cause of low frequency drifts is local magnetic field instabilities in a scanner and not motion or physiological noise. Smith *et al.* (1999) observed also that while high-pass filtering will reduce the effects of drifts it may decrease the power of the true signal as well.

The use of contextual clustering reduces the effect of random noise, but does not require spatial smoothing. Although combined smoothing and the spatial extent thresholding technique may have good detection rate on spatially extended activations, the accuracy at voxel level, i.e., the ability to accurately delineate, or segment, activation areas, will inevitably be deteriorated. In the experiments with simulated activation and measured noise, better segmentation accuracy was achieved with the contextual clustering technique in comparison to the combined spatial smoothing and spatial extent thresholding. The filter width used (FWHM = 8 mm) was two times the voxel size which has been recommended as a minimum for filter width (Worsley and Friston, 1995). However, in the spatial smoothing approach adjustment between spatial accuracy and sensitivity is possible by the choice of the smoothing filter width (Worsley *et al.*, 1996; Poline and Mazoyer, 1994). Another interesting and recent approach to the problem of spatial resolution is a 2-D smoothing on the convoluted manifold of the cortex (Andrade *et al.*, 2000).

An important difference of the contextual clustering method to the intensity thresholding technique is that the desired significance level does not determine the sensitivity as one-to-one. Higher sensitivity and significance level can be achieved simultaneously by increasing the β parameter if lower segmentation accuracy is accepted. In principle, the contextual information could be utilized by doing a spatio-temporal data restoration to the raw time-series data (Descombes *et al.*, 1998). An advantage of contextual clustering over the processing of the raw data is computational efficiency. The method uses as an input only the statistical parametric map (e.g., $64 \times 64 \times 16$) instead of the whole data (e.g., $64 \times 64 \times 16 \times 120$). The time required to segment a statistical parametric map is only a few seconds with a 500 MHz Pentium III PC. The computational efficiency allows even real time data-analysis. The use of the statistical parametric map as an input also enables the use of the algorithm with any modeling method that produces statistical maps. Thus it can be easily used with many existing fMRI analysis packages, e.g., with SPM99. Computational efficiency makes the estimation of false-positive rates possible by simulations, thus allowing hypothesis testing.

It is clear that the box-car type model for the signal in the simulation experiment was a very much simplified model for the activations. However, the statistical parametric maps were always calculated using the same linear model and only the way in which spatial information was used differed. Hence, this simplification should not affect the results. Instead, the spatial shape of the simulated activations is important. Simple boxes as activation objects would be too simple and would not reveal all differences in segmentation accuracy. Therefore, more complicated shapes were used, as well.

It is evident that the simulated statistical parametric maps are only rough approximations of the actual data. Hence, measured data were used to empirically estimate the actual false-positive rates. Measured fMRI noise was used in the sensitivity and segmentation accuracy studies, too. Genovese *et al.* (1997) presented a method to estimate false- and true-positive rates from replicated experiments. Unfortunately, the assumptions made are not valid for the purposes of our study. Especially, the independence assumption of the classification of the neighboring voxels is not valid. Instead, we estimated the whole volume false-positive rates by analyzing the noise data acquired during rest. The false-positives rate exceeded the expected theoretical 5% rate in most cases. It is evident that several

factors like the segmentation and imaging parameters, subject movements, the form of the general linear model and characteristics of temporal filters affect the false-positive rates. Indeed, it was possible to decrease the number of false positives by including the estimated realignment parameters as covariates of no interest to the general linear model.

We have studied the detection of fMRI activations by utilizing contextual information. We have shown that the results are more reproducible than with Bonferroni corrected intensity thresholding for unsmoothed data. In addition, by using a simulation example, we found that also the sensitivity is better than with Bonferroni corrected thresholding. We also illustrated how the spatial smoothing of the raw data decreases the segmentation accuracy more than the contextual clustering of a statistical parametric map.

APPENDIX

We model the statistical parametric maps using the locally dependent Markov random fields (MRF) (Besag, 1974, 1986). This means that the probability for activation at a voxel is specified conditionally on the activation pattern in the neighborhood of that voxel. We will use a two-class model with pairwise interactions. Considering two segmentations which differ only at voxel i, unordered classes and pairwise interactions the conditional probability of class m occurring at voxel i, given the classes of the neighborhood voxels is

$$p_i(m|\cdot) \propto \exp[\alpha_m + \beta u_i(m)] = \exp[\beta u_i(m)], \quad (7)$$

where $u_i(m)$ is the number of neighbors of the voxel i having class m and β is an interaction parameter (Besag, 1986). Variables α_m represent the *a priori* probability of the class m. This information is not needed in the hypothesis testing and the variables α_m are set to zero. It should be noted that a positive β discourages active and nonactive classes from appearing at neighboring voxels.

The nonactive voxels z_i of a statistical parametric z map follow a Gaussian distribution

$$f(z_i|m) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left[\frac{-(z_i - \mu_m)^2}{2\sigma_m^2}\right] \quad (8)$$

with mean $\mu_0 = 0$ and variance $\sigma_0^2 = 1$. A model distribution for the active regions is somewhat arbitrarily defined to be a Gaussian with $\mu_1 = 2T_{cc}$ and $\sigma_1^2 = 1$ (see the end of the Appendix).

Let z represent the values of the n voxels of a statistical parametric map and k its segmentation. Let us assume that the values of the statistical parametric map are conditionally independent so that the conditional density of z is

$$l(z|k) = \prod_{i=1}^{n} f(z_i|k_i), \quad (9)$$

where k_i is the true classification of the voxel i. One way to segment the image z is to find the most probable segmentation for the existing map values, i.e., maximize the conditional probability $P(k|z)$. In theory, this inverse problem can be solved by applying the Bayes rule and maximizing

$$P(k|z) \propto l(z|k)p(k), \quad (10)$$

where $p(k)$ represents the prior spatial model (Besag, 1986). However, the computational requirements for finding the global maximum of this would be enormous.

An approximation to the maximum of Eq. (10) can be found using the iterated conditional modes (ICM) algorithm (Besag, 1986):

(1) As an initial classification, values k_i are chosen so that $f(z_i|k_i)$ is maximized at each voxel i separately. This means that a voxel is considered as active if $z_i > T_{cc}$.

(2) The classification is updated at each voxel separately so that the new class has maximum conditional probability, given the value z_i and current classification information from the neighborhood of voxel i. Hence, the class m for voxel i must be chosen so that the probability

$$P[m|z_i, \hat{u}_i(m)] \propto f(z_i|k_i)p_i[m|\hat{u}_i(m)] \quad (11)$$

is maximized (Besag, 1986). Here, $\hat{u}_i(m)$ is the current number of neighbors of class m and is updated at every cycle. From Eqs. (7) and (8) we get that in order to maximize Eq. (11) the voxel i is must be classified as active if

$$z_i + \frac{\beta}{T_{cc}}[\hat{u}_i(1) - N_n/2] > T_{cc} \quad (12)$$

and otherwise as nonactive. Here N_n is the constant for the total number neighbors (e.g., 26). Step (2) is repeated until the classification does not change anymore or begins to oscillate between two states. The updating can be done simultaneously for all voxels (synchronous updating) or sequentially. We have used the synchronous updating scheme to avoid directional effects.

From the definition of the model activation distribution it follows that the algorithm will be optimal for image regions having activations whose z_i values fol-

low the model distribution ($\mu_1 = 2T_{cc}$, $\sigma_1^2 = 1$). By optimal it is meant that a local maximum of probability in Eq. (10) is found. In the hypothesis testing approach the maximization is not the goal. Instead the parameter T_{cc} is set to have as low value as the desired false-positive rate allows for the chosen β. For the controlling the false-positive rate it is only important that the voxels of the nonactive image regions follow the standard normal distribution.

ACKNOWLEDGMENTS

This work was supported by The Foundation of Technology in Finland (Tekniikan edistämissäätiö), Jenny and Antti Wihuri Foundation, Radiological Society of Finland, Academy of Finland, Clinical Research Institute of Helsinki University Central Hospital, Paavo Nurmi Foundation, Sigrid Jusélius Foundation, Cancer Organizations of Finland, and Helsinki University Central Hospital Research Grants, TYH-0313, TYH-8102, TYH-9102. The authors thank Sami Martinkauppi for assistance in acquisition of fMRI data.

REFERENCES

Aguirre, G., Zarahn, E., and D'Esposito, M. 1998. A critique of the use of the Kolmogorov-Smirnov (KS) statistic for the analysis of BOLD fMRI data. *Magn. Reson. Med.* **39:** 500–505.

Andrade, A., Kherif, F., Mangin, J.-F., Worsley, K. J., Simon, O., Dehaene, S., Le Bihan, D., and Poline, J.-B. 2000. Cortical surface statistical parametric mapping. *NeuroImage* **11:** S504.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc. B* **36:** 192–236.

Besag, J. 1986. On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B* **48:** 259–279.

Casey, B. J., Cohen, J. D., O'Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., Noll, D. C., Hu, X., Lowe, M. J., Rosen, B. R., Truwitt, C. L., and Turski, P. A. 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. *NeuroImage* **8:** 249–261, doi:10.1006/nimg.1998.0360.

Constable, R. T., McCarthy, G., Allison, T., Anderson, A. W., and Gore, J. C. 1993. Functional brain imaging at 1.5 T using conventional gradient echo MR imaging techniques. *Magn. Reson. Imag.* **11:** 451–459.

Descombes, X., Kruggel, F., and von Cramon, D. Y. 1998. fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. *NeuroImage* **8:** 340–349, doi:10.1006/nimg.1998.0372.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magn. Reson. Med.* **33:** 636–647.

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., and Turner, R. 1995a. Analysis of fMRI time-series revisited. *NeuroImage* **2:** 45–53, doi:10.1006/nimg.1995.1007.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D., and Frackowiak, R. S. J. 1995b. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2:** 189–210.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., and Evans, A. C. 1994. Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.* **1:** 214–220.

Genovese, C. R., Noll, D. C., and Eddy, W. F. 1997. Estimating test-retest reliability in functional MR imaging I: Statistical methodology. *Magn. Reson. Med.* **38:** 497–507.

Kruggel, F., von Cramon, D. Y., and Descombes, X. 1999. Comparison of filtering methods for fMRI datasets. *NeuroImage* **10:** 530–543, doi:10.1006/nimg.1999.0490.

Lowe, M. J., and Sorenson, J. A. 1997. Spatially filtering functional magnetic resonance imaging data. *Magn. Reson. Med.* **37:** 723–729.

Milton, J. S., and Arnold, J. C. 1995. *Introduction to Probability and Statistics. Principles and Applications for Engineering and the Computing Sciences*, 3rd ed. McGraw-Hill, New York.

Noll, D. C., Genovese, C. R., Nystrom, L. E., Vazquez, A. L., Forman, S. D., Eddy, W. F., and Cohen, J. D. 1997. Estimating test-retest reliability in functional MR imaging II: Application to motor and cognitive activation studies. *Magn. Reson. Med.* **38:** 508–517.

Poline, J. B., and Mazoyer, B. M. 1994. Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE Trans. Med. Imag.* **13:** 702–710.

Salli, E., Visa, A., Aronen, H. J., Korvenoja, A., and Katila, T. 1999. Statistical segmentation of fMRI activations using contextual clustering. *Lect. Notes Comput. Sci.* **1679:** 481–488.

Skudlarski, P., Constable, R. T., and Gore, J. C. 1999. ROC analysis of statistical methods used in functional MRI: Individual subjects. *NeuroImage* **9:** 311–329, doi:10.1006/nimg.1999.0402.

Smith, A. M., Lewis, B. K., Ruttimann, U. E., Ye, F. Q., Sinnwell, T. M., Yang, Y., Duyn, J. H., and Frank, J. A. 1999. Investigation of low frequency drift in fMRI signal. *NeuroImage* **9:** 526–533, doi:10.1006/nimg.1999.0435.

Worsley, K. J., and Friston, K. J. 1995. Analysis of fMRI time-series revisited—Again. *NeuroImage* **2:** 173–181, doi:10.1006/nimg.1995.1023.

Worsley, K. J., Marrett, S., Neelin, P., and Evans, A. C. 1996. Searching scale space for activation in PET images. *Hum. Brain Mapp.* **4:** 74–90.