Vertebrae Localization in Pathological Spine CT via Dense Classification from Sparse Annotations

Ben Glocker[1], Darko Zikic[1], Ender Konukoglu[2],
David R. Haynor[3], and Antonio Criminisi[1]

[1] Microsoft Research, Cambridge, UK
[2] Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, MA, USA
[3] University of Washington, Seattle, WA, USA

Abstract. Accurate localization and identification of vertebrae in spinal imaging is crucial for the clinical tasks of diagnosis, surgical planning, and post-operative assessment. The main difficulties for automatic methods arise from the frequent presence of abnormal spine curvature, small field of view, and image artifacts caused by surgical implants. Many previous methods rely on parametric models of appearance and shape whose performance can substantially degrade for pathological cases.

We propose a robust localization and identification algorithm which builds upon supervised classification forests and avoids an explicit parametric model of appearance. We overcome the tedious requirement for dense annotations by a semi-automatic labeling strategy. Sparse centroid annotations are transformed into dense probabilistic labels which capture the inherent identification uncertainty. Using the dense labels, we learn a discriminative centroid classifier based on local and contextual intensity features which is robust to typical characteristics of spinal pathologies and image artifacts. Extensive evaluation is performed on a challenging dataset of 224 spine CT scans of patients with varying pathologies including high-grade scoliosis, kyphosis, and presence of surgical implants. Additionally, we test our method on a heterogeneous dataset of another 200, mostly abdominal, CTs. Quantitative evaluation is carried out with respect to localization errors and identification rates, and compared to a recently proposed method. Our approach is efficient and outperforms state-of-the-art on pathological cases.

1 Introduction

Spinal imaging is an essential tool for diagnosis, surgical planning and follow-up assessment of pathologies such as curvature disorders, vertebral fractures, or intervertebral disc degeneration. Accurate 3D images of the spinal anatomy are commonly acquired using computed tomography (CT) for details of bony structures, and magnetic resonance imaging (MRI) for high soft tissue contrast. Automated methods which support the quantitative analysis of these images are of great importance, and in this context, localization and identification of individual vertebrae is a crucial component for many subsequent tasks. Applications which immediately benefit from a *locate-and-name* algorithm include vertebra

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 262–270, 2013.

body segmentation [1], fracture detection [2], longitudinal and multi-modal registration [3], and statistical shape analysis [4]. Furthermore, reliable vertebrae identification could greatly reduce the risk of wrong-level surgery [5].

The main difficulties for automatic locate-and-name methods arise from the high variability in clinical images due to pathologies and surgical implants. Abnormal curvature, such as high grade scoliosis, and metal implants, such as rods, alter both shape and appearance significantly. In addition, spine-focused scans are commonly taken with a restricted field of view, and the lack of broader contextual information adds to the difficulty of vertebrae identification. Some exemplary images from our spine-focused database are shown in Figure 1.

Most previous methods for vertebrae localization focus on a particular part of the spine [6,7] or rely on a priori knowledge about which part is visible [8,9,10] which makes them less applicable to general, varying image data. Methods which rely on statistical models of shape and appearance [11] can struggle with pathological, abnormal cases. The high variability of both shape and appearance makes it difficult to find appropriate parametric models. An advanced method proposed in [12] is based on a quite complex and computationally demanding chain of processing steps. It has been successfully applied to narrow field of view scans similar to ours. However, the identification phase which relies on a similarity measure evaluated between an appearance model and potential vertebra candidates might not be robust to abnormal appearance caused by implants or fractures. Hierarchical methods [13] relying on anchor vertebrae require the anchors to be present (and normal/healthy enough for reliable detection).

In order to overcome these limitations, we propose a vertebrae locate-and-name approach based on classification forests avoiding explicit parametric modeling of appearance. The demand of classification for dense annotations, which can be tedious to acquire, is overcome by employing a semi-automatic labeling strategy where sparse centroid annotations are transformed into dense probabilistic labels. Based on these labels, we learn a discriminative classifier exploiting local and short-range contextual features which shows robustness in the presence of typical characteristics of spinal pathologies and image artifacts. In our locate-and-name system, we make no assumptions regarding which and how many vertebrae are visible in a patient scan. We achieve good performance for high pathological cases where other approaches may fail.

After presenting technical details in Section 2, and providing extensive evaluation on over 400 CT scans in Section 3, we conclude our paper in Section 4.

2 Dense Classification from Sparse Annotations

As our vertebrae locate-and-name system is based on supervised, discriminative learning we start by formalizing the training data and introduce necessary notations. We assume the availability of a training database $\mathcal{T} = \{(I_k, \mathcal{C}_k)\}_{k=1}^{K}$ with K pairs of an image $I : \Omega_I \subset \mathbb{R}^3 \to \mathbb{R}$ and a set of annotated vertebrae centroids $\mathcal{C} = \{\mathbf{c}_v\}$ with $\mathbf{c}_v \in \Omega_I$ and $v \in \mathcal{V}$. The set of vertebrae is defined as $\mathcal{V} = \{C_1, \ldots, C_7, T_1, \ldots, T_{12}, L_1, \ldots, L_5, S_1, S_2\}$ which contains the regular 7 cervical, 12 thoracic, 5 lumbar and two additional centroids on the sacrum.

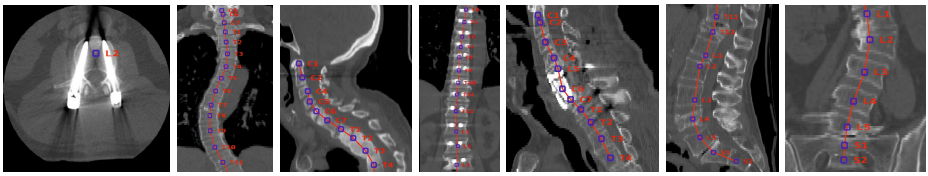


Fig. 1. Overview of the variability in our spine database with abnormal curvatures, small field of view, and surgical implants. Overlaid are localization results for our algorithm which is robust to pathologies and artifacts. Note, for simultaneous visibility of all vertebrae, we show warped slices for the sagittal and coronal views. Warping is performed with thin-plate splines which map the estimated centroids onto a plane.

A recently proposed method for vertebrae localization is making use of the above input for training a centroid regression forest with a subsequent parametric refinement step [11]. The key idea in that work is that every point in an image can vote for the position of all vertebrae. This seems to work quite well for non-pathological cases and "regular" CT scans where broad contextual information is available. However, this context is not available in spine-focused scans with narrow field of view (see Figure 1). Additionally, for pathological cases it is questionable if such an "outside-in" approach where non-spinal image points equally vote for vertebrae locations can capture abnormal curvature. Our experiments suggest that this is not the case.

In order to overcome these limitations, we propose a centroid classification approach which is quite different in nature. Instead of directly trying to regress the location, which seems to be a much harder problem in presence of pathologies, we learn a classifier based on local and short-range contextual features which is able to predict the probability for a given image point of being a particular vertebra centroid. Formally, we want to learn the posterior distribution $p(v|f(\mathbf{x}))$ where $f(\mathbf{x})$ are features extracted at $\mathbf{x} \in \Omega_J$ in a test image J.

A popular method for learning such distributions is randomized classification forests [14]. Taking as input a database of point-wise labeled images, during decision tree training a set of discriminative features is extracted automatically from a large pool of random appearance-based features. This feature extraction is performed through supervised, hierarchical clustering of the training data with respect to an objective function which favors compact clusters of image points having equal labels. At every level of the hierarchy a binary test on the most discriminative feature is determined such that the incoming subset of training data is split into more compact clusters. A standard approach is to employ Shannon entropy over the label distributions as a measure of cluster compactness. In the leaves of the decision trees empirical distributions over incoming subsets of training data are stored. At test time, the same hierarchy of feature tests is applied and for each tree an empirical prediction is made based on the leaf that is reached. The final probabilistic prediction of the forest is determined by

simple averaging over tree predictions. The randomness injected in the training phase yields decorrelated decision trees, which has been shown to improve generalization over alternative techniques [15].

Dense Labels from Sparse Annotations: A general problem with such a supervised classification approach is that point-wise labeled training data is required. Obtaining such data can be quite tedious and time-consuming. To overcome this costly demand, we employ a labeling strategy which transforms sparse centroid annotations into dense probabilistic labels. This strategy is similar to the ones used for semi-supervised forests previously used for video segmentation [16] and organ localization[17]. Given a set of annotated centroids $\mathcal{C}$ for a training image I, we define a centroid likelihood function for each vertebra as

$$\psi_v(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{c}_v - \mathbf{x}\|^2}{h_v}\right) \quad \text{with} \quad \mathbf{x} \in \Omega_I \ . \tag{1}$$

The parameter h_v controls the vertebra specific spatial influence of each centroid. The intuition behind this definition is that points close to a centroid get a high likelihood, while those values decrease with larger Euclidean distance. In order to be able to discriminate non-vertebra points, we introduce a likelihood function for a background label B as $\psi_B(\mathbf{x}) = 1 - \max_v \psi_v(\mathbf{x})$. Based on these likelihood definitions, we obtain a labeling distribution as

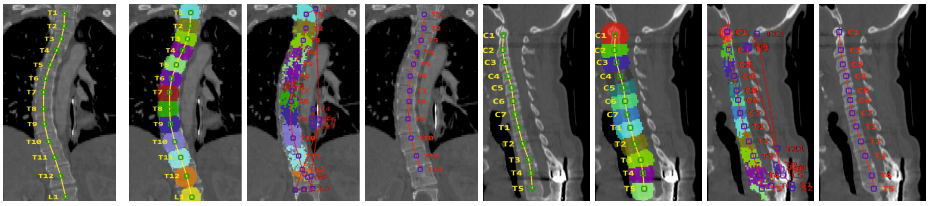
$$p(l|\mathbf{x}) = \frac{\psi_l(\mathbf{x})}{\sum_{m \in \mathcal{L}} \psi_m(\mathbf{x})} \quad \text{with} \quad l \in \mathcal{L} = \mathcal{V} \cup \{B\} \ . \tag{2}$$

This strategy allows us to generate the necessary annotations for training a classification forest. The dense labels are obtained from sparse centroid annotations, which are propagated to neighboring sites. The outcome of the labeling is shown for two examples in Figure 2(b,f). In contrast to [16], we found an overall best performance of our method when training is carried out on hard labels, *i.e.* the ones with highest probability given by $\hat{l} = \arg\max_l p(l|\mathbf{x})$, instead of using weighted soft labels. We believe a reason for this might be due to overly smooth empirical histograms yielding ambiguous optimal splitting choices during tree training. The effect of hard versus soft training labels in forest learning is interesting and worth further investigation.

Centroid Estimation: Applying the learned forest classifier on an unseen test image J produces a probabilistic label map $P : \Omega_J \to \mathbb{R}^{|\mathcal{L}|}$ where for each image point $\mathbf{x} \in \Omega_J$ we obtain vertebrae probabilities $p(v|f(\mathbf{x}))$. For localization of individual vertebrae centroids, we define a centroid density estimator as

$$d_v(\mathbf{x}) = \sum_{i=1}^{N} p(v|f(\mathbf{x}_i)) \, \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{h_v}\right) \ , \tag{3}$$

where $\{\mathbf{x}_i\}_{i=1}^N$ are image points for which $p(v|f(\mathbf{x}_i)) > 0$. Using a local mode seeking algorithm based on mean shift [18], we determine the vertebrae centroids



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. On two examples, we illustrate the different steps of our algorithm. **(a,e)** Manual centroid annotations are overlaid on the original CT image. **(b,f)** The corresponding dense labels with highest probability generated from the sparse centroids and used for training. **(c,g)** At test time, the classification forest (not trained on (b,f)) produces probabilistic label estimates for every image point. We show the labels with highest probability and overlay the centroid estimates after running mean shift. **(d,h)** Employing our false positive removal strategy, only confident centroids are kept for the final localization and identification results.

$\mathcal{C}$ as the image points with highest density $\mathbf{c}_v = \arg\max_{\mathbf{x}} d_v(\mathbf{x})$. Here, we are using the same vertebra specific kernel width h_v earlier defined in Equation (1). Due to the local nature of mean shift, we randomly initialize the mode seeking algorithm for each vertebra with different starting points. These points are drawn from a distribution of image points with a centroid probability $p(v|f(\mathbf{x}))$ above a certain threshold. We found that using 50 random seeds is sufficient for a robust estimation. Computational time can be reduced by limiting N to a few thousand points, the ones with the N highest centroid probabilities. Figure 2(c,g) illustrates the probabilistic maps and the outcome of the centroid estimation. Note that only the labels with highest probability are shown, while the mode seeking algorithm operates on the weighted soft labels obtained from the forest.

Confident False Positive Removal: Due to the probabilistic nature of the forest classifier, one can obtain spurious outputs indicating the presence of a vertebra which is actually not visible (false positives). In fact, a single image point obtaining a non-zero probability for a non-visible vertebra is sufficient to produce a incorrect centroid estimate. In order to robustly eliminate such wrong estimates, we employ a false positive removal strategy, which combines the centroid density estimates based on vertebra appearance, with a shape support term. The latter is based on learned global shape characteristics of the spine. Given a set of estimated centroids $\mathcal{C}$, we define the shape support term as

$$\phi(\mathbf{c}_v) = \frac{1}{|\mathcal{C}| - 1} \sum_{\mathbf{c}_w \in \mathcal{C} \setminus \{\mathbf{c}_v\}} \mathcal{N}\left(\|\mathbf{c}_v - \mathbf{c}_w\|_2 \,; \mu_{vw}, \sigma^2_{vw}\right) \quad . \tag{4}$$

Here, $\mathcal{N}(\mu_{vw}, \sigma^2_{vw})$ are normal distributions over Euclidean distances between two different vertebrae v and w. The mean and variance of each distribution is estimated using maximum likelihood on training annotations. The sum over all possible pairs determines the "plausibility" of the estimate $\mathbf{c}_v$ given the locations

of all other centroids $\mathcal{C}\backslash\{\mathbf{c}_v\}$. In combination with the centroid density estimates this yields our joint shape and appearance confidence measure:

$$\rho(\mathbf{c}_v) = \widehat{d}(\mathbf{c}_v)\,\widehat{\phi}(\mathbf{c}_v) \quad . \tag{5}$$

Here, $\widehat{d}(\mathbf{c}_v)$ and $\widehat{\phi}(\mathbf{c}_v)$ are normalized versions of Equations (3) and (4):

$$\widehat{d}_v(\mathbf{c}_v) = \frac{d_v(\mathbf{c}_v)}{\max_{v\in\mathcal{C}} d_v(\mathbf{c}_v)} \quad \text{and} \quad \widehat{\phi}(\mathbf{c}_v) = \frac{\phi(\mathbf{c}_v)}{\max_{v\in\mathcal{C}} \phi(\mathbf{c}_v)} \quad .$$

Based on a learned threshold κ, which is determined via cross-validation, we only accept centroid estimates with a confidence $\rho(\mathbf{c}_v) > \kappa$. The proposed false positive removal strategy is appealing as it is purely data-driven and makes use of the probabilistic nature of the underlying methods. Free parameters are learned from training data and thus, no manual tuning is necessary. In Figure 2(d,h) the outcome of the centroid estimation is shown after removal of false positives which are still present in Figure 2(c,g). Even when the majority of initial estimates are false positives, the removal strategy can detect them and preserves the true ones.

3 Experiments

We evaluate the performance of our algorithm on two different, large databases. The first one contains 224 spine-focused, *i.e.* tightly cropped, CT scans of patients with varying pathologies. These include abnormal curvature, such as high-grade scoliosis and kyphosis, fractures, and numerous post-operative cases where surgical implants cause severe image artifacts. Different images capture different parts of the spine depending on the pathology. In a few scans the whole spine is visible, while in most scans the view is limited to 5-15 vertebrae. The second database consists of 200, mostly abdominal, "normal" CT scans where the reason for imaging was not necessarily indicated by a spinal pathology. These scans exhibit varying field of view along the patient's body axis, and typically capture the entire anatomy in the axial view. Vertebrae centroids have been manually annotated in all 424 scans.

As the imaging characteristics are quite different between the two databases, we perform two separate evaluations. In both cases, the datasets are randomly split into two equally sized sets, where each of the sets is once used for training and once for testing. Parameters are fixed throughout all experiments. Each classification forest consists of 20 trees, trained to a maximum depth of 24. Tree growing is stopped earlier if the number of training points falls below a threshold of 8 samples to reduce overfitting. At each tree node, we evaluate 200 random features drawn from a global tree-specific pool of 2000 random features. Feature types correspond to commonly used local and contextual average intensity and intensity difference features efficiently implemented via integral images [17]. The range of the contextual features is limited to a radius of 4 cm.

We compare our method to a recently proposed regression approach which employs subsequent refinement using a parametric shape and appearance model

Table 1. Localization errors in mm for our method and a baseline RF+HMM [11]. Evaluation is carried out on two large databases. The "Normal CT" database consists of 200 mostly abdominal scans, while "Spine CT" includes 224 high pathological cases.

Method		Regression Forests + HMM [11]				Ours			
Data	Region	Median	Mean	Std	Id.Rates	Median	Mean	Std	Id.Rates
Normal CT	All	**5.4**	**9.7**	**11.2**	**80%**	7.6	11.5	14.1	76%
	Cervical	6.5	8.2	6.1	73%	**6.3**	**7.7**	**4.4**	**78%**
	Thoracic	**5.5**	**9.9**	**10.8**	**77%**	8.7	12.4	11.6	67%
	Lumbar	**5.3**	**9.4**	**12.0**	**86%**	6.6	10.6	16.9	**86%**
Spine CT	All	14.8	20.9	20.0	51%	**8.8**	**12.4**	**11.2**	**70%**
	Cervical	11.5	17.0	17.7	54%	**5.9**	**7.0**	**4.7**	**80%**
	Thoracic	12.7	19.0	20.5	56%	**9.8**	**13.8**	**11.8**	**62%**
	Lumbar	23.2	26.6	19.7	42%	**10.2**	**14.3**	**12.3**	**75%**

[11]. Parameters for this method, denoted as RF+HMM, are the same as in [11]. Both the regression forest and the shape and appearance model are trained on the same data as our classification approach. For quantitative evaluation, we compute localization errors in millimeters, and define identification rates as in [11]. A vertebra is defined as correctly identified if its estimated centroid is within 2 cm of the true one, and the closest one is the correct one.

Table 1 summarizes the quantitative results. On the "normal" CT database, the parametric approach RF+HMM performs slightly better than our method. As the spinal anatomy in these scans is rather healthy (*i.e.* non-spine patients), the restrictive shape and appearance model of [11] seems to help. Still, it is remarkable that we achieve similar results based on fewer assumptions. The outcome is quite different for the pathological spine CTs. Here, our method clearly outperforms the parametric approach, and we achieve localization errors which are not too far from the ones obtained on the normal database. The performance of RF+HMM degrades significantly, although the algorithm has been specifically trained on pathological data. It is worth mentioning that our method is also more efficient than RF+HMM, with only 1 minute total computation time for a typical scan ($512^2 \times 200$) when running our C# code on a standard desktop PC (Intel Xeon 2.27GHz, 12 GB RAM). Some visual results are shown in Figure 1.

4 Conclusion

We have shown that with a classification approach it is possible to achieve reasonable vertebrae localization and identification results in the presence of pathologies. In particular, making as few assumptions as possible about shape and appearance seems to be the right direction when dealing with abnormal cases. Future work will focus on improving the centroid estimation by employing intervertebral constraints. To facilitate research in the domain of spinal imaging, our spine CT dataset including manual annotations is available at `http://research.microsoft.com/projects/medicalimageanalysis/`.

References

1. Ben Ayed, I., Punithakumar, K., Minhas, R., Joshi, R., Garvin, G.J.: Vertebral Body Segmentation in MRI via Convex Relaxation and Distribution Matching. In: MICCAI 2012, Part I. LNCS, vol. 7510, pp. 520–527. Springer, Heidelberg (2012)
2. Yao, J., Burns, J.E., Munoz, H., Summers, R.M.: Detection of Vertebral Body Fractures Based on Cortical Shell Unwrapping. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 509–516. Springer, Heidelberg (2012)
3. Steger, S., Wesarg, S.: Automated Skeleton Based Multi-modal Deformable Registration of Head&Neck Datasets. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 66–73. Springer, Heidelberg (2012)
4. Lecron, F., Boisvert, J., Mahmoudi, S., Labelle, H., Benjelloun, M.: Fast 3D Spine Reconstruction of Postoperative Patients Using a Multilevel Statistical Model. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 446–453. Springer, Heidelberg (2012)
5. Hsiang, J.: Wrong-level surgery: A unique problem in spine surgery. Surg. Neurol. Int. 2(47) (2011)
6. Ma, J., Lu, L., Zhan, Y., Zhou, X., Salganicoff, M., Krishnan, A.: Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 19–27. Springer, Heidelberg (2010)
7. Oktay, A.B., Akgul, Y.S.: Localization of the Lumbar discs using machine learning and exact probabilistic inference. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 158–165. Springer, Heidelberg (2011)
8. Schmidt, S., Kappes, J., Bergtholdt, M., Pekar, V., Dries, S., Bystrov, D., Schnörr, C.: Spine detection and labeling using a parts-based graphical model. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 122–133. Springer, Heidelberg (2007)
9. Huang, S.H., Chu, Y.H., Lai, S.H., Novak, C.L.: Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. TMI 28(10), 1595–1605 (2009)
10. Kelm, B.M., Zhou, S.K., Suehling, M., Zheng, Y., Wels, M., Comaniciu, D.: Detection of 3D spinal geometry using iterated marginal space learning. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) MICCAI 2010. LNCS, vol. 6533, pp. 96–105. Springer, Heidelberg (2011)
11. Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E.: Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 590–598. Springer, Heidelberg (2012)
12. Klinder, T., Ostermann, J., Ehm, M., Franz, A., Kneser, R., Lorenz, C.: Automated model-based vertebra detection, identification, and segmentation in CT images. MedIA 13(3), 471–482 (2009)
13. Zhan, Y., Maneesh, D., Harder, M., Zhou, X.S.: Robust MR Spine Detection Using Hierarchical Learning and Local Articulated Model. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 141–148. Springer, Heidelberg (2012)

14. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
15. Caruana, R., Karampatziakis, N., Yessenalina, A.: An empirical evaluation of supervised learning in high dimensions. In: ICML, pp. 96–103 (2008)
16. Budvytis, I., Badrinarayanan, V., Cipolla, R.: Semi-supervised video segmentation using tree structured graphical models. In: CVPR, pp. 2257–2264 (2011)
17. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in CT volumes. In: MICCAI Workshop on Probabilistic Models for Medical Image Analysis (2009)
18. Cheng, Y.: Mean shift, mode seeking, and clustering. PAMI 17(8), 790–799 (1995)