CLASSIFICATION WITH REJECT OPTION USING CONTEXTUAL INFORMATION

Filipe Condessa[1,2,4], *José Bioucas-Dias*[1,2],
Carlos A. Castro[5], *John A. Ozolek*[6], *and Jelena Kovačević*[3,4]

[1]Inst. de Telecomunicações & [2]Dept. of ECE Inst. Superior Técnico
Technical University of Lisbon, Lisbon, Portugal
[3]Dept. of BME and Center for Bioimage Informatics, [4]Dept. of ECE
Carnegie Mellon University, Pittsburgh, PA, USA
[5]Dept. of Obstetrics and Gynecology, Magee-Womens Research Inst., Foundation Univ. of Pittsburgh
[6]Dept. of Pathology, Children's Hospital of Pittsburgh
University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

ABSTRACT

We propose a new algorithm for classification that merges classification with reject option with classification using contextual information. A reject option is desired in many image-classification applications requiring a robust classifier and when the need for high classification accuracy surpasses the need to classify the entire image. Moreover, our algorithm improves the classifier performance by including local and nonlocal contextual information, at the expense of rejecting a fraction of the samples. As a probabilistic model, we adopt a multinomial logistic regression. We use a discriminative random model for the description of the problem; we introduce reject option into the classification problem through association potential, and contextual information through interaction potential. We validate the method on the images of H&E-stained teratoma tissues and show the increase in the classifier performance when rejecting part of the assigned class labels.

Index Terms**—** image classification, reject option, discriminative random fields

1. INTRODUCTION

Classification is a ubiquitous image processing task that aims to separate a group of objects into different classes. In classification, small, unbalanced, or incomplete training sets can lead to low performance of the classifier. As in many applications, the need for high accuracy surpasses the need to classify all the samples, in those applications, we classify while rejecting a portion of the class labels [1]. To aid in the process, one can exploit the similarity between samples and the spatial context, providing useful cues when classifying. Classifying with rejection as well as using spatial contextual information are both applicable in the automated identification of tissues in histopathology [2, 3, 4], where the cost of creating a large and representative training set is high, the presence of unknown classes is a possibility, and the similarity between tissues belonging to the same class is high.

Our goal is thus to **improve classifier performance by adding a reject option and contextual information** as follows:

- Partition the image; perform feature extraction in each partition.
- Estimate class probabilities for each partition and the risk associated with each class.
- Classify each partition using contextual information and the reject option.

The paper is organized as follows: Section 2 provides the background on partitioning, feature extraction and classification techniques. Section 3 introduces the concept of rejection while Section 4 describes our classification method with reject option using contextual information. Section 5 presents experimental results and Section 6 concludes the paper.

2. BACKGROUND

We now briefly describe the background necessary for our work in terms of image partitioning, feature extraction procedures, and the classification methods used. A generic classifier consists of a feature extraction function that maps a high-dimensional input space into a lower-dimensional feature space, and a classification function that maps the feature space into the class label space. The partitioning of the image can be considered as a preprocessing method, mapping the image space into the partition space.

Partitioning To reduce the dimensionality of the problem, we partition the data, which also allows us to efficiently use graph-based methods. The partitioning of the image is performed by oversegmenting the image and creating superpixels [5]; this allows us not only to reduce the dimension of a problem by a factor of the order of 1000, but also guarantees that there is class and appearance coherence in each partition, due to the small dimension of each superpixel (each superpixel is of average size 5×10^2 pixels, corresponding to 4×10^3 partitions in a 1600×1200 image). One drawback of this partitioning method is the nonuniformity of the partitions in shape and size.

Feature Extraction We use two different types of features in our work: application-specific features and similarity features. We normalize features from each partition by its mean and standard deviation. As application-specific features we use the *histopathology vocabulary (HV)* containing features with physiological relevance,

The authors gratefully acknowledge support from the NSF through award 1017278 and the CMU CIT Infrastructure Award. Work partially supported by grant SFRH/BD/51632/2011, from Fundação para a Ciência e Tecnologia and the CMU-Portugal (ICTI) program.

designed based on expert knowledge [4, 3]: nucleus size (1D), nucleus eccentricity (1D), nucleus density (1D), nucleus color (3D), red blood cell coverage (1D), and background color (3D). Similarity features reflect similarities not taken into account by the HV; we use the image color.

Classification Given the set of partitions $\mathcal{S}$ and associated sets of features, we want to classify each partition into a single class. As we need probability estimates for that task, we use *multinomial logistic regression (MLR)* [6]. A linear combination of kernels is used as the nonlinear regression function to include both the application-specific as well as the similarity features.

Multinomial Logistic Regression. We model the *a posteriori* probability $p(y_i \mid f_i, W)$ with MLR, where f_i is the feature vector of the ith partition, $\mathcal{L} = \{1, 2, \ldots, N\}$ set of class labels that can be assigned to each partition, $y_i \in \mathcal{L}$ the class label assigned to the ith partition, and W the regression matrix with columns w_i, $i = 1, 2, \ldots, N$ (since W is translation invariant, we arbitrarily set $w_N = 0$). Let $\mathcal{T}$ be the set of indices of the partitions present in the training set and $F_\mathcal{T} = \{f_i\}_{i \in \mathcal{T}}$. Then, the MLR models *a posteriori* probabilities as

$$p(y_i = \ell \mid f_i, F_\mathcal{T}, W) \;=\; \frac{e^{w_\ell^T k(f_i, F_\mathcal{T})}}{\sum_{j=1}^N e^{w_j^T k(f_i, F_\mathcal{T})}},$$

where $k(f_i, F_\mathcal{T})$ is a kernel vector obtained by concatenating two length-$|\mathcal{T}|$ vectors of application-specific features $k_c(f_i, F_\mathcal{T})$ and similarity features $k_s(f_i, F_\mathcal{T})$.

LORSAL. To avoid overfitting and ensure generalization capacity of the classifier, we adopt an element-wise independent Laplacian prior for the MLR vector and compute the maximum a posteriori estimate of W by solving the optimization problem

$$\hat{W} \;=\; \arg\max_{W} \left(l(W) + \log p(W) \right), \tag{1}$$

with $\quad l(W) \;=\; \sum_{i \in \mathcal{S}} \log p(y_i \mid f_i, W), \qquad p(W) \;=\; \alpha\, e^{-\lambda \|W\|_1},$

where λ is the regularization parameter. Note that $\log p(W)$ is proportional to the negative of the ℓ_1 norm of W, which promotes sparse regression matrices, controlled by λ. The maximization (1) is performed with the *LOgistic Regression via Splitting and Augmented Lagrangian* (LORSAL) algorithm [7], which solves the equivalent problem

$$\arg\max_{W, \Omega} \left(l(W) + \log p(\Omega) \right), \quad \text{subject to } W \;=\; \Omega.$$

3. REJECT OPTION

To improve accuracy at the expense of not classifying all the partitions, we classify while rejecting. Let $\mathcal{L}' = \mathcal{L} \cup \{N+1\}$ be an extended set of partition class labels with an extra label. We consider two different class rejection concepts: *uninteresting class*, present in the training set, trained as a regular class and uninteresting to the observer (corresponding to the class label $y_i = 1$), and *unknown class*, which arises from an inability of the classifier to correctly classify all labels (corresponding to the class label $y_i = N+1$).

Let $\mathcal{C}$ and $\mathcal{I}$ be the sets of labels of correctly and incorrectly classified samples, respectively, and $r_\mathcal{C}$ and $r_\mathcal{I}$ be the rejection ratios for correctly and incorrectly classified samples, respectively. Define the accuracy of nonrejected samples as

$$A_{\text{nr}} \;=\; \frac{(1 - r_\mathcal{C})|\mathcal{C}|}{(1 - r_\mathcal{I})|\mathcal{I}| + (1 - r_\mathcal{C})|\mathcal{C}|},$$

with the rejection ratio

$$r \;=\; \frac{r_\mathcal{I}|\mathcal{I}| + r_\mathcal{C}|\mathcal{C}|}{|\mathcal{I}| + |\mathcal{C}|}.$$

Note that r, $r_\mathcal{C}$ and $r_\mathcal{I}$ are, respectively, the estimates for the probability of rejection, conditional probability of rejection given that the partition is correctly classified, and conditional probability of rejection given that the partition is incorrectly classified. A_{nr} is the estimate of the conditional probability of correctly classifying given that the partition was not rejected. Ideally, $r_\mathcal{I} = 1$ and $r_\mathcal{C} = 0$, that is, reject all of the incorrectly classified samples and none of the correctly classified ones.

Reject Option by Risk Minimization Let $\hat{p}(f_i, \hat{W}) = [\hat{p}(y_i = 1 \mid f_i, \hat{W}) \ldots \hat{p}(y_i = N \mid f_i, \hat{W})]^T$ be the estimate of the probability vector for all possible true class labels for the ith partition. Let $c(y_i) = [c_{y_i,1} \ \ldots \ c_{y_i,N}]^T$ be the cost vector of assigning the class label y_i to the ith partition, where $c_{y_i,j}$ is the cost of assigning the class label j to the ith partition with the correct class label y_i , and ρ the cost of the unknown class label. The expected risk of selecting the class label $y_i \in \mathcal{L}'$ in the partition is

$$r(y_i \mid f_i, \hat{W}) \;=\; \begin{cases} c(y_i)^T \hat{p}(f_i, \hat{W}), & y_i \neq N+1; \\ \rho, & y_i = N+1. \end{cases} \tag{2}$$

We can now obtain the vector of estimated class labels $\hat{y}$ by minimizing (2) over all possible partition labelings $\mathcal{L}'^{|\mathcal{S}|}$,

$$\hat{y} \;=\; \arg\min_{y \in \mathcal{L}'^{|\mathcal{S}|}} \sum_{i \in \mathcal{S}} \log(r(y_i \mid f_i, \hat{W})). \tag{3}$$

Note that if $c_{y_i,j} = 1 - \delta_{y_i - j}$, with δ_n the Kronecker delta function, minimizing (3) yields

$$\hat{y_i} \;=\; \left\{ \begin{array}{cl} \arg\max\limits_{y_i \in \mathcal{L}} \hat{p}(f_i, \hat{W}), & \max\limits_{y_i \in \mathcal{L}} \hat{p}(f_i, \hat{W}) > 1 - \rho; \\ N+1, & \text{otherwise.} \end{array} \right.$$

In other words, if the the maximum element of the estimate of the probability vector is rather large, we are reasonably sure of our decision and assign the label as the index of the element; otherwise, we are uncertain and thus assign the unknown-class label.

4. CLASSIFICATION WITH REJECT OPTION

We are now ready to include rejection concepts in an approach where contextual information is used in classifying with rejection.

Problem Formulation The entire classification problem can now be posed as an energy minimization problem of two potentials over the set of partitions $\mathcal{S}$, represented by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ the set of vertices each corresponding to an individual partition, and $\mathcal{E}$ the set of edges corresponding to the connections between partitions. Rejection is included in the association potential V_A and the contextual information in the interaction potential V_I,

$$\hat{y} \;=\; \arg\min_{y \in \mathcal{L}'^{|\mathcal{S}|}} \left[\sum_{i \in \mathcal{S}} V_\text{A}(y_i) + \sum_{(i,j) \in \mathcal{E}} V_\text{I}(y_i, y_j) \right]. \tag{4}$$

The minimization of energy in the discriminative random fields is performed by graph cuts [8, 9, 10].

Association Potential We introduce reject option into the classification problem through association potential. Given the risk function $r(y_i \mid f_i, \hat{W})$ from (2), the association potential is

$$V_{\mathrm{A}}(y_i) \;=\; \log(r(y_i \mid f_i, \hat{W})).$$

Interaction Potential We introduce contextual information into the classification problem through interaction potential [11] of neighborhoods. A *local neighborhood* $\mathcal{E}_{i,\ell} \subset \mathcal{E}$ is the set of edges connecting partition i and its immediate neighbors. A *nonlocal neighborhood* $\mathcal{E}_{i,n} \subset \mathcal{E}$ is the set of edges connecting partition i and other partitions with high similarity. This similarity is determined by the similarity function $s : (i, j) \to [0,\, 1]$ that assigns a similarity value between two partitions i and j. The similarity function is based on features that better assess similarity, which are not being present in the classifer. For example, we can feed the image into a Gabor filter bank and then make a similarity decision based on the Gabor coefficients of the two partitions.

Let ϕ be a transition function between two partitions, and ψ_ℓ, ψ_n the weights associated with local and nonlocal neighborhoods, respectively. Then, the interaction potential can be written as a sum of individual (possibly overlapping) interaction potentials,

$$\sum_{(i,j)\in\mathcal{E}} V_{\mathrm{I}}(y_i, y_j) \;=\; \psi_\ell \sum_{(i,j)\in\mathcal{E}_\ell} \phi(y_i, y_j) + \psi_n \sum_{(i,j)\in\mathcal{E}_n} \phi(y_i, y_j).$$

The transition function ϕ enforces (1) piecewise smoothness (in a given concept of neighborhood), (2) ease of transition to the unknown class ($y_i = N + 1$), (3) and ease of transition between the uninteresting ($y_i = 1$) and unknown ($y_i = N + 1$) classes,

$$\phi(y_i, y_j) \;\equiv\; \begin{cases} 1 - \delta_{y_i - y_j}, & y_i, y_j \neq N + 1; \\ 0, & \text{otherwise.} \end{cases}$$

The local neighborhood enforces a local piecewise smoothness of the labeling, while the nonlocal neighborhood enforces a nonlocal piecewise smoothness of the labeling.

5. EXPERIMENTAL RESULTS

For each image in the data set, the method is applied multiple times with randomized training sets. We test the method using training sets based on both a single image as well as the entire data set.

Data Set The data set consists of 36 1600×1200 images of H&E-stained teratoma tissues imaged at 40X magnification containing 23 classes. We show results on three representative images in Figure 1, using the entire data set for the training samples in the multiple-image training set approach.

Methods We analyze the performance based on the following criteria: the comparison between the initial accuracy A_{ini} obtained by the labeling resulting from a *maximum a posteriori* $\hat{p}(f_i, \hat{W})$ and the final accuracy A_{fin} obtained by the labeling (4); the improvement ratio $A_{\mathrm{imp}} = A_{\mathrm{fin}}/A_{\mathrm{ini}}$; the rejection rate r; and rejection ratios for correctly and incorrectly classified samples $r_{\mathcal{C}}$ and $r_{\mathcal{I}}$, respectively, and the ratio between the two, $r_{\mathcal{I}}/r_{\mathcal{C}}$.

For each image X_i we define four training sets: (1) T_{S1} consists of 60 randomly selected partitions of X_i ($\sim$ 1.5% of X_i). (2) T_{S2} consists of 600 randomly selected partitions of X_i ($\sim$ 15% of X_i). Let us consider an auxiliary training set T_{aux} consisting of 30 randomly selected partitions all the images except X_i. (3) T_{M1} consists of the union of T_{S1} and T_{aux}. (4) T_{M2} consists of the union of T_{S2} and T_{aux}.

Results From Table 1, we see that there is a substantial accuracy improvement by including the reject option and contextual information in the classification. Table 2 shows the effect of different rejection ratios on the accuracy of the labeling for a single example image. There is a tradeoff between accuracy improvement, the performance of the rejection option (quantified by $r_{\mathcal{I}}/r_{\mathcal{C}}$) and the overall rejection rate. Note that A_{ini} in Table 2 is obtained for a single instantiation of T_{S1}, and is thus different from that in Table 1, which is obtained as the average of 30 instantiations of T_{S1}. Figure 1 shows results on the three example images.

Images	A_{ini}	A_{fin}	A_{imp}	r	$r_{\mathcal{I}}$	$r_{\mathcal{C}}$
			Training set T_{S1}			
(1)	0.7872	0.8396	1.0671	0.0905	0.2930	0.0388
(2)	0.6053	0.7272	1.2043	0.2156	0.4178	0.0746
(3)	0.6169	0.7316	1.1884	0.1967	0.4098	0.0618
			Training set T_{S2}			
(1)	0.9068	0.9431	1.0401	0.0555	0.3726	0.0247
(2)	0.7856	0.8885	1.1310	0.1295	0.4913	0.0303
(3)	0.8256	0.9137	1.1068	0.1086	0.4997	0.0279
			Training set T_{M1}			
(1)	0.6628	0.8240	1.2448	0.2060	0.5417	0.0407
(2)	0.4463	0.5888	1.3392	0.3145	0.4701	0.1054
(3)	0.3620	0.5041	1.4111	0.3561	0.4998	0.1123
			Training set T_{M2}			
(1)	0.7275	0.8613	1.1940	0.1678	0.5391	0.0341
(2)	0.7176	0.8337	1.1651	0.1640	0.4727	0.0364
(3)	0.7296	0.8812	1.2131	0.1853	0.5943	0.0317

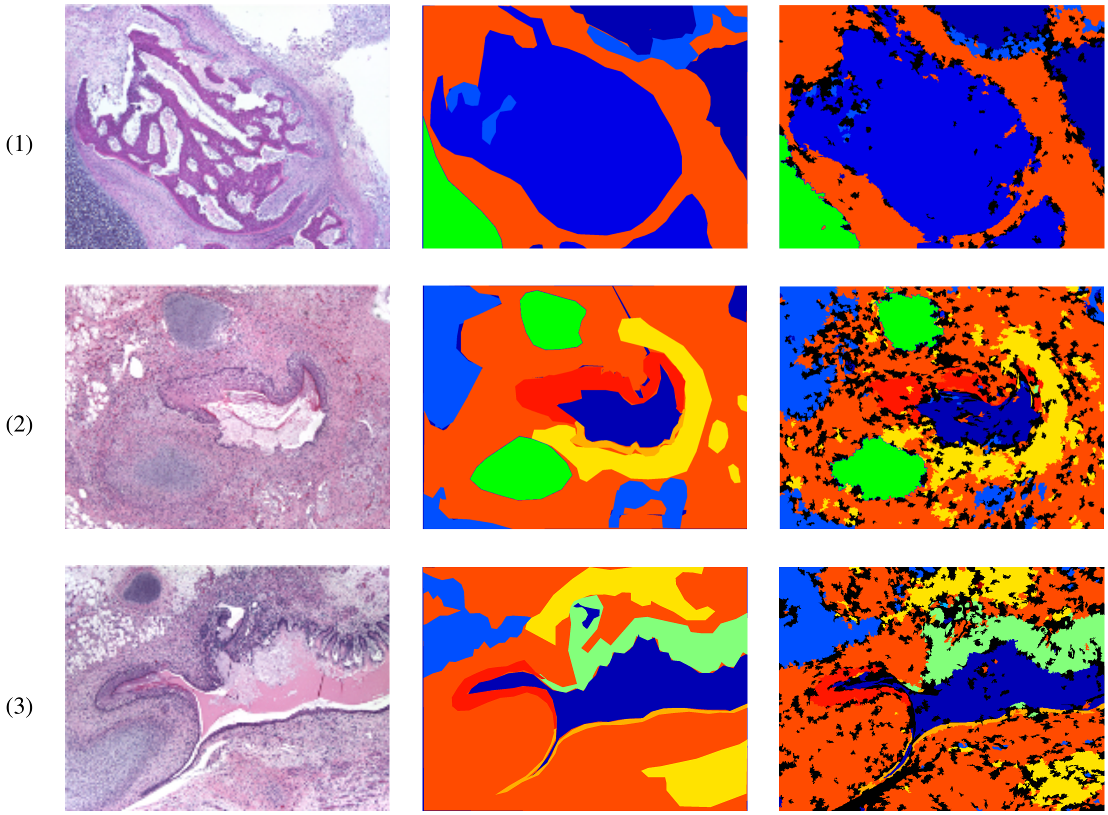
Table 1. Average values for the performance metrics for single-image and multiple-image training sets obtained after 30 randomized runs of the classification method corresponding to example images (1), (2) and (3) in Figure 1.

			A_{ini} = 0.8165			
ρ	10^5	0.9300	0.5300	0.1300	0.010	0.001
r	0.0000	0.0503	0.1009	0.1992	0.5715	0.6739
$r_{\mathcal{I}}$	0.0000	0.1677	0.3468	0.5403	0.9274	0.9677
$r_{\mathcal{C}}$	0.0000	0.0239	0.0457	0.1224	0.4915	0.6078
$r_{\mathcal{I}}/r_{\mathcal{C}}$		7.0167	7.5886	4.4142	1.8869	1.5921
A_{fin}	0.8390	0.8582	0.8687	0.8947	0.9689	0.9819

Table 2. Effect of changing the parameter ρ in (2) for Image (1) in the T_{S1} training set.

6. CONCLUSIONS AND FUTURE RESEARCH

The inclusion of a reject option using spatial contextual information greatly improves the accuracy of the classification. The relative high weight of the rejection rate for incorrect classifications compared to the rejection rate for correct classifications ($r_{\mathcal{I}}/r_{\mathcal{C}}$) points to a good behavior of the reject option. These encouraging results point towards potential utility in large-scale automated tissue identification of histological slices. Possible future research directions are: (1) Use of a partitioning algorithm that includes more specific features (with physiological relevance). (2) Modeling the transition function ϕ using expert knowledge of tissue transitions in normal



(a) Original image. (b) Ground truth. (c) Classification result.

Fig. 1. Example images of H&E stained samples of teratomas imaged at 40X magnification containing multiple tissues: (1) bone, cartilage and fat; (2) cartilage, fat, smooth muscle, epithelium, connective, and squamous tissue; and (3) fat, gastrointestinal, smooth muscle, epithelium, connective, and squamous tissue. Training set consists of T_{S2} for each example image. Dark blue denotes trained rejection (uninteresting class) and black denotes rejected partitions (unknown class).

and abnormal histological images. (3) Inclusion of multiscale partitioning, classification and rejection in the interaction and association potentials.

7. REFERENCES

[1] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inform. Th.*, vol. 16, no. 1, pp. 41–46, Jan. 1970.

[2] A. Chebira, J. A. Ozolek, C. A. Castro, W. G. Jenkinson, M. Gore, R. Bhagavatula, I. Khaimovich, S. E. Ormon, C. S. Navara, M. Sukhwani, K. E. Orwig, A. Ben-Yehudah, G. Schatten, G. K. Rohde, and J. Kovačević, "Multiresolution identification of germ layer components in teratomas derived from human and nonhuman primate embryonic stem cells," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Paris, France, May 2008, pp. 979–982.

[3] R. Bhagavatula, M. C. Fickus, J. W. Kelly, C. Guo, J. A. Ozolek, C. A. Castro, and J. Kovačević, "Automatic identification and delineation of germ layer components in H&E stained images of teratomas derived from human and nonhuman primate embryonic stem cells," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Rotterdam, The Netherlands, Apr. 2010, pp. 1041–1044.

[4] M. McCann, R. Bhagavatula, M. C. Fickus, J. A. Ozolek, and J. Kovačević, "Automated colitis detection from endoscopic biopsies as a tissue screening tool in diagnostic pathology," in *Proc. IEEE Int. Conf. Image Proc.*, Orlando, FL, Sept. 2012.

[5] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. Journ. Comp. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[6] D. Böhning, "Multinomial logistic regression algorithm," *Ann. of Inst. of Stat. Math.*, vol. 44, no. 1, pp. 197–200, Mar. 1992.

[7] J. Li, J. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new bayesian approach with active learning," *IEEE Trans. Geoscience & Remote Sensing*, 2011.

[8] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Trans. Patt. Anal. and Mach. Intelligence*, vol. 20, no. 12, pp. 1222–1230, 2001.

[9] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Patt. Anal. and Mach. Intelligence*, vol. 26, no. 9, pp. 1124–1137, Sept. 2004.

[10] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Patt. Anal. and Mach. Intelligence*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[11] S. Kumar and M. Hebert, "Discriminative random fields," *Int. Journ. Comp. Vis.*, vol. 68, no. 2, pp. 179–201, 2006.