# RECOGNIZING FOCAL LIVER LESIONS IN CONTRAST-ENHANCED ULTRASOUND WITH DISCRIMINATIVELY TRAINED SPATIO-TEMPORAL MODEL

*Xiaodan Liang*[⋆] *Qingxing Cao*[⋆] *Rui Huang*[*†] *Liang Lin*[⋆]

[⋆] Sun Yat-sen University [†] NEC Laboratories, China

**ABSTRACT**

The aim of this study is to provide an automatic computational framework to assist clinicians in diagnosing Focal Liver Lesions (FLLs) in Contrast-Enhancement Ultrasound (CEUS). We represent FLLs in a CEUS video clip as an ensemble of Region-of-Interests (ROIs), whose locations are modeled as latent variables in a discriminative model. Different types of FLLs are characterized by both spatial and temporal enhancement patterns of the ROIs. The model is learned by iteratively inferring the optimal ROI locations and optimizing the model parameters. To efficiently search the optimal spatial and temporal locations of the ROIs, we propose a data-driven inference algorithm by combining effective spatial and temporal pruning. The experiments show that our method achieves promising results on the largest dataset in the literature (to the best of our knowledge), which we have made publicly available.

***Index Terms*—** CEUS, FLLs, Spatio-Temporal Model,

## 1. INTRODUCTION

Liver cancer is the third cause of cancer-related death [1]. Visualization of Focal Liver Lesions (FLLs) has been attempted by employing various imaging techniques. Ultrasound is often performed in the diagnostics due to its low cost, efficiency and non-invasiveness. The use of Contrast-Enhanced Ultrasound (CEUS) can further assess the contrast enhancement (i.e., the intensity of the FLL area relative to that of the adjacent parenchyma) patterns of FLLs, which has markedly improved the accurate diagnosis of FLLs [1]. As shown in Fig. 1, temporal enhancement patterns typically characterize the benign or malignant FLLs (e.g., sustain enhancement in the last two vascular phases for benign and hypo-enhancement for malignant FLLs). On the other hand, spatial enhancement patterns during the arterial phase often characterize the specific types of FLLs.

Extensive research efforts have been made to assist the experts in diagnosing different types of cancers and, in particular, FLLs using ultrasound images [2]. However, the application of CEUS for differentiating FLLs is still a relatively new technique [3, 4, 5]. A cascade of Artificial Neural Networks [3] is employed to classify FLLs based on manually segmented lesion regions. Anaye et al. [4] analyzes the Dynamic Vascular Patterns (DVPs) of FLLs with respect to surrounding healthy parenchyma to differentiate between benign and malignant FLLs. Bakas et al. [5] track a manually initialized FLL and its surrounding parenchyma to characterize it as either benign or malignant based on its vascular signature.

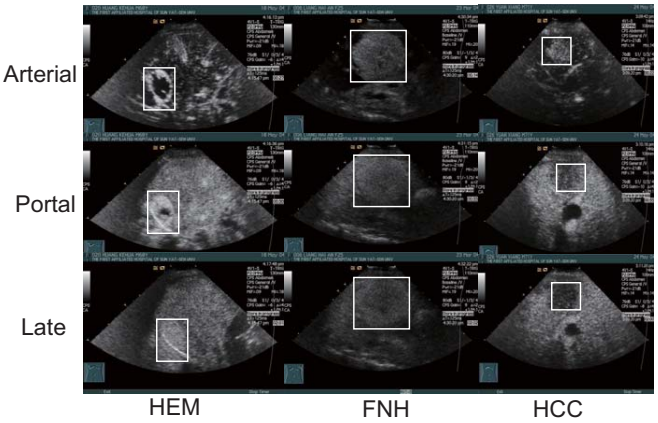


**Fig. 1**. The enhancement pattern ROIs of three different FLLs: Hemangioma (HEM), Focal Nodular Hyperplasia (FNH), Hepatocellular Carcinoma (HCC), in three different phases: arterial, portal venous and late phases. The HEM and FNH are benign FLLs and HCC is a malignant FLL.

In all these works, varying degrees of manual interactions are required to identify the Regions of Interest (ROIs) of FLLs or the normal parenchyma. The manual annotations are highly dependent on the skills and knowledge of the experts, leading to large variations in inter/intra-observer interpretations. Besides, the ever-increasing amount of CEUS data acquired and processed nowadays demands automatic computational systems that can save the radiologists' time and efforts. In addition, most of the previous works focused on differentiating between benign and malignant FLLs, or characterizing a specific type of FLLs. We, on the other hand, are trying to combine different enhancement patterns to recognize multiple different types of FLLs in a unified framework.

The main contributions of our work herein are threefold.

*Corresponding author. This work was supported by the Program of Guangzhou Zhujiang Star of Science and Technology (no. 2013J2200067), the Special Project on the Integration of Industry, Education and Research of Guangdong Province (no. 2012B091000101), the Guangdong Science and Technology Program (no. 2012B031500006).

First, we propose a fully automatic computational framework to recognize FLLs by modeling the locations of ROIs as latent variables in a discriminative model and combining both spatial and temporal enhancement patterns of the ROIs into the framework. Our model is then trained by a weakly supervised learning algorithm, which alternates between inferring the most probable spatial and temporal locations of the ROIs and optimizing the model parameters. Second, considering that most of the video frames and the regions in each frame contain redundant or irrelevant information for recognizing FLLs, the automatic detection of optimal locations of the ROIs is made very efficient by a novel data-driven inference method, which combines the spatial and temporal pruning techniques to disregard less discriminative frames and regions. The optimal ROI locations are then determined by dynamic programming. Last but not least, a new region representation for ROIs is presented to capture the important and relevant ultrasonic characteristics of FLLs, which is not necessarily limited to our framework.

We apply our method on a new dataset (namely SYSU-CEUS dataset) we collected and made public, which contains in total 353 CEUS video sequences of three types of FLLs (186 HCC, 109 HEM and 58 FNH), and is, to the best of our knowledge, the largest dataset in the literature. The experimental results demonstrate that our method achieves promising performance without manual interactions.

## 2. OUR MODEL

### 2.1. Region representation

The accurate classification of FLLs highly depends on the representation of the characteristics of the lesion regions (e.g., internal echo, morphology, edge, echogenicity and posterior echo enhancement). However, one single ROI $R$ is often insufficient to capture all the ultrasonic characteristics. For instance, the region inside the lesion, denoted as $R^-$, can capture the internal echo of the FLL; the lesion region $R$ can be used to observe the boundary and the morphology of the FLL; the tissue area surrounding the lesions, denoted as $R^+$, can be used to measure the posterior echo enhancement; and the echogenicity of the lesion can be measured by comparing the intensities of above regions. Therefore, given an ROI $R$, the regions $R^-$ and $R^+$ can be obtained by shrinking and enlarging $R$ by a small factor, respectively. We propose to describe the characteristics of the lesion region as following:

$$f(R) = [f^t(R^-), f^t(R), f^t(R^+), f^d(R^-, R), f^d(R, R^+)] \quad (1)$$

where $f^t$ extracts the appearance features of each region, such as the widely-used Grey Level Co-occurrence Matrix (GLCM) and Local Phase (LP); $f^d$ calculates the mean intensity difference of two regions. Consequently, the concatenation of all these features, $f(R)$, captures all kinds of ultrasonic characteristics of region $R$.

### 2.2. Model representation

Given a CEUS video sequence $\mathbf{x}$, $y$ is the corresponding class label of the FLL in this video, ranging over a finite set $\mathcal{Y}$ (e.g., $\mathcal{Y}$={HCC, HEM, FNH}). We assume that the FLL can be compactly represented by a set of ROIs (intuitively, the most discriminative regions for distinguishing different FLLs), $\{R_1, R_2, \ldots, R_m\}$, in three vascular phases, the arterial, portal venous, and late phase. Each ROI $R_i$ is a region extracted from video frame $t_i$, at spatial location $p_i = (x_i, y_i, s_i)$, where $x_i, y_i, s_i$ are the coordinates and scale of $R_i$. The latent variables $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$, where $h_i = (p_i, t_i)$, take values from a finite set $\mathcal{H}_i$ of all possible ROI locations. Given a video $\mathbf{x}$, its corresponding class label $y$, and latent variables $\mathbf{h}$, the conditional probability of the recognition problem is defined as,

$$\begin{aligned} p(y|\mathbf{x};\omega) &= \sum_{\mathbf{h}\in\mathcal{H}} p(y,\mathbf{h}|\mathbf{x};\omega) \\ &= \frac{\sum_{\mathbf{h}\in\mathcal{H}} \exp(\omega^{\mathrm{T}}\cdot\psi(\mathbf{x},\mathbf{h},y))}{\sum_{\hat{y}\in\mathcal{Y}}\sum_{\mathbf{h}\in\mathcal{H}} \exp(\omega^{\mathrm{T}}\cdot\psi(\mathbf{x},\mathbf{h},\hat{y}))} \end{aligned} \quad (2)$$

where $\omega$ is the model parameter vector, $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \times \mathcal{H}_m$, and $\psi(\mathbf{x}, \mathbf{h}, y)$ is a feature function depending on video sequence $\mathbf{x}$, class label $y$, and latent variables $\mathbf{h}$. We then define $\omega^{\mathrm{T}} \cdot \psi(\mathbf{x}, \mathbf{h}, y)$ as the combination of two potentials,

$$\begin{aligned} \omega^{\mathrm{T}}\cdot\psi(\mathbf{x},\mathbf{h},y) = &\sum_{i\in m} \alpha_i^{\mathrm{T}}\cdot\phi^u(\mathbf{x},y,h_i) \\ &+ \sum_{(i,j)\in\mathcal{E}} \beta_{i,j}^{\mathrm{T}}\cdot\phi^p(\mathbf{x},y,h_i,h_j) \end{aligned} \quad (3)$$

where $\phi^u(\cdot)$ is the unary potential function of variable $h_i$ and $\phi^p(\cdot)$ the pairwise potential function of $(h_i, h_j)$. $\mathcal{E}$ is the set of neighboring latent variables (defined for the pairs of temporally adjacent ROIs).

1) Unary potential $\alpha_i^{\mathrm{T}} \cdot \phi^u(\mathbf{x}, y, h_i)$: This singleton potential function $\phi^u(\cdot)$ models the compatibility between class label $y$ and appearance of region $R_i$ (note that $R_i = \mathbf{x}(h_i)$).

$$\alpha_i^{\mathrm{T}}\cdot\phi^u(\mathbf{x},y,h_i) = \sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{H}_i} \alpha_i^a\cdot\delta_y(a)\cdot\delta_{h_i}(b)\cdot f(\mathbf{x}(h_i)) \quad (4)$$

where $f(\mathbf{x}(h_i))$ is the feature vector describing the appearance of the region, as defined in section 2.1. The indicator function $\delta_y(a)$ is equal to 1 if $y = a$, 0 otherwise. Similarly, $\delta_{h_i}(b)$ is equal to 1 if $h_i = b$, 0 otherwise. The parameter $\alpha_i$ is simply the concatenation of all $\alpha_i^a$.

2) Pairwise potential $\beta_{i,j}^{\mathrm{T}} \cdot \phi^p(\mathbf{x}, y, h_i, h_j)$: The potential function $\phi^p(\cdot)$ models the compatibility between class label $y$ and the temporal transition of a pair of neighboring latent variables $(h_i, h_j)$.

$$\begin{aligned} \beta_{i,j}^{\mathrm{T}}\cdot\phi^p(\mathbf{x},y,h_i,h_j) = &\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{H}_i}\sum_{c\in\mathcal{H}_j} \\ &\beta_{i,j}^a\cdot\delta_y(a)\cdot\delta_{h_i}(b)\cdot\delta_{h_j}(c)\cdot f^p(\mathbf{x},h_i,h_j) \end{aligned} \quad (5)$$

where $f^p(\cdot)$ includes two components: the appearance variance feature, measuring the difference between $f(\mathbf{x}(h_i))$ and $f(\mathbf{x}(h_j))$, and the spatial displacement feature, measuring the Euclidean distance between the spatial coordinates of $h_i$ and $h_j$. The parameter $\beta_{i,j}$ is simply the concatenation of all $\beta_{i,j}^a$.

### 2.3. Learning

Given a training set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, the model parameter $\omega$ can be learned by maximizing the conditional log-likelihood on the training samples:

$$\begin{aligned}\omega^* &= \arg\max_{\omega} \mathcal{L}(\omega) = \arg\max_{\omega} \sum_{i=1}^{N} \mathcal{L}^i(\omega) \\ &= \arg\max_{\omega} \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i;\omega) \\ &= \arg\max_{\omega} \sum_{i=1}^{N} \log(\sum_{\mathbf{h}\in\mathcal{H}} p(y_i, \mathbf{h}|\mathbf{x}_i;\omega))\end{aligned} \tag{6}$$

where $\mathcal{L}^i(\omega)$ denotes conditional log-likelihood of the $i^{th}$ training sample (defined in Eq(2)) and $\mathcal{L}(\omega)$ conditional log-likelihood of the whole training set. The objective function is not concave due to the latent variables $\mathbf{h}$. We adopt the latent structural SVM framework [6, 7], which alternates between inferring latent variables $\mathbf{h}$ and optimizing model parameter $\omega$. The problem of inferring $\mathbf{h}$ can be solved efficiently using a data-driven inference algorithm (Sec. 2.4), and the parameter optimization is a standard structural SVM training problem, solved by the cutting-plane algorithm. We use the one-vs-one binary classification strategy for the multi-class problem, and repeat the above two steps until convergence.

Given a learned model, the classification is achieved by picking the FLL class with the highest SVM score given the optimal locations of the ROIs, which are also inferred during the classification:

$$y^* = \arg\max_{y\in\mathcal{Y}} \max_{\mathbf{h}\in\mathcal{H}} \omega^{\mathrm{T}} \cdot \psi(\mathbf{x}, y, \mathbf{h}) \tag{7}$$

### 2.4. Data-driven inference

The inference task is to find the optimal locations of the ROIs (i.e., the latent variables $\mathbf{h}$). However, the searching space will be very large if we consider all regions in all frames. Thus, we propose a data-driven inference algorithm, which combines effective spatial and temporal pruning techniques to disregard less discriminative frames and regions. The optimal locations of the most discriminative ROIs can then be determined using dynamic programming in a limited space.

1) Temporal pruning: In a CEUS video, the appearance of ultrasound frames often varies slowly and smoothly according to the hemodynamic, and the most discriminative frames are usually those with the largest contrast changes compared with neighboring frames. Thus, a small set of candidate frames, which have local maximum of the contrast change, are automatically selected. In particular, for each frame $I_t, (t = 1, \cdots, T)$ in a video $\mathbf{x}$, we compute the contrast feature $v_t$ from the co-occurrence distribution $C_t$ defined over $I_t$ [8]. Let $\nabla\mathbf{v}$ be the gradient of the contrast vector $\mathbf{v} = [v_1, v_2, \ldots, v_T]$, the candidate frame set $B$ is formed by finding the frames at the local maximum of $\nabla\mathbf{v}$.

2) Spatial pruning: After temporal pruning, we also prune the less important regions by considering two priors: saliency prior and location prior. First, we believe that *salient* regions (e.g., having higher contrast or containing typical structures) have more discriminative information, and thus are more likely to be candidates of ROIs. Second, we observe that FLLs often appear in or close to the center of the images, probably because a skilled ultrasound operator usually places the liver area in the middle of the display. According to these two observations, we evaluate all the regions with different scales in each candidate frame $I \in B$ (sliding window protocol), and only select the regions with prior probability larger than a threshold $\tau$ as ROI candidates. The prior probability of a region $r$ being an ROI is,

$$p(r) = \mathcal{S}(r)\mathcal{N}(C^r|C^I, \sigma) \tag{8}$$

where $\mathcal{S}(r)$ is the normalized mean saliency of the region $r$ in the saliency map $\mathcal{S}$ computed, e.g., by the quaternion-based spectral saliency method [9], on image $I$. $C^r$ and $C^I$ are the centers of region $r$ and image $I$, respectively. $\mathcal{N}(C^r|C^I, \sigma)$ is a Gaussian distribution.

It is worth noting that the spatial pruning in the last two vascular phases (portal and late) can be more aggressive. This is because the contrast between FLLs and normal tissues is often very low, and the locations of FLLs do not change much since the arterial phase. Thus, in the last two phases, we only search the regions in a spatial neighborhood around the locations of ROI candidates found in the arterial phase. Finally, given the model parameters and the observations, the latent variables $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$ form a hidden Markov model, and can be solved by the Viterbi algorithm [10].

## 3. RESULTS

We test our method on the SYSU-CEUS dataset collected from the First Affiliated Hospital, Sun Yat-sen University [1]. The equipment used was Aplio SSA-770A (Toshiba Medical System). The dataset consists of three types of FLLs: 186 HCC, 109 HEM and 58 FNH instances (i.e., 186 malignant and 167 benign instances). All these instances with resolution $768\times576$ were taken from different patients, with large variations in appearance and enhancement patterns (e.g., size, contrast, shape and location) of FLLs. We adopt the 5-fold cross validation training strategy and the sensitivity for each class and mean accuracy as the evaluation criteria, similar to [4]. In our implementation, we extract four statistics (i.e., Contrast, Correlation, Energy, Homogeneity) of GLCM [8] with four orientations ($\theta = 0^o, 45^o, 90^o, 135^o$), and one distance "1", to represent the texture feature $f^t$ (Sec 2.1). Three scales of regions (i.e., $64 \times 64$, $128 \times 128$, $200 \times 200$) and step length 20 are used for sliding windows, and $\tau = 0.6$ and $\sigma = 0.5$ are used for spatial pruning. The experiments are carried out on a PC with Core I7 3.4GHz CPU, and the average processing time for a 4-min CEUS video is about 100 seconds (dozens of minutes without spatial and temporal pruning).

We first report the sensitivities and mean accuracies of our method in differentiating benign and malignant FLLs in Table. 1. The average accuracy (89.7%) is comparable, if not superior, to the results reported in previous studies on smaller

[1] https://github.com/lemondan/Focal-liver-lesions-dataset-in-CEUS

| | $\text{Sens}^{Benign}$ | $\text{Sens}^{Malignant}$ | Accuracy |
|---|---|---|---|
| Ours | **85.7%** | **93.4%** | **89.7%** |

**Table 1**. Sensitivities and mean accuracies on characterizing benign and malignant FLLs. Sens means the sensitivity of the specific class.

| | $\text{Sens}^{HCC}$ | $\text{Sens}^{HEM}$ | $\text{Sens}^{FNH}$ | Accuracy |
|---|---|---|---|---|
| DDI | **88.9%** | 81.0% | 63.6% | **82.4%** |
| manual | 86.1% | **85.7%** | **72.7%** | **83.8%** |
| bruteforce | 83.3% | 80.1% | 36.4% | 75.0% |
| baseline | 78.9% | 22.0% | 10.3% | 49.9% |

**Table 2**. Sensitivities and mean accuracies in the different experiment settings.

datasets [4, 5]. The second experiment in Table. 2 shows the effectiveness of our data-driven inference algorithm by altering the procedure to determine the ROIs. Our data-driven inference algorithm ("DDI") is compared with 1) "manual": the ROI of each instance in the arterial phase is manually selected and the inference only performed in the portal and late phase; 2) "bruteforce": the liver region is labeled and the optimal ROIs are searched in the entire region of liver, without pruning; 3) "baseline": the ROIs are randomly selected in the images of three phases. The results demonstrate that our fully automatic inference algorithm achieves comparable performance to the "manual" method, and performs better than "bruteforce" and "baseline". Note that the performance of our algorithm on FNH is worse because the amount of training data of FNH is relatively small.

Finally, in Table.3 we compare the region representation of our framework with other state-of-the-art methods: Multiple-ROI [11], $\text{ROI}^{posterior}$ [12] and $\text{ROI}^{out}$ [13]. Each region representation is tested with three most popular low-level features used for the ultrasound image: GLCM, Law's texture, and Local Phase, similar to [11]. Note that we ignore the shape features because FLLs often show a wide variety of shapes or have unclear boundaries due to the low contrast. We manually select ROIs in three phases as required in previous works (note here we do not consider the performance of the inference algorithm), and use linear SVM as the classifier. The results show that our region representation obtains superior performances in general.

## 4. CONCLUSIONS

In this work we propose a fully automatic computational framework for characterizing different types of FLLs in CEUS, which effectively combines the diverse information of spatial and temporal enhancement patterns. Besides, a weakly supervised learning algorithm is utilized, which alternates between inferring the latent variables (i.e. the locations of ROIs) and optimizing the model parameters. A data-driven inference algorithm is then proposed to efficiently determine the optimal locations of ROIs. The method is shown to achieve promising results and have the potential of being developed for real-time clinical applications. In the future, an interactive system will be developed to allow the radiologists to revise the intermediate output, e.g., the locations of ROIs, to improve the accuracies.

| | $\text{Sens}^{HCC}$ | $\text{Sens}^{HEM}$ | $\text{Sens}^{FNH}$ | ACC1 |
|---|---|---|---|---|
| $[11]^{GLCM}$ | 85.9 % | 75.9% | 36.2% | 74.7% |
| $[12]^{GLCM}$ | **88.1%** | 67.5% | 51.7% | 75.8% |
| $[13]^{GLCM}$ | 82.1% | 61.1% | 34.4% | 67.8% |
| $\text{Ours}^{GLCM}$ | 87.2% | **83.5%** | **67.2%** | **82.7%** |
| $[11]^{Law's}$ | 82.7 % | 75.9% | **72.4%** | 78.9% |
| $[12]^{Law's}$ | 75.6% | 77.7% | 62.0% | 74.2% |
| $[13]^{Law's}$ | 69.7% | 72.2% | 56.9% | 68.4% |
| $\text{Ours}^{Law's}$ | **84.3%** | **87.2%** | 67.6% | **82.4%** |
| $[11]^{LP}$ | 85.9 % | 67.5% | 63.7% | 76.5% |
| $[12]^{LP}$ | 80.0% | 69.4% | 55.1% | 72.7% |
| $[13]^{LP}$ | 78.9% | 50.9% | 48.2% | 65.2% |
| $\text{Ours}^{LP}$ | **86.1%** | **73.4%** | **63.8%** | **78.3%** |

**Table 3**. Comparisons of region representation methods by applying different feature descriptors.

## 5. REFERENCES


[1] M. Claudon et al., "Guidelines and good clinical practice recommendations for contrast enhanced ultrasound in the liver - update 2012," *Ultrasound in Med and Bio*, vol. 39, no. 2, 2013.

[2] J.cA. Noble, "Ultrasound image segmentation and tissue characterization," *Proc Inst Mech Eng H.*, vol. 224, no. 2, 2010.

[3] J. Shiraishi et al., "Computer-aided diagnosis for the classification of focal liver lesions by use of contrast-enhanced ultrasonography," *Med Phys*, vol. 35, no. 5, 2008.

[4] A. Anaye et al., "Differentiation of focal liver lesions: usefulness of parametric imaging with contrast-enhanced US," *Radiology*, vol. 26, no. 1, 2011.

[5] S. Bakas et al., "Histogram-based motion segmentation and characterisation of focal liver lesions in CEUS," *Annals of the BMVA*, vol. 7, 2012.

[6] X. Wang et al., "Dynamical and-or graph learning for object shape modeling and detection," in *NIPS*, 2012.

[7] X. Liang et al., "Learning latent spatio-temporal compositional model for human action recognition," in *ACM MM 2013*.

[8] R. Haralick et al., "Textural features for image classification," *Systems, Man and Cybernetics*, vol. SMC-3, no. 6, 1973.

[9] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *ECCV*, 2012.

[10] D. Koller and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*, MIT Press, 2009.

[11] JH Jeon et al., "Multiple ROI selection based focal liver lesion classification in ultrasound images," *Expert Systems with Applications*, vol. 40, no. 2, 2013.

[12] SH Kim et al., "Computer-aided image analysis of focal hepatic lesions in ultrasonography: preliminary results," *Abdom Imaging*, vol. 34, no. 2, 2009.

[13] G. Xian, "An identification method of malignant and benign liver tumors from ultrasonography based on GLCM texture features and fuzzy SVM," *Expert Systems with Applications*, vol. 37, no. 10, 2010.