Context-based region labeling for event detection in surveillance video

S. Javanbakhti, S. Zinger, P. H. N. de With
Video Coding and Architectures Research Group (SPS-VCA),
Electrical Engineering Faculty, Eindhoven University of Technology, the Netherlands
{S.Javanbakhti, S.Zinger, P.H.N.de.With}@tue.nl

***Abstract*—Automatic natural scene understanding and annotating regions with semantically meaningful labels, such as road or sky, are key aspects of image and video analysis. The annotation of regions is a considered helpful for improving the object-of-interest detection because the object position in the scene is also exploited. For a reliable model of a scene and associated context information, the labeling task involves image analysis at multiple, both global and local, scene levels. In this paper, we develop a general framework for performing automatic semantic labeling of video scenes by combining the local features and spatial contextual cues. While maintaining a high accuracy, we pursue an algorithm with low computational complexity, so that it is suitable for real-time implementation in embedded video surveillance. We apply our approach to a complex surveillance use case and to three different datasets: WaterVisie [1], LabelMe [2] and our own dataset. We show that our method quantitatively and qualitatively outperforms two sate-of-the-art approaches [3][4].**

I. Introduction

Video understanding has a demanding but essential application in the video surveillance domain. Automatic natural scene understanding and labeling regions with semantically meaningful labels (e.g., road, sky, etc.), have increasingly attracted attention, since they are key aspects in image and video understanding. One dominant application direction for scene understanding and region labeling is classifying pictures in a large database. Alternatively, in our case in surveillance, the region labeling and video understanding can improve the analysis of events or contribute to more reliable object detection.

For a reliable model of a scene and associated context information, the labeling task involves image feature analysis at global and local scene levels. Although local features such as color and texture per pixel or region are instrumental for understanding, they are typically not uniquely determining the semantic meaning of such a region (e.g. sky and water). In general, local features can be influenced by the presence of other objects as well as by the overall context of the scene. In most real-world object recognition tasks, the context provides a rich source of information that can help to improve the involved object recognition performance and reduce ambiguities of very local scene information [5].

There is no common understanding on the correct classification of different types of context into meaningful groups and categories. In a 3D spatio-temporal space, context refers to any *spatio-temporal* information [6]. In [7], three main types of contextual information that can be exploited in computer vision solutions are proposed. First, *probability (semantic)* information refers to the likelihood of an object being found in some scenes but not in others. Second, *size (scale)* exploits the fact that objects have a predetermined size in relation with other objects in the scene. Third, the *position (spatial)* corresponds to the likelihood of finding an object at specific positions in the scene, with respect to other objects. In this paper, to reduce ambiguities of local image information, we propose an automatic region labeling system based on the observation that each region is more likely to be found at a specific vertical position. Besides feature extraction, we intend to exploit the use of scene information in a *learning* method to relate image features and labels to each other. This may also lead to an automatic region labeling approach. The Support Vector Machine (SVM) classification provides a better performance compared with the other approaches [8]. Therefore, we adopt this technique for further use in our system. In conclusion, we want to design a system that exploits both feature extraction based on the above properties, while also employing SVM-based learning methods to constitute a powerful general framework for region labeling.

The starting point of our system design is a previously developed statistical model based on SVM [3], which is extended here by adding contextual information. We assign to each region one of the 5 labels: sky, vegetation, construction, road, water plus the class “unknown”. We apply our approach to a complex surveillance use case and to three different datasets: WaterVisie, LabelMe [2] and our own dataset. We also benchmark our algorithm against two other approaches [3][4], which aim at finding similar types of regions.

The paper is organized as follows. Section II describes our approach for region labeling. Section III presents our results and their application to a surveillance use case. Conclusions and discussion are provided in Section IV.

II. Region labeling approaches

Our algorithm contains three stages, as depicted in Figures 1 and 2, which will be discussed now.
Stage 1: Uniform regions. The image is divided into several regions with uniform color using graph-based segmentation.

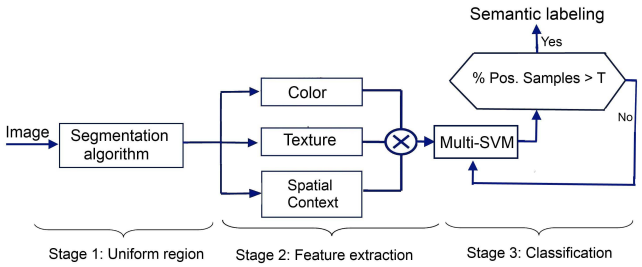


Figure 1: Our gravity-based region labeling approach.

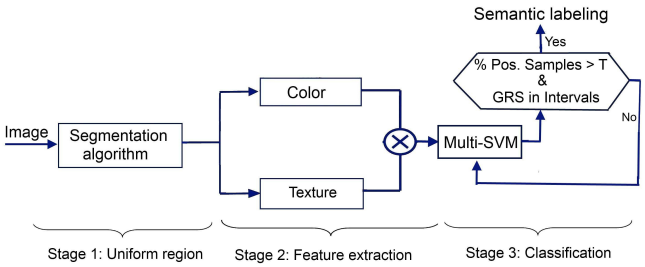


Figure 2: Our GRS-based region labeling approach.

Stage 2: Feature extraction. The region-based feature (vertical position) and pixel-based features (HSV color space and a group of Gabor features) of each segmented region are extracted. Regarding global feature extraction, two methods are proposed: (1) Spatial Context in which the normalized vertical position for each pixel is calculated; (2) Global Region Statistics (GRS) in which intervals for mean and standard deviation of vertical positions for each specific region are obtained.
Stage 3: Classification. Our algorithm employs two concepts in a sequential order.

- *Multiple-SVM (one vs. all)*. For each region class, an off-line separately trained SVM is used to classify that region. Given the feature analysis of the previous stage, color, texture and spatial context (normalized vertical position) are used for learning each region class separately. We call the extensive use of vertical information a *gravity-based model*. However, this is not sufficient for region classification.
- *Assigning labels*. For each class, we measure the percentage of pixels classified as belonging to this class in a given region. We assign a specific label to a region when the percentage of positively classified pixels in this region is above a threshold.

Figures 1 and 2 depict the two instantiations of our region labeling approach. The first approach adds Spatial Context in the form of the gravity-based model to the feature extraction stage. The second approach operates without gravity-based model in the feature extraction, but uses Global Region Statistics (GRS) in the classification stage. This alternative is motivated by the property that it avoids training of the Spatial Context information beyond color and texture, which are standard features for SVM. Note that in our discussion, we addressed GRS under feature extraction, but we use it here in *Stage 3* for classification. This is a way to include position information at a region level without using the local gravity-based model. Section II-C2 describes the GRS-based region labeling approach of Figure 2 in more detail.

A. Stage 1: Uniform regions

We employ an efficient graph-based segmentation as pre-processing in our region labeling to achieve two objectives: (a) distinguish each region from other objects while preserving the overall characterization of the region itself, (b) perform fast segmentation to support real-time application in surveillance systems. Details of our graph-based method can be found in [1].

B. Stage 2: Feature extraction

To train a reliable and robust SVM classifier, it can be sufficient to use only local features such as color and texture. However, when classes have similar characteristics, complications arise, which can be solved by adding spatial context. It involves the vertical position of the regions in the image, e.g. the sky tends to be at the top of the image and the water at the bottom. Summarizing, we combine the locally calculated pixel-based features and the region-based features to achieve a more reliable region labeling approach.

1) Pixel-based features: Color and *Texture*. We expect that significant region information is carried by pixel color. Here, we use the HSV color space [9]. The texture feature leads to a better classification by analyzing the local neighborhood variation [10]. Gabor features in addition to accurate time-frequency location provide robustness against varying brightness and contrast of images [11]. Kamarainen *et al.* [12] describe the basics of the 2D Gabor filter.

2) Spatial Context (SC): Spatial Context (SC). This information become specific for the region when a vertical position is used. This builds a gravity-based model and helps to overcome the ambiguities of using only color and texture [13]. For each pixel (i, j), we calculate its normalized vertical position $SC_{ij} = i/n$, where i is the row number, j the column number and n is the row count of the region.

C. Stage 3: Classification approaches

After segmenting the image and extracting the features, we proceed to obtain the labeling results. The labeling is performed by a classification system based on an off-line trained SVM. Here, we present two approaches for region classification, as depicted in Figures 1 and 2.

1) Fast classification using the gravity-based model: In this approach, color, texture and Spatial Context are used to train the SVM for each region class individually, to achieve unitary-category classification (i.e. an individual SVM is trained for each region type). Later, we randomly sample 100 pixels from a segmented region. The previously trained SVM for the considered class assigns labels to each pixel as positive or negative, depending on the classification

results. We calculate the percentage of positive samples in that region. Then, we label the region as belonging to the considered class (e.g. we find the segment depicted by sky), if this percentage of positive samples is higher than an empirically defined threshold. Our fast unitary-category classification is described in the inner part of Algorithm 1.

For multi-category labeling, we assign to each segment one of the 5 labels: sky, vegetation, construction, road, water plus the class "unknown". To this end, we classify each segment by 5 SVMs using our unitary classification (Algorithm 1) and obtain 5 numbers, indicating the percentages of positive pixels for each SVM. Finally, a segmented region is assigned to a particular class if its percentage is higher than the empirical threshold, which we from now on call T_e. Algorithm 1 illustrates our multi-category classification algorithm with the unitary algorithm embedded into it. The empirical threshold T_e for each region is set to 0.5.

```
for 5 classes do
    Define the next class type;
    for a segmented region do
        Randomly choose 100 samples in this region and use
        the SVM classifier to label the samples;
        Calculate the percentage of positive samples in this
        region;
        if the percentage of positive samples is higher that T_e
        then
            Set this region is positive;
        end
        Compare results to threshold T_e;
        Label the current region;
    end
end
```

Algorithm 1: Our fast multi-category classification

2) Classification using the GRS-based model: We define the *GRS* as the standard deviation and mean of the region position. Let us assume that we have M regions of a particular type, for example sky, in the training set of images. For each region, we calculate mean values μ_k $(k = 1, ..., M)$ of the vertical positions of its pixels. We also calculate the standard deviation σ_k of the vertical pixel positions for each region. Then we take minimum and maximum values for all means and standard deviations for this region type: $\mu_{min} = min(\mu_1, .., \mu_M)$, $\mu_{max} = max(\mu_1, .., \mu_M)$, $\sigma_{min} = min(\sigma_1, .., \sigma_M)$ and $\sigma_{max} = max(\sigma_1, .., \sigma_M)$. In this way, we obtain intervals for mean and standard deviation for the region position. We assume that the mean value of vertical pixel positions lies in the interval (μ_{min}, μ_{max}) and standard deviation within $(\sigma_{min}, \sigma_{max})$. We find these intervals for each of the 5 region types described in this paper. For a correctly labeled region, the region borders are in the typical interval values for mean and standard deviation of the vertical positions. Therefore, for assigning a label to a region, we check that both following conditions are satisfied: (1) the percentage of positively classified pixels exceeds the threshold T_e; (2) the mean and standard deviation of the vertical positions of the pixels lie in the intervals as discussed above.

III. Region labeling results

A. Initial test with still images

We have constructed a broad dataset which consists of images from multiple Internet datasets and a personal archive. It contains 5 classes (sky, vegetation, road, water, construction) plus the class "unknown". The dataset contains 255 images: 121 images for training, 134 for testing. The parameters for graph-based segmentation are the same as in [1]. The means and standard deviations in the GRS-based model are calculated based on 51 images from the training set. For benchmarking, we have tested our algorithm on the LabelMe [2] and WaterVisie [1] datasets. We have randomly selected 142 images from LabelMe and divided them into 102 training and 40 test images. We have also trained our gravity-based model on 111 frames and tested it on 16 videos of WaterVisie dataset. Figure 3 shows the original images of the datasets and the corresponding results of the gravity-based model. Figure 4 illustrates a challenging image along with the comparison results of three different reference approaches to highlight the differences between the labeling algorithms. This image contains several regions of interest and the color information is quite poor with only small color differences between neighboring regions. It can be observed that our gravity-based model achieves results that better correspond to the ground truth.

To evaluate the performance of the region labeling algorithm, we use the Coverability Rate (CR), which measures how much of the true region is detected by the algorithm [15]. In order to benchmark our approach, we compare our gravity-based model with two state-of-the-art approaches: Bao *et al.* [3] and Millet *et al.* [4]. We have extended the unitary-category classification of Bao *et. al.* [3] into multi-category classification and applied contextual information as an additional feature. The rule-based approach proposed by Millet *et al.* [4] relies on preknowledge on the relative spatial positions between regions. Table I shows the results of applying the gravity-based model and GRS-based model approaches, compared to Bao's and Millet's algorithms on 112 images of our dataset. We can observe that the gravity-model approach results in a higher CR. Our gravity-based approach outperforms the recently published algorithm of Bao *et. al.* with approximately 2%. Our approach also surpasses Millet *et. al.* [4], while preventing any preset rules which reduce flexibility of the method. Unlike Millet, our approach does not need to be rebuilt if a new region is added. The average of CR over six regions for the gravity-based labeling approach is 93% for our dataset, 94% for LabelMe and 96% for WaterVisie. This shows a significant improvement on LabelMe dataset compared to the 59% reported by Jain *et al.* [16]. We note though that Jain *et al.* [16] aimed at a clearly higher number of semantic region types, which is more difficult.

B. Video surveillance use case

In this study, we combine an existing group behavior analysis application [9] with our context detection of the scene. The group detection approach [9] locates a people

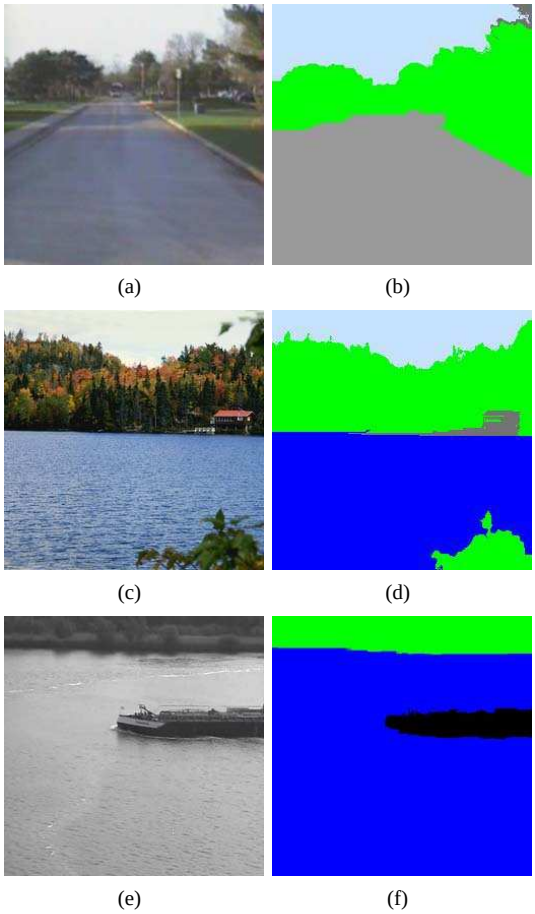


Figure 3: (a) Image from our dataset, (b) The gravity-based region labeling of (a), (c) Image from LabelMe [2], (d) The gravity-based region labeling of (c), (e) Image from WaterVisie [1], (f) The gravity-based region labeling of (e).

Region	Gravity-based model	GRS-based model	Bao *et. al.* [3]	Millet [4]
Sky	96	96	95	94
Construction	88	88	87	84
Water	93	91	89	89
Road	92	92	92	89
Vegetation	90	85	85	87
Unknown	95	96	97	99
Average	93	91	91	90

Table I: Coverability Rate (%) comparison for several semantic labeling algorithms.

group in a video based on human motion, computed by an optical flow approach. Road region is labeled by our proposed gravity-based approach. For finding a car, we apply a simple motion detection consisting of thresholding the difference between consecutive frames. If the detected motion does not belong to the group of moving people, then it is related to a car movement. Afterwards, if a car moves on the road where a group is detected, then it is considered as a dangerous event. Without using context, the group behavior analysis indicates an abnormal situation when the car is so

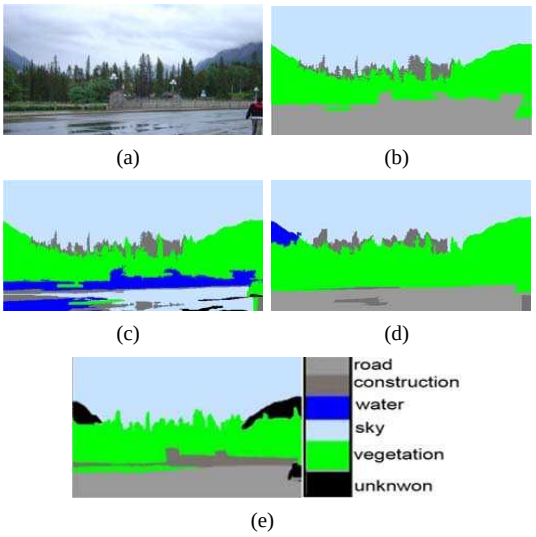
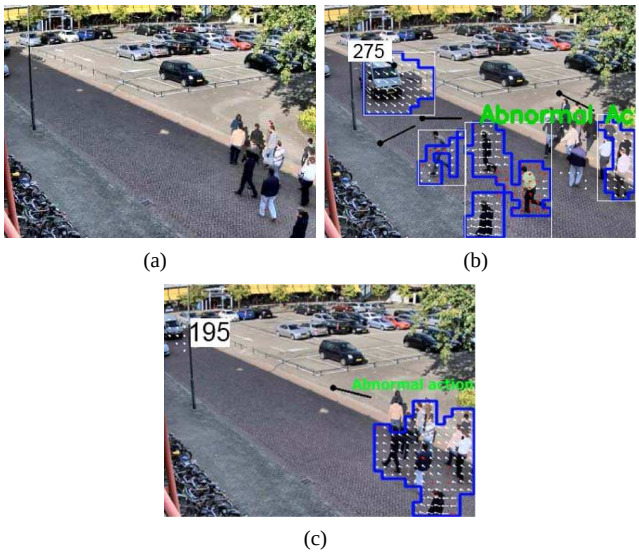


Figure 4: (a) Image from our dataset, (b) The gravity-based region labeling, (c) Region labeling from Bao *et. al.* [3], (d) Region labeling from Millet *et. al.* [4], (e) Ground truth of (a).

close to the group that it causes the group members to panic and disperse into different directions. Such late detections are not sufficiently preventive in surveillance applications. This has motivated us to apply contextual information in order to detect the danger much earlier. Experimental results show that using the contextual information together with group behavior detection leads to preventive detection of abnormal events as described above. For context-based event detection, we have used surveillance videos of our campus where several abnormal situations are simulated by a group of volunteers. This involves e.g. running of several people from the middle of the scene, due to an approaching car to simulate accident situations on a road. The spatial resolution of an original video frame is 800×600 pixels with 25-Hz frame rate, which is typical for broadcast-TV surveillance. Figure 5 illustrates an example of an abnormal situation within our surveillance video with the results from our group, car and gravity-based detection. Figures 5 (b) and 5 (c) illustrate the group behavior detections without and with the use of context, respectively, in particular the frame where the abnormal event is detected for the first time. It can be observed that our approach produces an alarm clearly earlier than the other approach from Figure 5 (b).

IV. Conclusion and Discussion

In this paper, we have presented our ongoing research on context analysis for outdoor video surveillance, where the context should provide additional background information to the already existing foreground object detection, in order to increase the scene understanding and improve complex object detection, like person groups. We have discussed an improved region labeling approach, featuring besides color and texture also the vertical position as part



(a) (b)

(c)

Figure 5: (a) Sample frame of our video sequence, (b) First abnormal group behavior detection [9] without using context at frame No. 275, (c) First abnormal event detection when using context of the video, at frame No. 195.

of a gravity-based model. For local feature extraction, we have selected a group of Gabor filters combined with the color feature. For fast classification, we have applied a random sampling for each segment and the subsequent multiple-SVM classification is based on a probability model of the segment to be classified as a specific region type. As an alternative, we have presented a system without the gravity-based approach, but with a novel Global Region Statistics (GRS) based model. This model involves the computation of mean and standard deviation of the vertical region positions. Our experimental results show that the gravity-based model gives the best results and outperforms two other existing region labeling algorithms, which are also suitable for context analysis. Our major contribution is introducing a general framework based on spatial context (in our case vertical position information) for labeling each region and its significant effect on surveillance event detection. The experimental result show that our method provides both qualitative and quantitative gains, while maintaining low complexity and high adaptivity to new semantic region types. Furthermore, we show in a case study that semantic region labeling as additional background information improves the result of moving group behavior analysis in surveillance applications, which increases the safety of people. Our framework is generic and does not depend on the type of scene, while our fast algorithms allow real-time execution.

REFERENCES

[1] X. Bao, S. Javanbakhti, S. Zinger, R. G. J. Wijnhoven and P. H. N. de With, *Context modeling combined with motion analysis for moving ship detection in port surveillance*, Journal of Electronic Imaging, vol. 22, 2013.

[2] B. C. Russell, A. Torralba, K. P. Murphy and W. T. Freeman, *LabelMe: a database and web-based tool for image annotation*, International Journal of Computer Vision, pp. 157-173, vol. 77, 2008.

[3] X. Bao, S. Zinger, R. G. J. Wijnhoven and P. H. N. de With, *Water Region Detection Supporting Ship Identification in Port Surveillance*, Advanced Concepts for Intelligent Vision Systems, pp. 444-454, Brno, Czech Republic, 2012.

[4] C. Millet, I. Bloch, P. Hède and P.-A. Moëllic, *Using relative spatial relationships to improve individual region recognition*, The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies, pp. 119-126, London, UK, 2005.

[5] O. Marques, E. Barenholtz and V. Charvillat, *Context modeling in computer vision: techniques, implications, and applications*, Multimedia Tools and Applications, vol. 51, pp. 303-339, 2011.

[6] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T-S. Chua and J. Li, *Hierarchical Spatio-Temporal Context Modeling for Action Recognition*, IEEE Conference on Computer Vision and Pattern Recognition, pp. 2004-2011, Singapore, 2009.

[7] C. Galleguillos and S. Belongie, *Context based object categorization: a critical survey*, Computer Vision and Image Understanding, vol. 114, pp. 712-722, Elsevier, 2010.

[8] N. Nguyen and Y. Guo, *Comparisons of Sequence Labeling Algorithms and Extensions*, Proceedings of the 24th International Conference on Machine Learning, pp. 681-688, New York, USA, 2007.

[9] I. M. Creusen, S. Javanbakhti, M. J. H. Loomans, L. B. Hazelhoff, N. Roubtsova, S. Zinger and P. H. N. de With, *ViCoMo: visual context modeling for scene understanding in video surveillance*, Journal of Electronic Imaging, vol. 22, 2013.

[10] G. Rellier, X. Descombes, F. Falzon and J. Zerubia, *Texture feature analysis using a Gauss-Markov model in hyperspectral image classification*, IEEE Transactions on Geoscience and Remote Sensing, vol. 42, pp. 1543-1551, 2004.

[11] W. K. Kong, D. Zhang and W. Li, *Palmprint feature xtraction using 2-D Gabor filters*, The Journal of the Pattern Recognition, vol. 36, pp. 2339-2347, Elsevier, 2003.

[12] J. Kamarainen, V. Kyrki and H. Kalvianen, *Fundamental frequency Gabor filters for object recognition*, In Int. Conf. on Pattern Recognition, Vol. 1, pp. 628-631, Quebec City, Canada, 2002.

[13] J. Shotton, J. Winn, C. Rother and A. Criminisi, *TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context*, International Journal of Computer Vision (IJCV), 2009.

[14] D. A. Migliore, M. Matteucci and M. Naccari, *A revaluation of frame difference in fast and robust motion detection*, Proceedings of the 4th ACM International Workshop on Video surveillance and Sensor Networks, pp. 215-218, New York, USA, 2006.

[15] S. Javanbakhti, S. Zinger and P. H. N. de With, *Fast sky and road detection for video context analysis*, Proceedings of the 33rd WIC Symposium on Information Theory in the Benelux, pp. 210-218, Boekeloo, Netherlands, 2012.

[16] A. Jain, A. Gupta and L. S. Davis, *Learning what and how of contextual models for scene labeling*, The 11th European Conference on Computer Vision (ECCV), pp. 199-212, Crete, Greece, 2010.

[17] C. Benedek and T. Sziranyi, *Study on color space selection for detecting cast shadows in video surveillance*, Special Issue on International Journal of Imaging Systems and Technology, John Wiley and Sons, Vol. 7 (3), pp. 190-201, 2007.