Real-Time Human Motion Detection and Tracking with Learning based Representation

Sandar Win, Thin Lai Lai Thein
University of Computer Studies, Yangon
sandarwin@ucsy.edu.mm, tllthein@ucsy.edu.mm

Abstract

Nowadays, real-time information is very important and learning based human motion has fascinated range from detection to tracking state in Computer Vision. In this system, the real-time videos are used to detect, track, and classify object or events in order to understand a real-world scene. Video based real time human motion detection and tracking is a complex and challenging task due to variation in human pose, shape variation, illumination changes and background appearance. A real-time mechanism is to detect the person and their moving within an environment from the video camera. This paper proposes human motion detection from video sequences. The proposed method includes three stages: human detection, motion tracking and accuracy result based on learning approach. The result is to become an efficient detection system for real-time human motion. Motion detection and tracking is determined by using Histogram of Oriented Gradients (HOG) feature extractor and Support Vector Machine (SVM) detector with learning human pattern which is well performed human detection and tracking in video sequences. Detailed analysis is carried out on the performance and accuracy of the system with the various test videos to show the results. The experimental results demonstrate *the efficiency of the method*

Keywords: *Human detection, Histogram of Oriented Gradients (HOG), Support Vector Machine (SVM)*

1. Introduction

Real-Time human motion detection, tracking and activity recognition is one of the most important system in computer vision. It interprets visual information from the surrounding environment. Human detection system can improve system's performance in fields such as security, safety, human activity monitoring in many environments. Detection of human from an image or video is an important step in human motion analysis due to the numerous variations of the human postures and the complexity of the surrounding environment. Object detection method can be organized into learning method approach and template method approach. [7] In the learning method, object features are obtained from training using positive or negative samples. In the template method, objects are expressed with templates and object detection process becomes to find the best matching result from an input image. The templates can be represented as intensity, shape and color image of the objects. Templates using geometric shapes are suitable for tracking objects because whose pose is not varied during tracking. Some of the templates are often seem too specific and less of generalization because the object is varied its location with respect to its background and illumination changes [6]. And people interact with each other, overlapping groups and may move in different directions. This requires a well-defined method which manages the different motions, different situations without being influenced by changes of environment features. To overcome changes in the environment monitoring by the system is required. The propose system is described based on adaptive background model and a robust human full-body model that without being altered by illumination changes such as sunny, rainy, windy, etc. And, the object is processed on a gray level to enhance picture quality. Then it is processed on a high level. Our goal is to develop a robust and efficient approach to detect and track for human. **Figure 1** shows an example of human detection and tracking result.

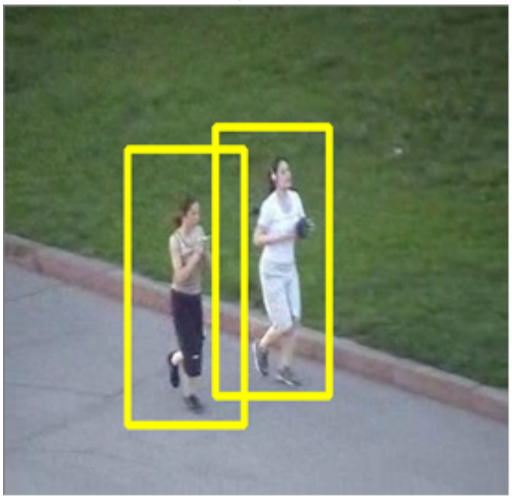

Figure 1: Example result of human detection and tracking

This paper is arranged as follow: Section 2 describes related work. Section 3 expresses processing stages for the detection of human motion. Section 4 shows human motion tracking. In session 5 describes experiment and accuracy results in detail. Session 6 presents the conclusion and future work.

2. Related Work

There are many approaches for learning based object detection using different features such as color, shape and texture or different methods. Basically color, texture and shape have been used to get good results and different methods have been proposed for human motion detection and tracking. Ismail Haritaoglu et al. [2] proposed 2D body modeling using silhouette boundaries shape-based analysis and dynamic appearance model defined on centroid, major axis, contour of its boundary. The real-time videos are captured by using an infrared camera. It can detect single person, multiple people, carrying different things, but color cues didn't use in this system. Viola et al. [9] described combination of image intensity information and appearance of motion information. They trained and tested with dynamic and static human motion pattern using Cascade classifier. But, if less than (20 x15) pixel window, it cannot detect human motion. N. Dalal and B. Trigs [5] proposed HOG feature extraction process for human detection. Detection process is based on the contrast of silhouette shape on background. It also described overlapping area for human detection. J. Grahn and H. Kjellstrom [3] stated the problems of classifying the different size of image patterns in video sequences. It used Linear Spatial-Temporal difference filters and SVM to detect human motion, but some problems are the person is walking away from the camera and between two humans can occur false positive state. Our approach continues this work also contain with different data that using human motion patterns to detect more robust system in our environment.

3. Human Motion Detection

Detection of human from an image or video is a crucial step in many application areas. The system takes monocular sequences of a scene and identifies the moving objects that are human or non-human. Object detection is defined as to discover and identify the existence of objects in the defined-class.

Object detection can be achieved by building a representation of the scene called the background model and then finding deviations from the model for each incoming frame. Any significant change in an image region from the background model signifies a moving object. The human detection is formulated as follows: given the image IW of a detection window W, determine whether the window contains a human or not by evaluate the following conditional probability as in (1)

$$P(Human\backslash I_w) \geq \theta \qquad (1)$$

Where θ is a threshold.

3.1 Background Subtraction

The background subtraction is a widely used approach for moving objects in videos. It involves absolute difference between current image and reference frame updated background over a period of time. This reference frame is called as background image. The background image is nothing when the representation of the scene with no moving object. The reference image needs to be updated regularly that it must adjust to the fade motion due to the leaves of the trees, water flowing, flag waving in the wind and other different variations.

3.2 Foreground Region Extraction

Foreground region extraction is to discriminate foreground regions from background area to detect any moving object. A natural approach is to segment those regions of the image that are moving relative to the background. The system makes a model of the background image over time. Those pixels with near zero are defined as background and other pixels with a larger result are defined as foreground. For any given video frame, the system subtracts background image pixels from the detected region. The background of the moving pixel is defined by using the probability density function (pdf) that expressed as in (2).

$$P_r(x_t) = \frac{1}{N}\sum_{i=1}^{N}\coprod_{j=1}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{(x_{tj}-x_{ij})^2}{2\sigma_j^2}} \qquad (2)$$

The pixel is regarded as a foreground pixel if $P_r\ (x_t) \leq th$ and a pixel is regarded as a background pixel if $P_r\ (x_t) > th$. An example of step by step background subtraction procedures are shown in **Figure 2 and 3**. After background subtraction and thresholding to obtain binary image, a region of interest is extracted and representing a moving person. After that morphological filtering is performed to reduce noise and shadow effects. Each image is normalized to represent in the form of a row vector such that the dimension of the vector is equal to number of pixels in the image. This algorithm is simple, efficient, and easy to implement.

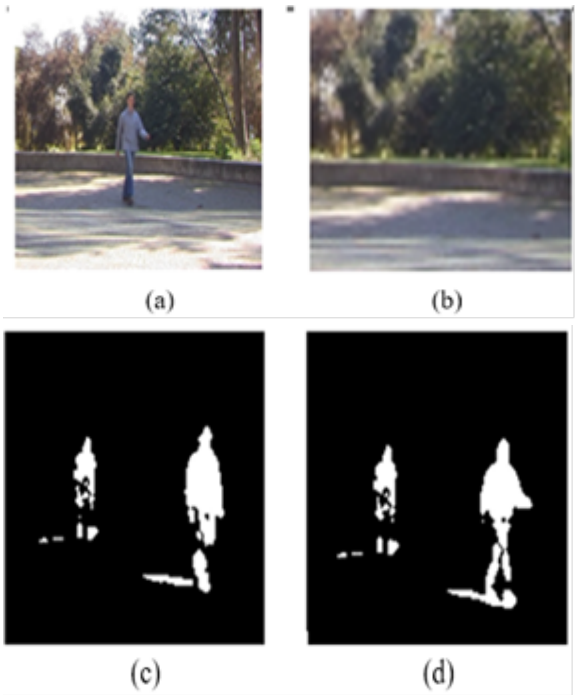

Figure 2: (a) Original image, (b) Background, (c) and (d) describe the correspondence foreground region with different frames

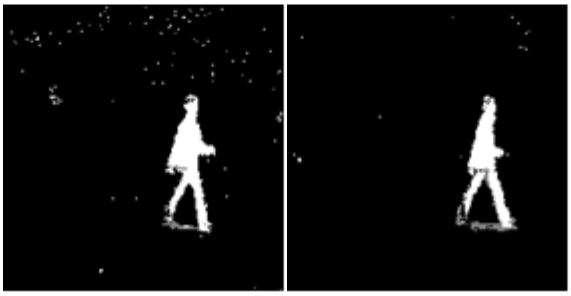

Figure 3: The result of foreground extraction using Background Subtraction algorithm based on Gaussian Mixture Model

3.3 Human Motion Learning Model

The learning of human motion from training examples have already been successfully exploited for parameterized motion estimation and activity recognition [1]. Similarly, we learned model for full-body human motion from training examples operated on motion captured data. Here we operated different appearance of human motions. Human Motion relies on the probabilistic generative models learned by training examples [11].

For example, a given human motion model M, θ(M) denotes the related view angle, $N_B(M)$ be the number of normal vectors and $B(M)=\{B_k(M), k\in \{1, N_B(M)\}\}$ the eigenvectors for the human motion. Then, we defined a phase as φ and a magnitude as γ, the system generates corresponding to the human motion model is computed by (3). Human motion learning model with different view angles are shown in **Figure 4**.

$$P_r(x_t) = \frac{1}{N}\sum_{i=1}^{N}\coprod_{j=1}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{(x_{tj}-x_{ij})^2}{2\sigma_j^2}} \qquad (3)$$

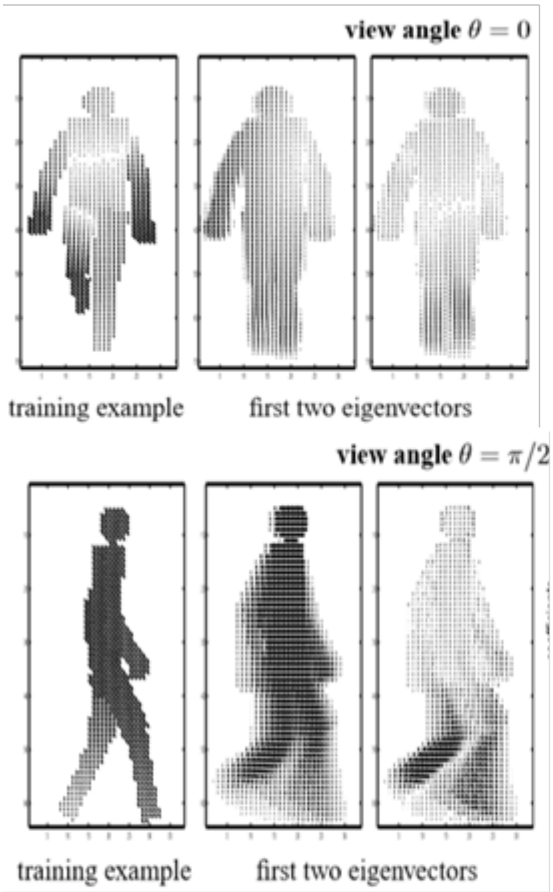

Figure 4: Human motion model with different view angles

The performance of real-time detection and tracking system in human motion depends on configuration of learning model and its ability to detect feature of moving objects (such as color, shape, contour, geometric pixel values, wavelets, etc.) in the observed environment.

To reduce the error detection rate, we employed cascade classifier [10] where each stage classifies different position and scale in the frame as “human” or “non-human”. **Figure 5** shows the stages of Cascade Classifier. If a sample has not contained a human being, it is thrown out immediately, saving valuable execution time for other samples. Samples that pass through all stages are considered to contain humans and the video frames with camera position, noise removal stage, low resolution and different illumination changes to detect true positive results. The examples of true positive and false positive results are shown in **Figure 6**.

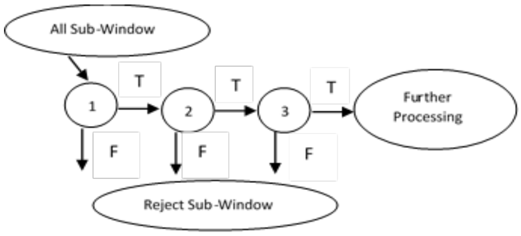

Figure 5: The stages of Cascade Classifier

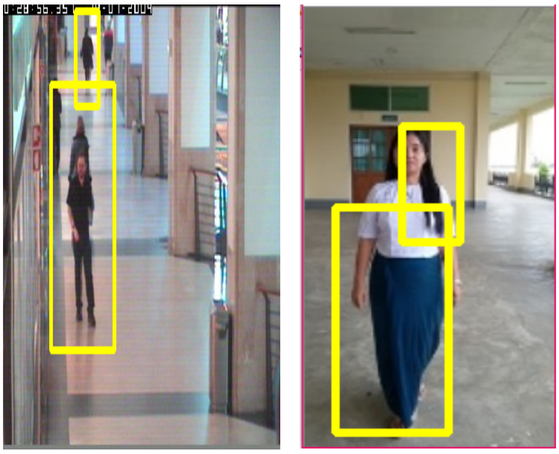

Figure 6: The examples of true positive and false positive results

4. Human Motion Tracking

The main goal of the motion tracking is to determine when object is entering in the scene. That motion region is a foreground region that has a difference between the current frame and previous frame more than a specified threshold. Each motion region is cluttered into the object. Each object is classified and become the template. We use this template to track the corresponding object in the next frame. The aim of object tracking is to find moving object in a frame extracted from video input. An object tracking system consists of three steps: object extraction, object recognition and tracking.

The task of a defined object tracker is to search region and identify them in each frame. In video sequence, an object is said to be as a motion with frame, it is changing its location with respect to its background. And then tracking algorithm is applied to estimate the position of each object. This system gives a high precision detection and tracking for moving object concerned with variation of appearance such as the presence of noise in foreground image, different poses in human, and changes of size, shape and scene in indoor or outdoor environment.

4.1 Feature Extraction

The main role of feature extraction is to transform visual information into the vector space which is used in computer vision. Because of the input image has comprised extra information that is not necessary for classification. We can find features that are more reliable for the system. The system uses HOG feature extraction algorithm which converts an image of fixed size to a feature vector of fixed size. This technique computes the flow of gradient orientation in localized portions of an image.

The image gradients are described by the gradient vectors that are associated with the gradient magnitude. Then, we apply mask patterns to calculate the auto-correlations of gradient vector which is weighted by the gradient magnitude. The system is scaling and normalizing the image to a standard size object. Extracted spatial information combined with the features in time domain that represent the trajectory of tracked object. In HOG feature extraction, 1st order differential coefficients are computed by (4).

$$\begin{cases} G_x(i,j) = f(i+1,j) - f(i-1,j) \\ G_y(i,j) = f(i,j+1) - f(i,j-1) \end{cases} \qquad (4)$$

$where\ f(i,j) means\ luminance\ at\ (i,j)$

The magnitude of the gradients m and the direction θ are computed respectively in (5) and (6)

$$m(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \qquad (5)$$

$$\theta(i,j) = \arctan\left(\frac{G_x(i,j)}{G_y(i,j)}\right) \qquad (6)$$

After that, it can be normalized with(7)

$$v = \frac{v_k}{\|v_k\| + \varepsilon} \qquad (7)$$

where V_k is the vector corresponding to a combined histogram for the block, ε is a small constant, and v is the normalized vector, which is a final HOG feature. The example results of HOG features extraction are shown in **Figure 7**.

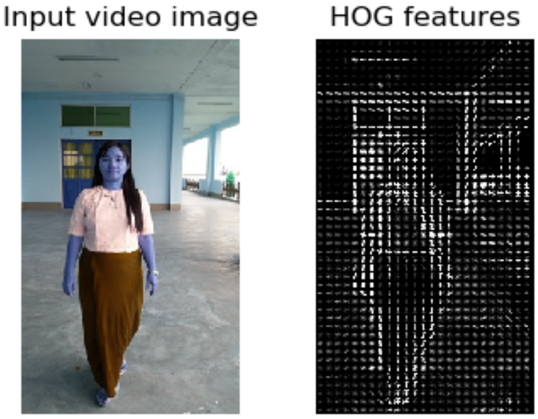

Figure 7: The example results of HOG features extraction

4.2 Learning with SVM

The system used a Histogram of Oriented Gradients (HOG) feature extractor to generate feature vectors. The generated feature vectors are classified with Support Vector Machine [4, 8]. The learning a binary classifier can be expressed as that of learning function g : R^n → ±1 that maps patterns x onto their correct classification y as y = g(x). In the case of SVM, the function g takes the form in (8)

$$g(x) = \sum_{j=1}^{n} y_j \beta_j k(x, x_j) + a \qquad (8)$$

where (x_j,y_j) is training pattern j wit classification, n is the number of training patterns, β and a are learned for weights, and k is a kernel function with (9)

$$k(x, x_j) = e^{-\|x-x_j\| /_{2\sigma^2}} \qquad (9)$$

Here, the patterns for which $\beta_j > 0$ are denoted as support vector.

The surface g(x) = 0, it is defined a hyperplane through the feature space by the kernel while the distances from this hyperplane to the support vectors are maximized as(10)

$$L_D = \sum_{j=1}^{n} \beta_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \beta_j \beta_j k(x_i , \qquad (10$$

In our system, we tested a number of sample videos using the different appearances of human motion based on indoor and outdoor environments.

5. Experiment Results

The results obtained in the implementation are shown in this section by using KTH dataset consists of 600 videos. The videos' frame rate are 25fps and their resolution is 160x120. The true positive results of the detection algorithm are presented and showed the tracking conditions, in which the approach works reasonably well as described in **Figure 8**. Some false detection and tracking results are shown in **Figure 9**.

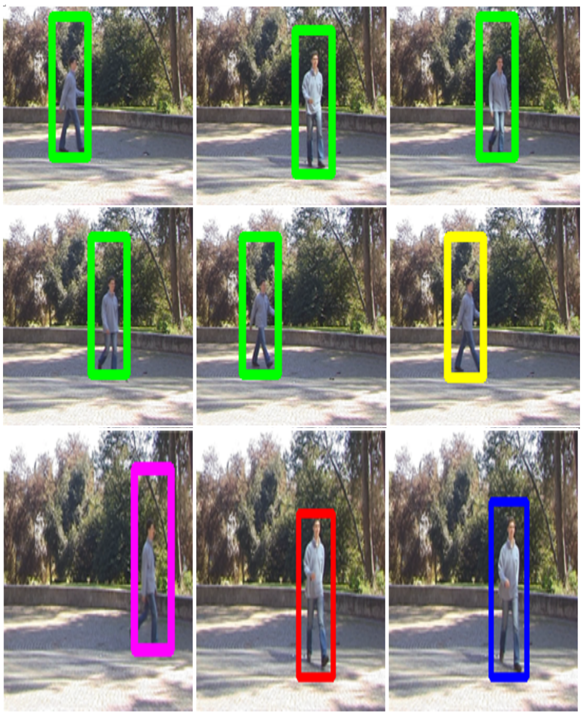

Figure 8: Example of detection and tracking results

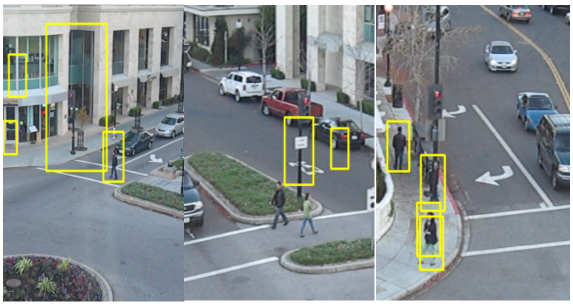

Figure 9: Example of some false detection and tracking results

5.1 Performance Evaluation

The video data sequences are split into training and testing sequences where 50% is used for training set and next 50% for testing set. The same person is never present in both training and testing sequences. From the training sequences, 174 positive and 126 negative examples are extracted and positive and negative features are used to train with SVM. Better results have been obtained, but occurred false positive between humans when multiple humans walk in close proximity to each other. Other false negative classifications are found when unrelated to any occurrences of humans. Performance score are measured by using Cross-validation method. The accuracy is expressed in **Figure 10** and **Figure 11** shows performance score that may be the range of Cross-validation score.

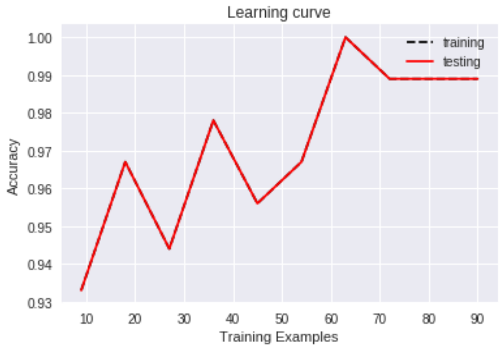

Figure 10: Accuracy result on training and testing

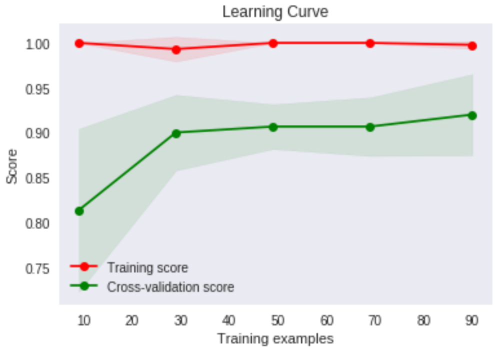

Figure 11: Performance evaluation score on experiment

6. Conclusion and Future Work

In reality, real-time information is very important and require an efficient detection and tracking method for human motion in order to know a real-world scene. There are many challenges concerned with different variation in human pose, shape, illumination changes and background appearance. In this paper, the system is implemented by using Histogram of Oriented Gradients feature extractor and Support Vector Machine detector with learning human pattern approach. The experimental results have been concluded that the method have a big dependency with different backgrounds, camera calibration and illumination changes. We trained and tested video data on different changes that are significantly increased the detection and tracking rate of our results.

Future research directions will continue 3D reconstruction for moving object to provide real scene of human motion detection, tracking and activity recognition system.

References

[1]. C.Bregler. “Learing and reconizing human dynamics in video sequences”. CVPR, pp. 568-574, 1997

[2]. Ismail Haritaoglu, David Harwood and Larry S. Davis, “W4: Real-Time Surveillance of people and their Activities”, Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22, No. 8, pp.809-830, August 2000.

[3]. J.Grahn and H. Kjellstrom “Using SVM for Efficient Detection of Human Motion” IEEE International Workshop of Tracking and Surveillance, pp. 231-238, Beijing, China2005.

[4]. N. Cristianini and J. Shawe-Taylor. “An Introduction to Support Vector Machines,” Cambridge University Press, Cambridge, UK, 2000.

[5]. N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection”, Proceedings of the Conference on Computer Vision and Pattern Recognition, Sam Diego, California, USA, pp. 886-893, 2005.

[6]. Nizar Zarks, Aaidi Al Halah, Rada Deeb “Real-Time Human Motion Detection and Tracking,” Telecom department, Higher Institute for Applied Sciences and Technology (HIAST), Faculty of Information Technology University of Damascus, ICTTA 2008, Syria.

[7]. Qifei Wang “ A Survey of Visual Analysis of Human Motionand Its Applications” Dept. of EECS, University of California, Berkeley CA 94720, USA, 2016.

[8]. S. Kang, H. Byun, and S-W. Lee. “Real-time pedestrian detection using support vector machines”. International Journal of Pattern Recognition and Aritical Intelligence, 17(3):405-416, 2003.

[9]. Viola, M. J. Jones, and D. Snow. “Detecting pedestrains using patterns of motion and appearance”.In ICCV, pages 734-741, 2003.

[10]. Viola, M. J. Jones, “Rapid object detection using a boosted cascade of simple features”. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 1,pp.511-518(2001).

[11]. Weiming Hu, Liand Wang, Tieniu Tan, and Huazhong Ning, “Automatic gait recognition based on statistical shape analysis”, IEEE Transations on Image Processing, vol.12, no.9, 2009.