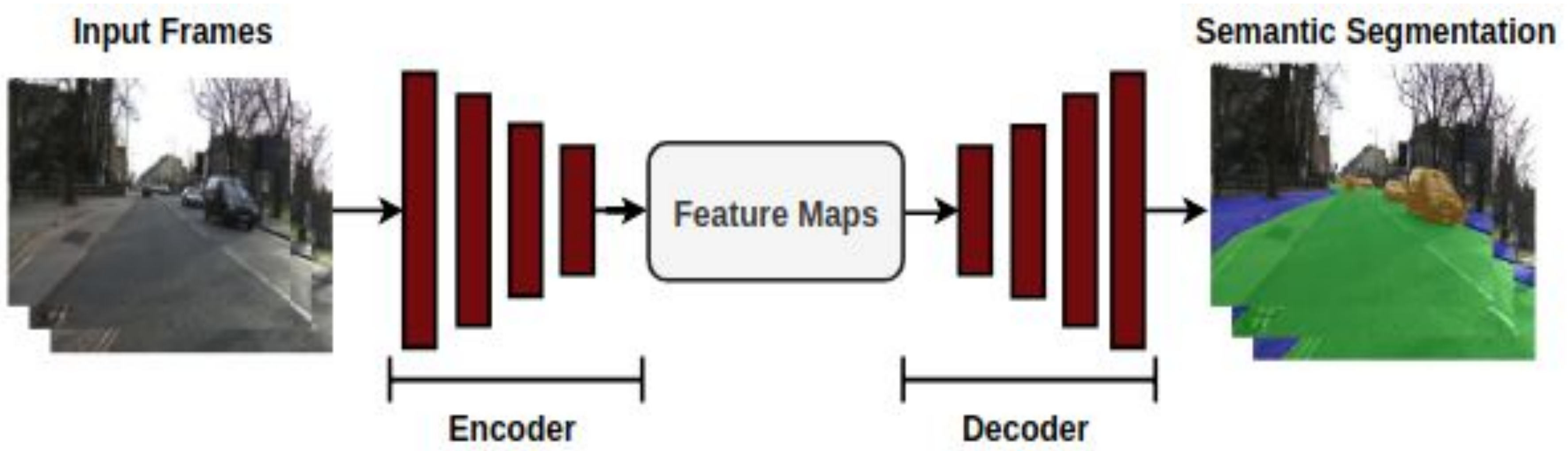
Semantic Segmentation in Compressed Videos

Ang Li, Yiwei Lu,Yang Wang
Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

Introduction



Problem:

Existing approaches for semantic segmentation in videos usually extract each frame as an RGB image, then apply standard image-based semantic segmentation models on each frame. This is time-consuming.

Goal:

We aim at building a faster semantic segmentation model by directly processing compressed videos.

Contributions:

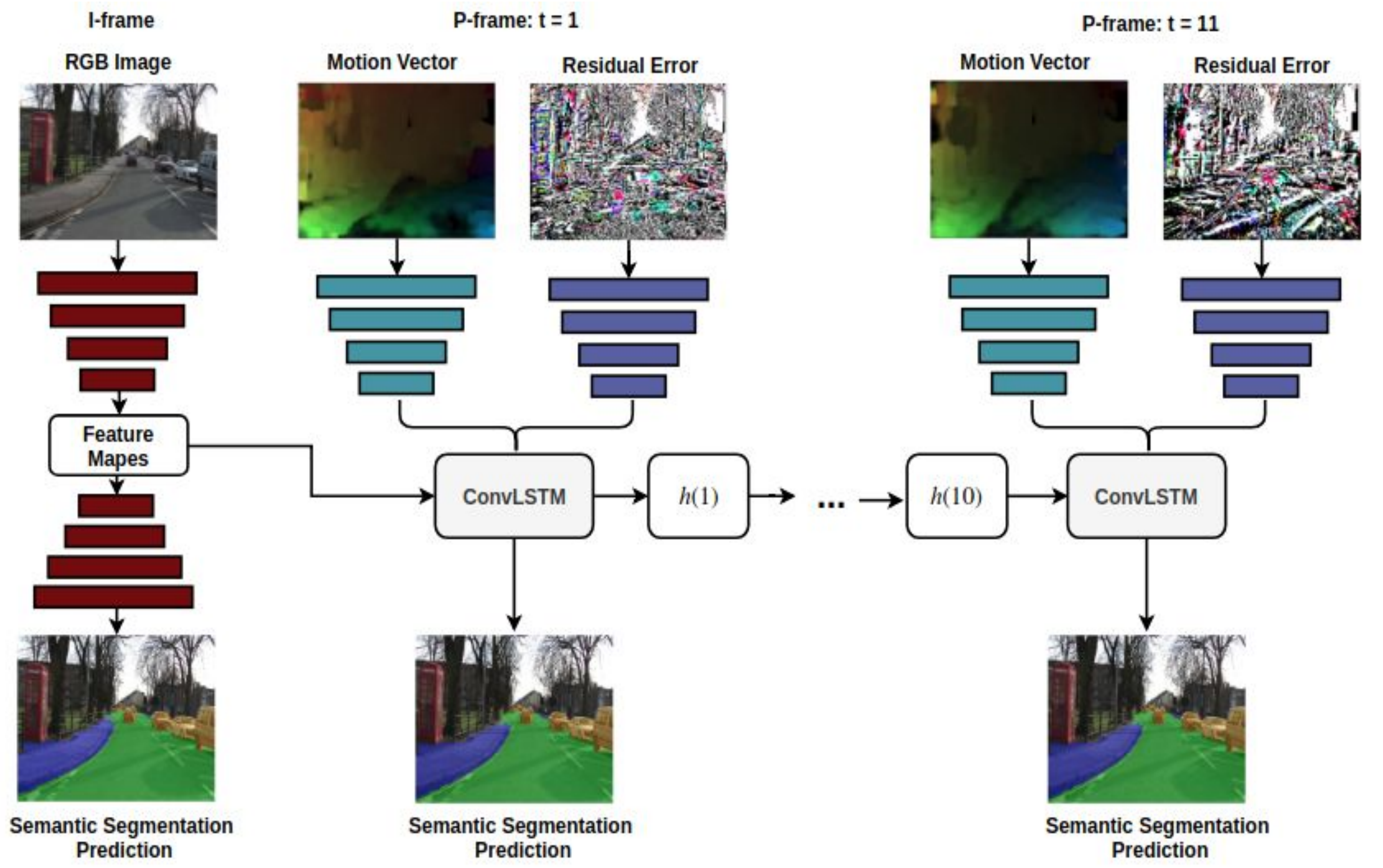
1. We propose a ConvLSTM model that propagates the temporal information from I-frame to succeeding P/B-frames for semantic segmentation.
2. Our experimental results show that the proposed method performs either better or on-par with standard frame-based methods. But the proposed method can run at a much faster speed.

Approach

Compressed Videos:

A compressed video contains three types of frames, I-frames,P-frames, and B-frames. I-frames are represented as regular images, P-frames are represented as motion vectors and residual errors, and B-frames are bidirectionally frames that can be regarded as a special case of a P frame.

Proposed Method:



We divide frames in an entire video into several groups, while each group contains one I-frame and several P-frames, represented by the collection $\{I, P_1, P_2, \ldots, P_T\}$.

Given the ground-truth semantic segmentation masks, our learning objective function can be described below:

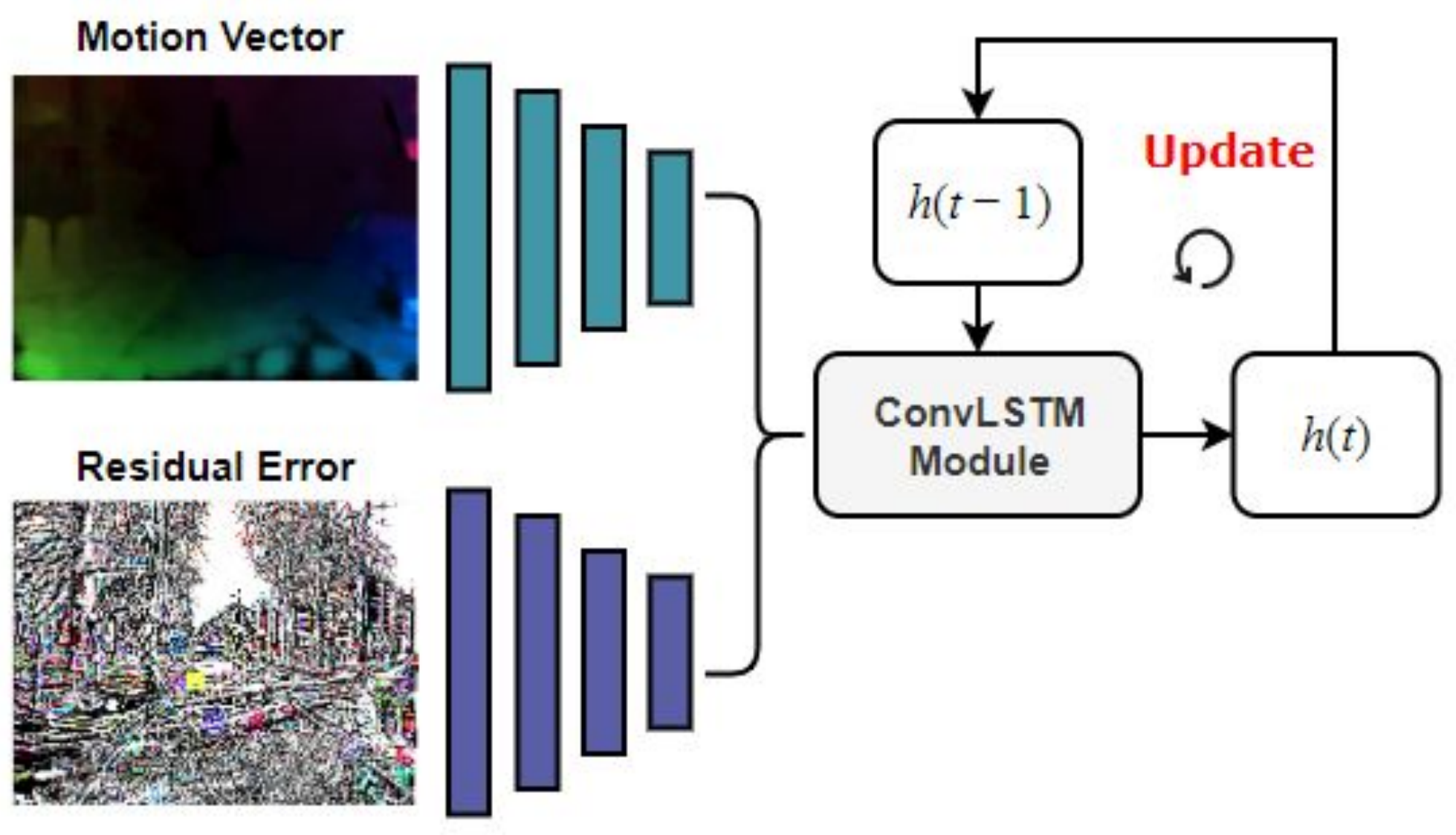
$$L = L_{ce}(GT_I - f_s(I)) + \sum_{t=1}^{T} L_{ce}(GT_{P_t} - f_s(P_t))$$

I-frames:

In order to obtain the semantic segmentation of an I-frame, we use a standard encoder-decoder architecture for semantic segmentation.

P-frames:

we apply a ConvLSTM module to accumulate the information of previous frames.



Experiments

Comparison of Performance:

Network	Pixel Accuracy	MeanIoU
FCN-32s [5]	91%	46.1%
FCN-8s [5]	92.6%	49.7%
ResNet [5]	95%	53%
Ours	**94%**	**51%**

Comparison of Inference Time:

Network	Inference time (ms per frame)
FCN-32s	42.5
FCN-8s	56
ResNet	168
Ours	**17**

Another Baseline:

This baseline first produces the semantic segmentation map on an I-frame. For remaining P-frames in the group, this baseline simply uses the semantic segmentation map from this I-frame as the prediction for each P-frame.

Comparison of Performance on this Baseline:

CamVid

Network	Pixel Accuracy	MeanIoU
Baseline	89%	25%
Ours	**94%**	**51%**

Cityscapes

Network	Pixel Accuracy	MeanIoU
Baseline	80%	22%
Ours	**87%**	**34%**

Our experimental results show that the proposed method performs on-par with frame-based methods in terms of accuracy. But our method can perform at a much higher speed during inference time. We believe our method can potentially be used in real-time applications where the efficiency is crucial.