Mohammad Pivezhandi

Wayne State University
Department of Electrical and Computer Engineering

November 25, 2025

1. Introduction
2. Feature Extraction Methods
3. CNN Accelerators
4. Vision Transformers
5. Spiking Neural Networks
6. Edge AI and TinyML
7. Conclusion

Problem Statement:

- Deep neural networks demand massive computational resources
- GPUs offer high throughput but consume significant power
- ASICs are efficient but lack flexibility
- Edge deployment requires low power and real-time processing

FPGA Advantages:

- Reconfigurable hardware
- Lower power than GPUs
- Customizable datapath
- Parallel processing capability
- Bridge between software and hardware

This survey covers FPGA accelerators for:

1. **Feature Extraction Methods**
   - Harris Corner Detection, SIFT, SURF
2. **Convolutional Neural Networks**
   - LeNet, AlexNet, VGG, ResNet, MobileNet
3. **Vision Transformers**
   - ViT, DeiT, Swin Transformer
4. **Spiking Neural Networks**
   - Neuromorphic computing, Event-based vision
5. **Edge AI and TinyML**
   - Low-power deployment, On-device learning

Harris Corner Detection:

- Detects corners using autocorrelation
- Stream processing architecture
- Low dynamic power optimization

SIFT Algorithm:

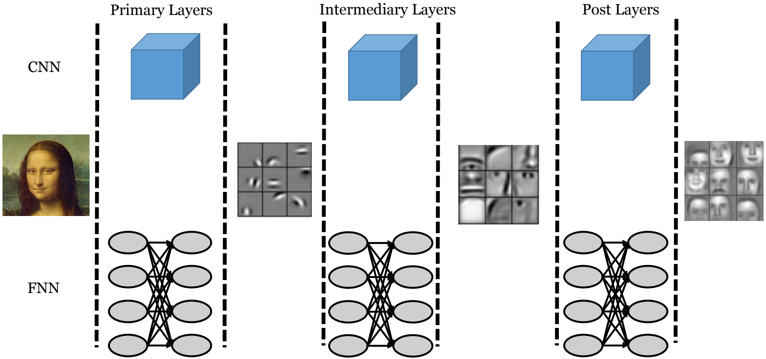
- Scale-invariant feature transform
- DoG pyramid construction
- Keypoint orientation assignment

SURF Algorithm:

- Speeded-up robust features
- Integral images for fast computation
- Hessian-based interest points

FPGA Benefits:

- Real-time processing
- Parallel filter computation
- Pipeline architecture



Key Components:

- Convolutional layers: Feature extraction with learned filters
- Pooling layers: Dimensionality reduction

Data Type Organization:

- Fixed-point vs. floating-point
- Custom precision (6-bit exponent, 5-bit mantissa)
- Dynamic quantization per layer

Storage Management:

- On-chip BRAM limitations
- Off-chip DRAM bandwidth
- Parameter encoding schemes

Bandwidth Optimization:

- Batch processing
- Weight reuse strategies
- Roofline model analysis

Loop Optimization:

- Loop unrolling
- Loop tiling
- Loop interchange

Algorithmic Optimization:

- FFT-based convolution
- Winograd transform
- Layer fusion

Quantization Methods:

- Dynamic fixed-point
- Binary Neural Networks (BNN)
- Incremental network quantization

Model Compression:

- Structured pruning
- Channel-wise pruning
- Sparse neural networks

Design Automation:

- High-Level Synthesis (HLS)
- OpenCL frameworks
- Hardware-aware NAS

Key Components:

- **Patch Embedding:** Divide image into 16×16 patches
- **Multi-Head Self-Attention (MHSA):** Compute attention between all patch pairs
- **Feed-Forward Network (FFN):** Two-layer MLP with GELU activation
- **Layer Normalization:** Applied before each sub-layer

Computational Challenge:

- Attention complexity: $O(N^2 \cdot D)$ where $N =$ number of patches
- For 224×224 image with 16×16 patches: $N = 196$

Work	FPGA	Model	TOPs/s	Power(W)	Eff.
ViTA'21	ZCU102	ViT-B	1.4	6.8	206
AutoViT'22	Stratix 10	DeiT-S	3.15	21.5	147
FlightBERT'23	Alveo U280	ViT-B	5.8	37.2	156
SwinAcc'23	Versal VCK190	Swin-T	4.7	38.3	123
MobileViT'24	ZCU104	MobileViT-S	0.89	3.2	278

Table: FPGA Vision Transformer implementations comparison (Efficiency in GOPs/W)

Biological Inspiration:

- Mammalian cortex: 100 billion neurons
- Each neuron: 1,000–10,000 synapses
- Spike-based communication
- Graceful degradation

Large-Scale Systems:

- IBM TrueNorth
- Intel Loihi
- Stanford Neurogrid
- Manchester SpiNNaker

SNN Advantages:

- Event-driven computation
- Energy efficiency
- Temporal processing
- Self-supervised learning (STDP)

FPGA Benefits:

- Higher throughput than CPUs/GPUs
- Flexibility vs. ASICs
- Rapid prototyping
- Technology independence

Dynamic Vision Sensors (DVS):

- Asynchronous pixel-level brightness change detection
- High temporal resolution (μs level)
- Low power consumption
- Natural fit for SNNs

FPGA Applications:

- Real-time optical flow calculation
- Object detection and tracking
- Frequency extraction from rotating objects
- Histogram creation for event processing

Constraints for Edge Deployment:

- **Power Budget:** <5W for battery-powered, <15W for powered devices
- **Latency:** Real-time requirements (1–100ms)
- **Memory:** Limited on-chip BRAM (few MBs)
- **Cost:** Low-cost FPGAs (Zynq-7000, Artix-7, Cyclone V)
- **Model Size:** Compressed models <10MB

Key Techniques:

- Sub-4-bit quantization
- Hardware-aware Neural Architecture Search
- On-device incremental learning

Application	FPGA	Model	FPS	Power(W)	Acc.(%)
ImageNet	PYNQ-Z2	MobileNetV2	47	2.1	71.8
Detection	ZCU104	YOLOv5s	42	6.8	44.2
Face Rec.	Artix-7	ArcFace	89	1.4	99.3
Segmentation	Zynq-7020	U-Net	37	2.8	94.1
Re-ID	Cyclone V	OSNet-BNN	156	0.9	91.2

Table: Edge AI FPGA implementations across different application domains

FPGA Advantages:

- High parallel processing capability
- Customizable datapaths
- Energy efficiency vs. GPUs
- Reconfigurability vs. ASICs

Optimization Strategies:

- Quantization (INT8 to binary)
- Pruning and sparsity
- Algorithmic transforms (Winograd, FFT)
- Hardware-aware NAS

Emerging Trends:

- Vision Transformers on FPGAs
- Edge AI deployment
- On-device learning
- Event-based vision

Future Directions:

- Neuromorphic computing
- Self-supervised learning
- Sub-mW vision systems
- Hybrid CNN-Transformer architectures

FPGAs provide an optimal balance between:

- Performance and Power Efficiency
- Flexibility and Customization
- Real-time Processing and Accuracy

Neuromorphic architectures represent the future of vision algorithms for real-time, energy-efficient processing.

Key references available in the survey paper:
Vision FPGA Accelerators: A Comprehensive Survey

Thank You!

Questions?

Mohammad Pivezhandi
Wayne State University
mpvzhndi@wayne.edu