The goal of this project is to produce a low energy, memory-bandwidth efficient embedded solution for realtime 3D reconstruction of dynamic environments from monocular video.

OVERVIEW

- **Problem**: Perceptually responsive system require realtime 3D knowledge of our social environment (e.g., pedestrians, cars)
- **Approach**: We will develop the theory and practice required to reconstruct, in realtime, the 3D scene structure and 3D camera motion from monocular video
- **Impact**: Enables technology for perceptually aware robotics
  - Allows robots to safely co-habit environments with humans
- **Applications**:
  - 3D video “tagging”: What is every pixel looking at
  - Collision avoidance: Spatial proximity of objects
  - Human robot interface: Safely co-habit the human world

**Example of Long-term Success**: Every video camera with embedded capability to ‘tag’ videos in 3D

SIFT MATCHING

- The overwhelming performance bottleneck in reconstruction
  - A SIFT (Scale Invariant Feature Transform) descriptor is a 128-dimension vector
  - HD video generates ~30,000 target descriptor (TD) per sec.
  - Find for each TD its two nearest neighbors (by L2-norm) in a corpus of 10,000 to 1,000,000 (depending on complexity of scene) reference descriptors (RD)
  - Realtime requires 0.3-30 billion L2-norm calculations per second against a very large memory-footprint corpus

**FPGA Accelerator Architecture**:

- **Throughput**:
  - Compute-bound: $R \cdot N / \text{corpus size}$
  - Memory-bound: $T \cdot BW / \text{corpus size}$

**Scene Dynamism**

- **State-of-the-art Algorithms**
  - Algorithmic improvement (e.g., use of priors)

- **Performance Requirement**
  - Current Performance
  - Target Performance

- **Input Images**
  - Static 3D Structure
  - Dynamic 3D Structure

- **Performance**
  - **Requirement**
  - **Scene Dynamism**

- **Social Robot**
  - **Social Environment**