Distributed Vision Processing in Smart Camera Networks

ICASSP 2009
Taipei
April 20, 2009

Hamid Aghajan
Stanford University, USA
aghajan AT stanford.edu

Part 2/4: Smart Cameras

http://wsnl.stanford.edu/ICASSP09/
Our Lab

Students (credits for the results presented):
Chen Wu
Jingyu Cui
Amir Khalili
Nan Hu, Tianshi Gao
Stephan Hengstler, Huang Lee, Itai Katz
Tommi Maataa (Philips, TU Eindhoven, Netherlands)
Linda Tessens, Marleen Morbee (Ghent U., Belgium)

EE Dept. - WSNL: Wireless Sensor Networks Lab

Y2E2 iRoom
• Smart cameras
• Case Study – Human pose analysis
Technology Cross-Roads

Image Sensors
- Rich information
- Low power, low cost

Sensor Networks
- Wireless communication
- Networking

Signal Processing
- Embedded processing
- Collaborative methods

Vision Processing
- Scene understanding
- Human gesture

Architecture? Algorithms? Applications?

Potential impact on design methodologies in each discipline
Vision

• Rich content
  – Window to the world

• Unobtrusive interface
  – Non-user-wearable

• Context-based processing
  – Many applications: Versatile high-level interfacing with common vision blocks

Assisted living

Gaming

Retail ads

Avatars

Face profile: Remote gaming
Multi-Camera Vision

• Added coverage
  – Areas of interest
  – Occlusion handling

  Smart homes: user behavior modeling

• 3D reconstruction

• Added confidence
  – Event interpretation

  Telepresence

  • Role selection
    Large-area view: Location
    Close-up view: Pose

Assisted living
Smart Environments

- Observe → interpret → build up behavior models → react

- Quantitative knowledge + Qualitative assessment
  - Sensing
  - Processing
  - Context
  - Behavior Model

- Responsive to events
  - Adapt services
  - Employ additional sensors
  - Send alerts

- Interactive
  - Based on gesture, location, region of interest of user

- Self configure, discover the interests, adapt to user

Vision can play an enabling role
Vision - Potentials

- Assistive technologies
  - Response systems
  - Companion robots

- Robotics

- Surveillance
  - Event detection
  - Identification / Tracking
  - Large-scale deployments

Enabling technologies:
- Vision processing
- Wireless sensor network
- Embedded computing
- Signal processing

Vision and Multi-modal Sensor Network

- Tele-presence
  - Virtual reality
  - Gaming over network

Multimedia

- Human Computer Interaction
  - Immersive virtual reality
  - Non-restrictive interface
  - Occupancy sensing

Distributed Vision Processing
Rich design space driven by application requirements

Camera Node

Vision System:
- Mono or stereo?
- Resolution?
- Field-of-View?

Data aggregation?

Data Exchange:
- Type of data?
- Traffic load?

Energy consumption?

Task:
- Tracking?
- Counting?

Distributed Observations

Camera orientation?
- Placement?

Which cameras sense?

Application Requirements:
- Accuracy?
- Coverage?
- Network Lifetime?

Network topology?

Vision algorithm:
- Local vs. central processing

WSNL - Stanford
Classical Multi-Camera Application: Surveillance
Classical Multi-Camera Application: Surveillance

<table>
<thead>
<tr>
<th>Network Intelligence</th>
<th>Network Objective</th>
<th>Required Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>Event description</td>
<td>low</td>
</tr>
<tr>
<td>medium</td>
<td>Object description</td>
<td>~ 10 KB/s</td>
</tr>
<tr>
<td>~none</td>
<td>Object detection</td>
<td>~ 1 MB/s</td>
</tr>
<tr>
<td></td>
<td>Moving scenes</td>
<td>~ 10 MB/s</td>
</tr>
<tr>
<td></td>
<td>Raw video stream</td>
<td>high</td>
</tr>
</tbody>
</table>

WSNL - Stanford
Smart Cameras

Image Sensor
CIF (352x288)
VGA (640x480)

Processing Unit
32-Bit RISC, 20-200 MHz

Radio
Data Rate kB/s to MB/s
802.11 and 802.15.4

Energy Source

Storage
SRAM & Flash (MBs)

Resource constraints:
Computation, energy, communication bandwidth
Big Picture

- Process locally … Fuse globally
  - Move away from streaming raw video
  - “Smart” cameras: Local processing power

Algorithm levels implemented in different processors
• Process locally … Fuse globally

– Algorithm design dependent on system and application:
  • Network’s scale and size
  • Available bandwidth
  • Processing powers (embedded vs. central)
  • Application requirements (accuracies, latency, data fusion level)

Multi-camera hardware & network

• Local processing and centralized processing
  • Communication bandwidth
  • Latency of real-time results
  • Resolution in image view and time
  • Temporal alignment (synchronization)
  • Camera view overlaps, data redundancies
  • Data exchange methods

Vision algorithms
Big Picture

• Process locally … Fuse globally
  – Different levels of local processing:
    • Extract generic features (e.g. silhouette, edges)
      – Low order of magnitude data reduction from raw video
    • Report mid-level objects (e.g. segment area of interest)
      – High order or magnitude data reduction
  • Decision-level processing (e.g. classify an action)
    – Small number of information bytes
The Issue of Privacy

– Cameras:
  • Offer a non-wearable sensing option (unobtrusive ..)
  • However, are often regarded as rather *invasive* sensing

– Privacy concerns MUST be addressed for home applications

  Added motivation for “Smart” Cameras
The Issue of Privacy

• Smart cameras + a multi-layered privacy handling approach:
  – Turn video into text in normal state (as well as at alerts)
  – Map person’s gesture onto: silhouette, avatar

• Alert mechanism:
  • Implement multi-level alert system (green - yellow – red)
  • Activate voice communication first to check status
  • Image query only possible by authorized nurse / family
  • Raw video saved locally for post-event analysis / diagnosis
Distributed Processing

Distribution across processors

Distribution across space

Vision processing tasks

CAM 1  CAM 2  CAM 3

Algorithm levels implemented in different processors

Low  Intermediate  High
Distributed Processing

Distribution across processors

High Level
- High-level Reasoning
  - Event detection
  - Posture estimation
  - Tracking

Intermediate Level
- Object Processing (Image Processing / Low-level Reasoning)
  - Segmentation
  - Shape analysis / coding

Low Level
- Pixel Processing (Image Processing)
  - Filtering
  - Template matching
  - Background subtraction
  - Pixel grouping

Algorithm levels implemented in different processors
Distributed Processing

WiCa1.1

ZigBee Channel

ZigBee coordinators

Distribution across space

WiCa

IC3D
(SIMD processor)

DRAM

8051
(General processor)

WiCa 1

WiCa 2

IC3D
(SIMD processor)

DRAM

8051
(General processor)

WiCa 3

IC3D
(SIMD processor)

DRAM

8051
(General processor)

PC

Receiver

Processor

Distribution across processors

WiCa: NXP Semiconductor Research
Layered Processing

Description Layers

- **Description Layer 4**: Actions, labels
- **Description Layer 3**: Poses, attributes
- **Description Layer 2**: Low-level features
- **Description Layer 1**: Image / video

Processing Layers

- **Processing Layer 3**: Interpretative
- **Processing Layer 2**: Collaborative
- **Processing Layer 1**: Distributed

Multi-camera networks: Distribution across space
Fusion and Feedback

- Initialize in-node feature extraction
- Focus on what is important
- Assign tasks to cameras

Active vision

Feedback

- Estimate fusion
- Decision fusion
- Model-based fusion

WSNL - Stanford
The Big Picture

Network Feedback (robustness, efficiency)
- Low-level vision: appearance
- High-level: activity interpretation
Interfacing Vision

- What accuracies / observation frequencies are needed?
- Task assignment to cameras
- Priorities of parameters to extract
- Process based on available contextual information
Interfacing Vision

Description Layers
- Description Layer 1: Image / video
- Description Layer 2: Low-level features
- Description Layer 3: Poses, attributes
- Description Layer 4: Actions, labels

Processing Layers
- Processing Layer 1: Distributed
- Processing Layer 2: Collaborative
- Processing Layer 3: Interpretative

Behavior analysis

Instantaneous action

Low-level features

Posture / attributes

Model parameters

AI reasoning

Vision Processing

Feedback
- Queries
- Context
- Persistence
- Behavior attributes

Interpretation levels
Syllabus

- Smart cameras
- Case Study – Human pose analysis
Posture Estimation – Review

- **Discriminative** -> *template-based*
- **Generative** -> *model-based*
  - **Bottom-up**
  - **Top-down**
- **Combined**
  - Discriminative for body parts
    - Detect each body part as a unit
  - Generative for whole-body configuration
    - Find best model to match composition of all parts

**Multi-View Issues**

**Opportunities:**
- Complementary info
- Occlusion handling
- Outlier rejection
- Distributed processing

**Challenges:**
- Correspondence
- Redundant data
- Misleading info in some images
- Communication (bandwidth, latency)
Posture Estimation

Bottom-up

Top-down
Multi-View Camera Networks

• Combine bottom-up and top-down approaches
  – Powerful local image processor
  – Limited communication

  - Generative (model-based) for whole-body configuration
  - Discriminative (template-based) for body parts

- Vision processing options:
  – Segmentation with generic features
  – Opportunistic segmentation -- detection of body parts
Pose Estimation – Top-Down Approach

- 3D model -> 2D projections of edges and silhouettes
- Validate 2D projections with image observations
  + Easy to handle occlusions
  - Difficult to optimize: non-convex
  - Time consuming in calculating projections and evaluating them
Pose Estimation – Bottom-Up Approach

- Look for body part candidates in images
- Assemble 2D/3D models from body part candidates

+ Distribute more computation in images (body part candidates, local assemblage)

- Difficult to handle occlusions without knowing relative configuration of body parts
- Not direct to map from 2D assemblage to the 3D model
Model-based Fusion

- Motivation to build a human model:
  - A concise reference for merging information from cameras
  - Universal interface for different gesture interpretation applications
  - Allows new viewing angles in virtual domain
  - Facilitates active vision methods:
    - Focus on what is important
    - Exchange descriptions only relevant to the model
    - Develop more detail in time
    - Initialize next operations (segmentation, motion tracking, etc)
  - Helps address privacy concerns in various applications

http://wsnl.stanford.edu/videos/gesture/rotate2.avi
Case Study: Pose Analysis

What is the problem we try to solve?
- Reconstruct ion of detailed dynamic body model
  - Has to be real-time?
- Detect ion of certain poses (gesture control, fallen, …)
  - How critical is missed detection? Or false alarm?
- Extraction of long-term behavior routines
  - Afford to make short-term mistakes? Can ignore low-confidence frames?

System constraints?
- Real-time, frames-per-second
- Local versus central processing power
- Communication bandwidth
- Latency
Case Study: Pose Analysis

Graphical Model

Optimal solution for body model reconstruction

Kinematic edges:
Angle and distance constraints

Silhouette + Edge

Optimization

Silhouette + Edge

Silhouette + Edge

Distributed Central

Image processing power (operations per pixel)

Each camera sends silhouette and edge maps

Graphical model algorithm

400 (for ZigBee channel)

80k

Not to scale

Communicated data (bytes per frame)
**Requirements**

- **Real-time:**
  - 30 fps
  - Latency of 10 ms

- **Wireless link:**
  - 100 kbps data per channel / 30 fps ~ 400B/frame

Joint work with NXP Semiconductors, The Netherlands
Case Study: Human Pose

Generic Features
Generic Features

Generic features for body parts

- More processing power at cameras
- Limited communication bandwidth

Distributed Processing: Segmentation

- Background subtraction
- Rough segmentation
- Refinement
- Ellipse fitting

3D human body model

Collaborative Processing: Model Fitting

- Check stop criteria
- Configuration using previous geometric configuration and motion
- Generate test configurations

Update 3D model (color/texture, motion)

Ellipse parameters are sent to central processor
- Reduced data communication load

WSNL - Stanford

Distributed Vision Processing
Segmentation - Generic Features

- Images
- Optical flow estimation
- Background subtraction
- Markers for the person
- Watershed segmentation
- Foreground
- Background
- Body part segments
- Ellipse fitting and attributes extraction
- Watershed
- Info from model
- K-means clustering (color)
Distributed Processing

Initialize from model, or refresh (k-means)

Refine color models (adaptivity)

Enforce spatial connectivity for ambiguous pixel colors

Concise description of segments

Color segmentation and ellipse fitting in local processing

Distributed Processing: Segmentation

Background subtraction → Rough segmentation → POEM: refine color models → Watershed segmentation → Ellipse fitting

Feedback

Previous color distribution

Maintain current model

Update 3D model (color/texture, motion)

3D human body model

Previous geometric configuration and motion

Combine 3 views to get 3D skeleton geometric configuration

Collaborative Processing: Model Fitting

Check stop criteria

Generate test configurations

Local processing from other cameras

WSNL - Stanford

Distributed Vision Processing

39
Distributed Processing
Collaborative Model Fitting

- Exchange segments and attributes, combine to reconstruct a 3D model
- Subject’s information mapped and maintained in the model:
  - Geometric configuration: dimensions, lengths, angles
  - Color / texture / motion of different segments
Collaborative Model Fitting

- Red: projection of skeleton on image plane
- Green: region of arms grown from red lines
- Blue: ellipses from segmentation
  - Score = Area (ellipses falling within green polygons) / Area (green polygons)
Model-based Pose – Generic Features

Frame 105

[Images of a person in various poses with corresponding 2D and 3D plots showing motion and pose features]
Model-based Pose – Generic Features
• Generic features are used to detect body parts
  – Can improve using more specific features for each body part?
Case Study: Human Pose

Opportunistic Features
Opportunistic Segmentation

In the 2D image

- Chamfer-based matching
- Skin blob detection
- Histogram of x-axis projection of foreground
- Hierarchical hough-line detection
- Skeletonizing thighs and calves

Head candidates
- Hands candidates
- Torso width
- Line segments delineating upper body
- Skeletons of thighs and calves

Opportunistic Reconstruction

- Different features for body parts
- Limited communication bandwidth

http://wsnl.stanford.edu/videos/gesture/features3.avi

Image observation
- Chamfer-matching score $s$

penalty
- Short-term observation (decay: lack of observation for $n_1$ frames)

reward
- Short-term observation (inertia: consecutive observations for $n_1$ frames)
- Long-term observation (Number of frontal face $n_3$ presences)
Collaborative Processing

- Multi-camera validation
  - Outlier rejection
  - Occlusion handling

- Model construction

Camera 3 will not participate in hand modeling
Collaborative Model Construction

In 2D images

- Camera 1, 2, 3: Head candidates → Back projection → Kalman filtering → Head position
- Camera 1, 2, 3: Torso width → Ratio between views → Torso orientation
- Camera 1, 2, 3: Hands candidates → Back projection → Kalman filtering → Hands positions

Sampling of arms in selected views

- Camera 1, 2, 3: Distance maps from line segments delineating upper-body → Chamfer-based matching → Filtering → Arms angle configurations
- Camera 1, 2, 3: Orientation of thighs and calfs → Back projection → Legs angle configurations

In 3D space
Collaborative Model Construction

Back projection ➔ Kalman filtering

Sampling ➔ Chamfer matching

Torso angle

In the 2D images ➔ In the 3D space

Camera 1, 2, 3: Head candidates ➔ Back projection ➔ Kalman filtering ➔ Head position
Camera 1, 2, 3: Torso width ➔ Ratio between views ➔ Torso orientation
Camera 1, 2, 3: Hands candidates ➔ Back projection ➔ Kalman filtering ➔ Hands positions

Sampling of arms in selected views ➔ Chamfer-based matching ➔ Filtering ➔ Arms angle configurations
Camera 1, 2, 3: Distance maps from line segments delineating upper-body ➔ Orientation of thighs and calves ➔ Back projection ➔ Legs angle configurations

WSNL - Stanford
Distributed Vision Processing
Collaborative Model Construction

http://wsnl.stanford.edu/videos/gesture/combine1.avi
Collaborative Model Construction

http://wsnl.stanford.edu/videos/gesture/rotate2.avi

http://wsnl.stanford.edu/videos/gesture/jogging1.avi
Collaborative Model Construction

http://wsnl.stanford.edu/videos/gesture/pang.avi
Communication Load

• Data record per frame:

<table>
<thead>
<tr>
<th>Body part/Feature</th>
<th>Head</th>
<th>Skin blobs</th>
<th>Torso</th>
<th>Edge line segments</th>
<th>Legs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Size (bytes)</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>~336</td>
<td>32</td>
<td>~382</td>
</tr>
</tbody>
</table>

Each pixel requires many processing passes → Line memory limitation
Embedded Implementation

• Further hardware constraints:
  – Multiple image passes are required
  – Line memory available on WiCa allows ~1 pass per full frame
  – This imposes severe limit on the algorithm

• Process a small subset of features:
Real-Time WiCa Implementation

Face detection

Skin color adaptation

Hand detection

Shirt color adaptation

Shoulder detection

2D body part positions

WSNL - Stanford
Distributed Vision Processing
30 Frames per Second

http://wsnl.stanford.edu/videos/gesture/realtime1.avi
Ping Pong
Ping Pong

ICDSC (International Conference on Distributed Smart Cameras)
Sept 2007, Vienna, Austria

http://wsnl.stanford.edu/videos/gesture/realtime2.avi
Spatiotemporal Smoothing

From smart camera 1

Parse ZigBee Msg

2D body part coordinates

Noise filtering in a single camera

Median filtering in a sliding window

prediction

From smart camera 2

Parse ZigBee Msg

2D body part coordinates

Noise filtering in a single camera

Median filtering in a sliding window

prediction

Noise mitigation from 2 cameras

2D to 3D lifting

No smoothing

Two-camera feature fusion and temporal smoothing

http://wsnl.stanford.edu/videos/ballgame/comparison32.avi
WiCa Implementation

Image processing power (operations per pixel)

- Graphical model algorithm
- Collaborative model construction
- WiCa implementation

Not to scale
Case Study: Pose Analysis

- May only need high-level posture state in some applications
  - E.g. assisted living
Publications

http://wsnl.stanford.edu
Distributed Vision Processing in Smart Camera Networks

ICASSP 2009
Taipei
April 20, 2009

Hamid Aghajan
Stanford University, USA
aghajan AT stanford.edu

Part 2/4: Smart Cameras

http://wsnl.stanford.edu/ICASSP09/