Computer Vision for Augmented Reality on Mobile Platforms

Research Problem and Initial Plans

PhD Thesis – Collection of Papers – University of Jyväskylä
Matti Johannes Eskelinen
matti.j.eskelinen@jyu.fi

Author

- Matti Johannes Eskelinen *matti.j.eskelinen@jyu.fi*
- 31 years, married, two sons (3 and 4 years)
- M.Sc. in Computer Science and Mathematics from Jyväskylä University in 2005
- Working as software engineer for 5 years after that
- Last 3.5 years in mobile software industry
- Ph.D. student since 2010 at Jyväskylä University
- Working on first publications with COMAS funding
- Target for finishing set at 2014

Supervisors

- Tommi Kärkkäinen, professor

- Tuomo Rossi, professor

- Ville Tirronen, Ph.D.

- All from Jyväskylä University, Faculty of Information Technology, Department of Mathematical Information Technology

Long-term interests

- High-level computer vision

- Understanding general visual scenes

- Utilizing that understanding for augmented reality and personal robotics

- 'Artificial Visual Cognition'

- Time-frame 10-20 years

- Not general intelligence, but more 'mechanical' probabilistic visual understanding that is a key survival method of many quite simple creatures as well, not only highly developed mammals like humans

Problem of vision

- Examine a complex scene, not directly, but by analyzing *reflected light* collected by a sensor

- Using a *2D projection* of a 3D scene

- Spatio-temporal data *represented as a matrix* of changing light intensity values

- No *unique solution* is possible to infer!

  ➔ Probabilistic processing required

- Still, humans and numerous other creatures do this quite well, which gives us a proof that vision is, indeed, possible, with highly useful results.

Augmented reality

"Augmented reality (AR) is a term for a live direct or indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input, such as sound or graphics." *http://en.wikipedia.org/wiki/Augmented_reality*

- Enhancing one's current perception of reality

- Ronald Azuma 1997: *A Survey of Augmented Reality*

  - Combining real and artificial (virtual) data

  - Interactive in real time

  - Registered in 3D

Between reality and virtuality

- Paul Milgram and Fumio Kishino 1994:
A Taxonomy of Mixed Reality Visual Displays
Reality-Virtuality Continuum:

Reality   Mixed Reality   Virtuality

- Person 'looking at' computer
- Separate perception of both reality and virtuality
- Hard to focus on both

- Person 'inside' computer
- Perception of virtuality only
- Hard to switch focus

Augmented Reality   Augmented Virtuality

- Person looking at reality or virtuality 'through' a computer
- Allows immersion and focus to relevant aspects of real and virtual

Possible applications

- Medical: 'virtual x-ray vision' by overlaying images and test data over patient during surgery

- Service and maintenance: overlaying sensor data and service instruction diagrams over large machinery

- Navigation and discovery: overlaying locations of friends and landmarks, visualizing routes

- Games, magic books, augmented archeological sites and museums, multi-party videoconferencing around virtual table, remote collaboration on virtual model...
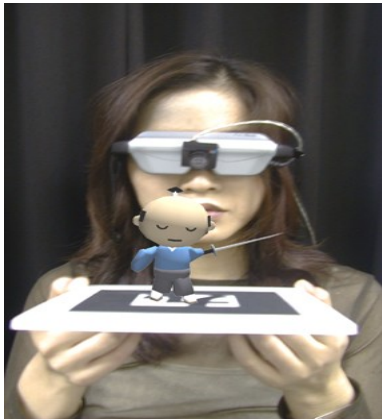
- Enables a whole new interface paradigm

Possible applications

Magic Book
http://www.hitlabnz.org/MagicBook

http://www.edopter.com/trends/Augmented_Reality

http://www.layar.com/

http://www.vuzix.com/ar/products_wrap920ar.html

http://thenextweb.com/2009/06/23/augmented-reality-beginning-tourism/

Terms

- Tracking: establishing position and optionally orientation of interesting objects in the scene

- Pose estimation: establishing viewpoint, position and orientation of user's head in case of head-mounted display, and device in case of hand-held display

- Registration: alignment of real and virtual information, has to be precise for not breaking the illusion
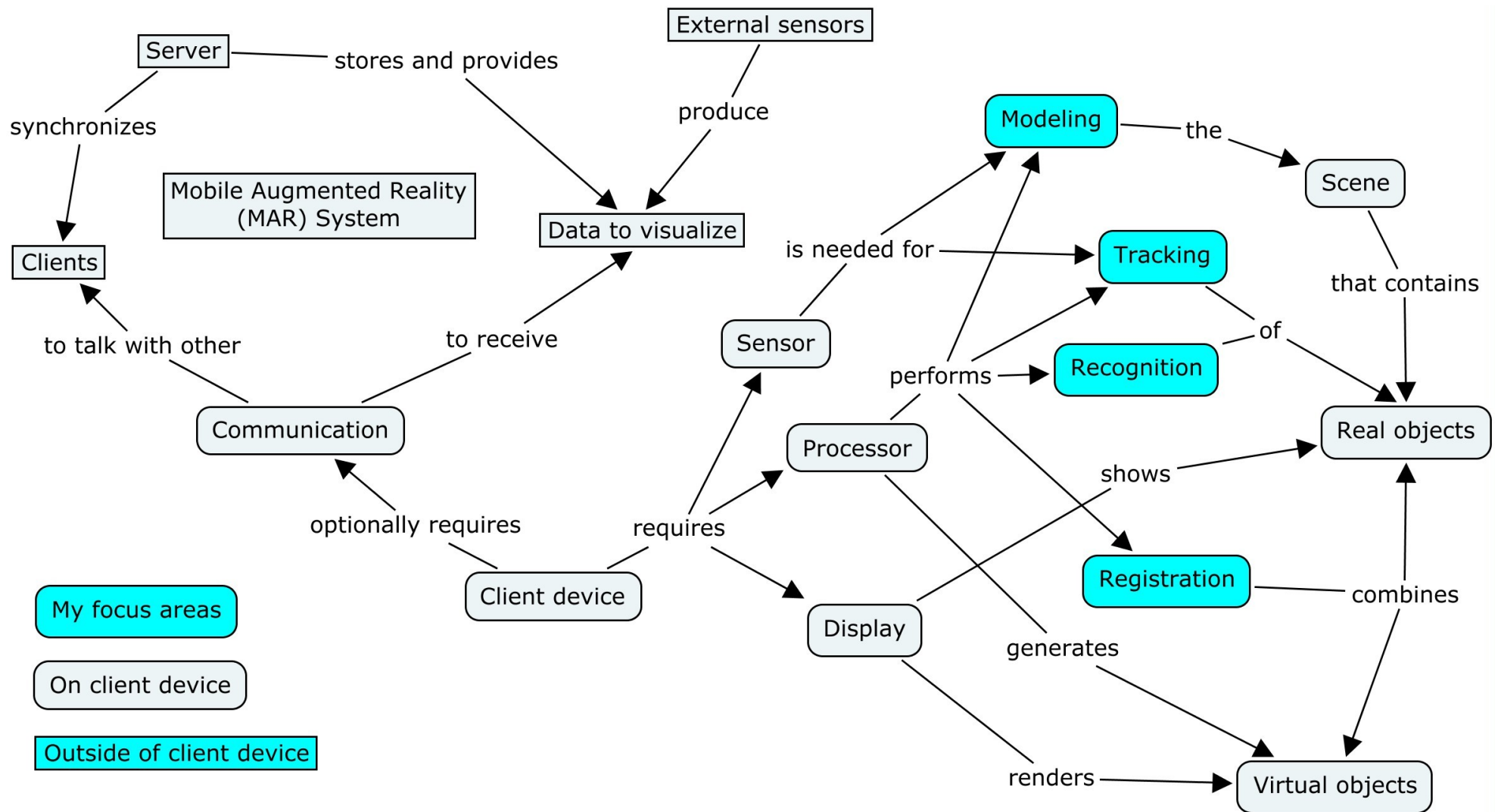
Requirements for AR systems

- Sensor for tracking and registration

  - Ultrasound sonar, laser range-finder, inertial sensor, camera...

- Processing unit for analyzing sensor data and generating the augmentations

- Display for showing the combined view

  - Head-mounted display, see-through display, hand-held display

- Data communication is also often useful

- Current smartphones provide all of these...

Why mobile platforms?

- Smartphones and other mobile devices are widely available and affordable, and demand for useful applications is high

- Decent cameras, integrated inertial sensors, compass and GPS, big displays, and increasingly powerful processors enable augmented reality applications

- Enable socially acceptable 'magical lens' or 'window to augmented world' metaphors, create interest in augmented reality, encourage development of more sophisticated algorithms – ultimately generate demand for immersive headset-based systems
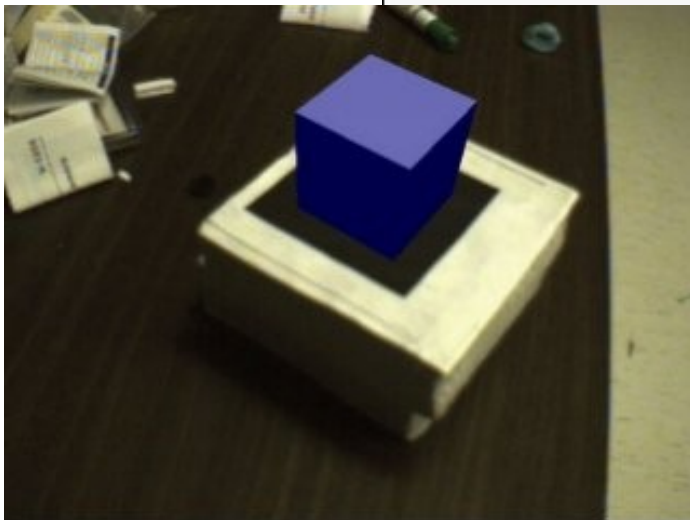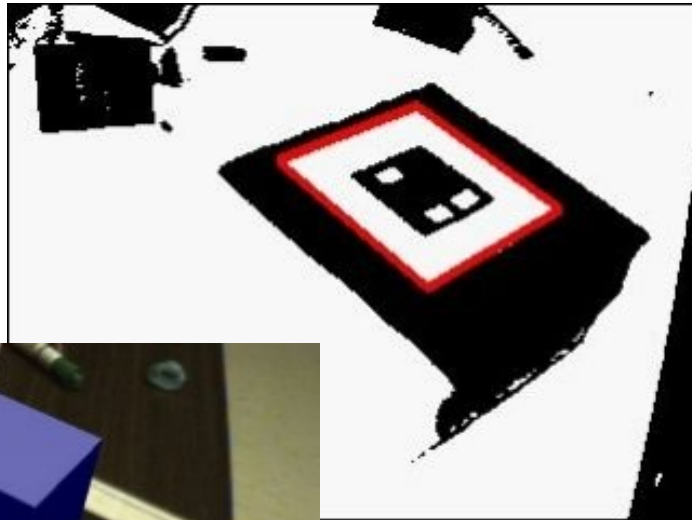
Mobile augmented reality system

Server — stores and provides → Data to visualize

External sensors — produce → Data to visualize

Server — synchronizes → Clients

Clients — to talk with other → Communication

Communication — to receive → Data to visualize

Client device — optionally requires → Communication

Client device — requires → Processor

Client device — requires → Display

Processor → Sensor

Data to visualize — is needed for → Tracking

Modeling — the → Scene

Scene — that contains → Real objects

Tracking — of → Real objects

performs → Recognition

performs → Registration

Modeling

Display — shows → Real objects

Processor — generates → Virtual objects

Registration — combines → Real objects / Virtual objects

Display — renders → Virtual objects

Mobile Augmented Reality (MAR) System

My focus areas (cyan)

On client device

Outside of client device (cyan)

13

Cmaptools: http://cmap.ihmc.us/

Traditional solutions

- Placing planar fiducial markers in the scene

  - Tracking the markers is easy as they are designed for maximum contrast

  - Known marker shape makes pose estimation easy

  - Helps in establishing a coordinate frame

- Creating a CAD model of the location

  - Identifying landmarks extracted from rendered model in the viewed scene

  - Effectively aligning a CAD model to real scene

Traditional solutions

http://www.hitl.washington.edu/artoolkit/

Gerhard Reitmayr et al.: Going out: Robust Model-based Tracking for Outdoor Augmented Reality

15

My target

- Visual tracking and registration in unmodified scene with minimal prior information

- A robust and working software

- Running on a mobile device (Nokia N900 initially)

- Provides a visual framework for AR applications

  - Analyze an unknown scene

  - Model the environment at some level of detail

  - Find interesting objects

  - Enable various schemes for recognizing and identifying objects in the scene

Problems to overcome

- Moving picture and dynamic scene

- Estimating the pose of camera

- Tracking the position and orientation of objects

- Handling objects at various scales

- Comprehending and modeling the scene

- All this with limited computing resources

Computer vision methods

- Salient point and blob detection

- Scale-space methods

- Template-matching methods

- Edge and boundary detection

Salient point and blob detection

- Salient point: small area of image with some discriminative characteristics that makes it different from rest of image, allowing to detect and track it

- Blob: a region of image with uniform intensity and with some characteristic shape

- A number of detectors exist to find salient points and blobs in image: in general they all find *local maxima* of some *response function*

Scale-space methods

- Salient point and blob detectors typically operate at specific scale, looking at the image through a fixed-size window

- Examining scaled-down versions of image allows to find responses at multiple scales

- Scale-space methods typically involve smoothing the image with gaussian filters of increasing size, thus generating a scale-space; maxima of response functions are then searched both by location and scale

Template-matching methods

- Based on example images, a representative template of object to find is created

- Template is fitted on various image locations and a distance measure is calculated

- Local minima of the distance measure are candidate locations for the object

- This approach is often used for face detection

- Scale-space can be utilized for scale-invariance

- Salient point, blob, and template detectors can find mainly point-like, rigid objects

Edge and boundary detection

- Finding object contours from image is difficult and time-consuming, but several methods exist

- Gradient-based methods detect areas where image intensity changes abruptly

- Wavelet-based methods detect areas that respond to waves with specific orientation and frequency

- Similarity-based methods detect regions that are statistically similar according to some measure

- Edges and boundaries are harder to localize than salient points, blobs, and template matches

Towards robust object recognition

- Current problem: looking mainly at local features in images, unable to consider the big picture

- Higher-level scene classification is needed

- Let us first review various approaches used for describing objects

Object recognition strategies

- Find salient points from image and check if similar collection of them is found as in sample image

- Find object contour and calculate some invariant properties by its shape

- Find individual parts of a complex object, and check if the found parts are in correct relations to each other according to the object model

- Generate all allowed transformations of an object model and check if such object can be found in scene

- Problem reduces to managing search space, finding a good distance measure, and minimizing the distance

Shimon Ullman: High-Level Vision – Object Recognition and Visual Cognition

Manageable high-level objects

- Complex and hierarchical object models pose challenges:

  - Resulting objects reside in very high-dimensional spaces

  - Search spaces are extremely large and sparse

  - Different objects may not reside in the same mathematical space, which makes comparison difficult

- Similar problems have been approached using category theory, a branch of mathematics, that allows to examine mappings between different subspaces

S. Clark, B. Coecke and M. Sadrzadeh: Diagrammatic Reasoning about Meaning of Sentences

Localization and mapping

- If salient points and blobs can be identified in multiple subsequent frames, backprojection can be used to estimate the 3D location of these landmarks in the scene, and a map of the scene can be created

- Some method is needed to evaluate the estimation error, in order to minimize it

- Using sensors to measure camera movement helps

- Simultaneous Localisation and Mapping (SLAM) is a widely studied method for mobile robot navigation, that will become feasible also in mobile augmented reality when processing power increases

Andrew I. Comport et al.: Real-Time Markerless Tracking for Augmented Reality: The Virtual Visual Servoing Framework

Probabilistic aspects of mapping

- Like stated previously, it is impossible to find a unique solution for object location: we are left with an estimate, with associated estimation error

- Mapping algorithms need to model the probability of object locations, mutual dependences of the probabilities, and the confidence level of the probability estimates

- Various kinds of Bayesian networks are used for modeling, Kalman filters being a typical example

- Problem is, that models are non-linear

Mapping a scene

Cmaptools: http://cmap.ihmc.us/
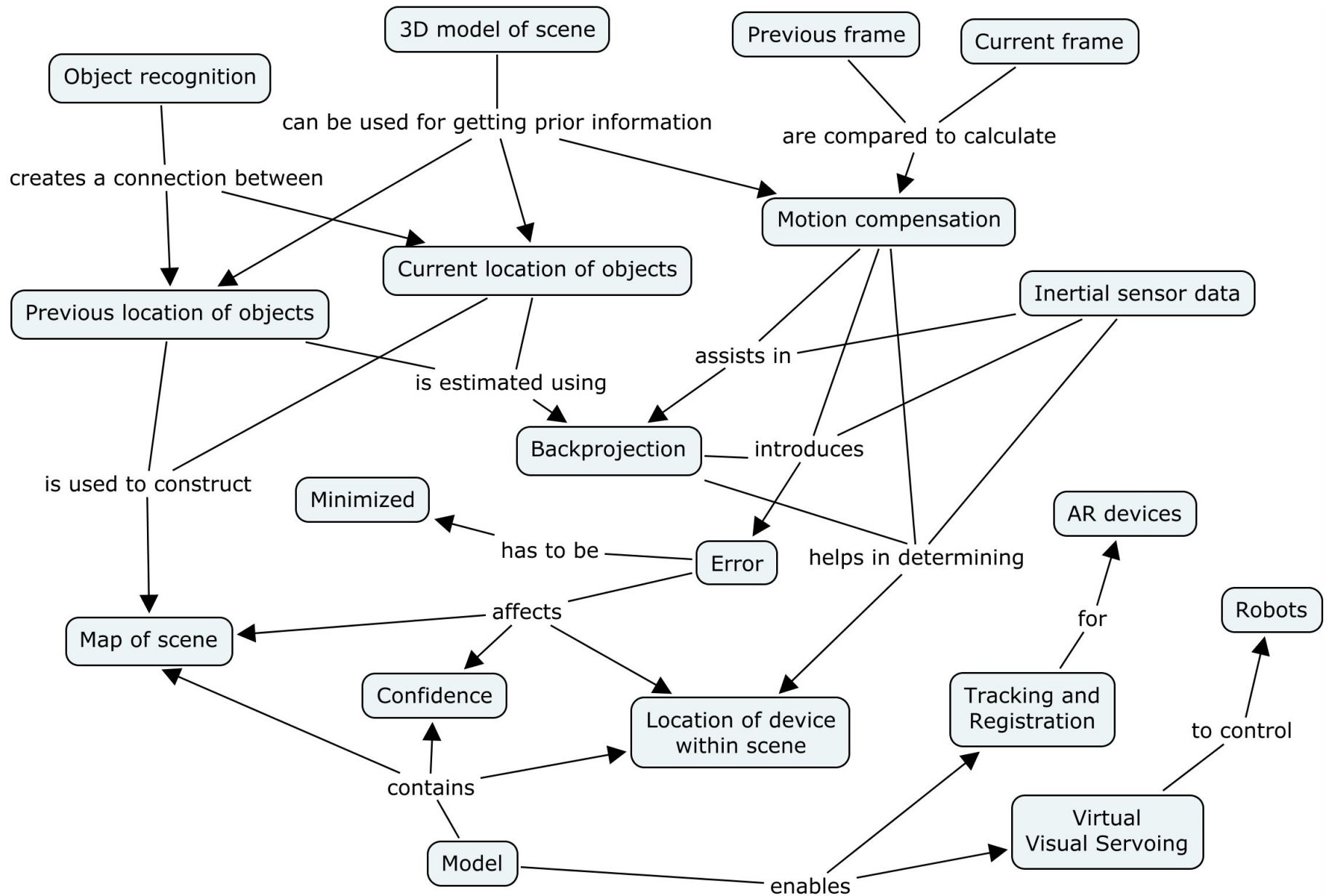
Mapping in augmented reality

- Even a coarse map of the scene, coupled with probabilistic spatial reasoning, allows some level of scene understanding and narrowing down the search space when recognizing objects

  - Location-specific objects – if we are in a specific building and room, we can expect to find only certain objects

  - Context-specific objects – if we are looking at a table, we can expect to see only certain objects

- For visualising sensor data, it is useful if the information can be shown in appropriate location

Sketching a visual AR framework

- Low-level image acquisition using FCam

- Noise reduction with statistical model

- Motion compensation utilizing inertial sensors

- Efficient scale-space generation

- Salient region detection and tracking in scale-space (corners, edges, boundaries, blobs)

- Localisation and mapping of salient regions, creating a representation for scene understanding

- Generic interface for high-level object recognition

Considerations for mobile devices

- Balance frame size and frame rate

- Take advantage of sensor and motion compensation data to concentrate on changed parts of picture

- Imitate biological eye and use full resolution only on small part of each frame – select interesting part of scene

- Prepare to utilise future dual-core processors

- Possibly utilise graphics processors to remove load from main processor

Evaluating the results

- Achieved framerate and frame size

- Accuracy of object detection (objects detected correctly / correct objects detected)

- Size of scene that can be handled

- Confidence levels and accuracy of modeling related to a ground truth (manually generated model with correct distances)

Initial results

- Traditional marker tracking with ARToolkit is feasible on VGA resolution

- Scale-space creation and salient point detection is infeasible on VGA resolution, testing with QVGA (320x240) to reach acceptable framerates

- Traditional SLAM using extended kalman filters allows only limited map sizes, investigating general Bayesian networks and Markov random fields to achieve better performance

Next steps

- Continue to study the key methods in more detail

- Continue to implement the framework

- Prepare the first conference presentation

  - Target to propose my first presentation this spring (subject to progress)

  - For example ISMAR

  - IEEE International Symposium on Mixed and Augmented Reality - http://www.ismar11.org/

Possible subjects for publication

- Registering a mobile camera with a fixed camera or a fixed multi-camera system

  - Utilise a server that analyses the data from fixed cameras in real time

  - Mobile device analyses the scene and asks the server to help in localisation

- Hybrid registration, tracking and mapping using markers, visual features and inertial sensors

  - Solve partial or complete occlusion of marker

  - Find boundaries and occlusion planes in scene

Thank you for your interest!
Any questions?