VISUAL POSE ESTIMATION FOR A MOBILE MANIPULATOR

With Aaron Walsman and Siddhartha Srinivasa
Personal Robotics Lab, Carnegie Mellon University

Shushman Choudhury
IIT Kharagpur

MOBILE MANIPULATOR

- A robotic system built around one or more manipulator arms and a base for locomotion.
- Used mainly for assembly in factories but have been used for domestic tasks.
- Say hello to HERB (Home Exploring Robotic Butler) !

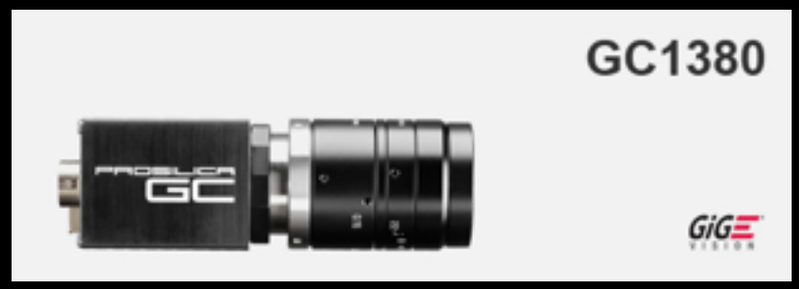
VISION SYSTEM REQUIREMENTS

- Profile of environment (domestic environments may be cluttered, unlike in factories)
- Recognition of various household objects
- Estimation of the pose of objects of interest (for grasping and tracking)
- Considerable accuracy of readings due to close-range interaction

VISION SYSTEM - SENSORS

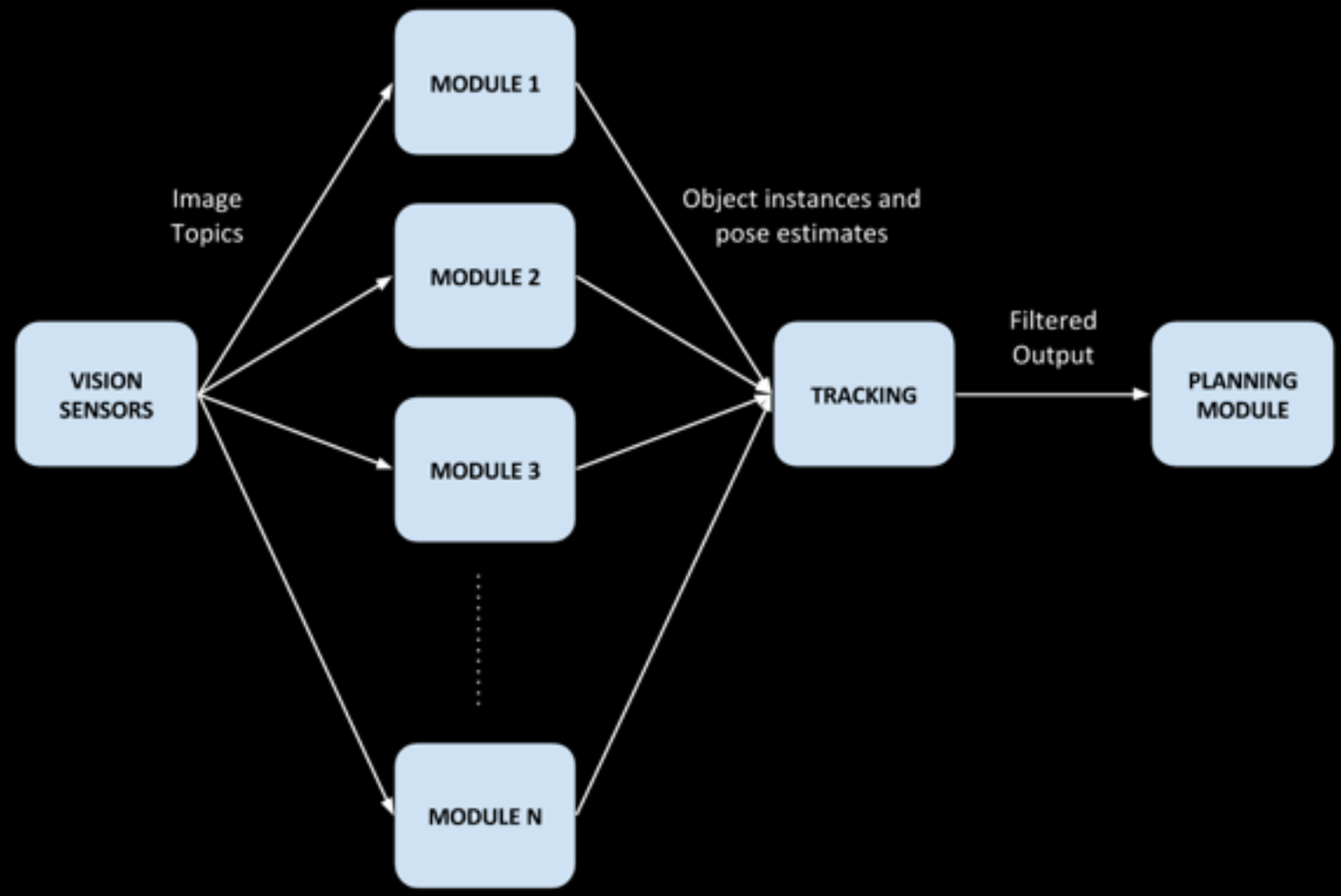
- Asus Xtion for depth information

- Prosilica hi-def RGB camera for texture information



VISION SYSTEM - ARCHITECTURE

- Based on ROS
- Tracking is slightly different for each module, but the workflow is similar
- Some of the modules on HERB - AprilTags, MOPED, ROCK

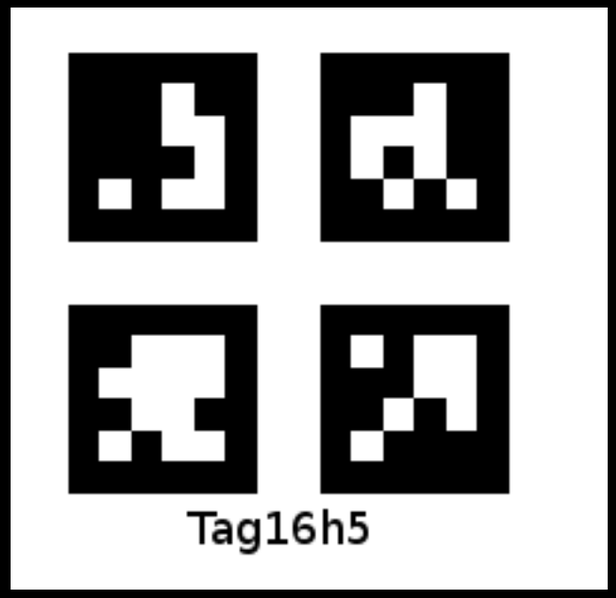


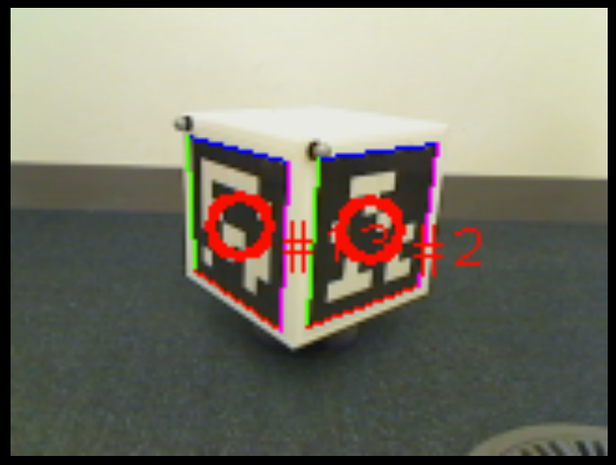
POSE (VISION)

- Combined term for position and orientation of an object.
- Represented as $P_{4*4} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}$ where R_{3*3} is the rotation matrix and T_{3*1} is the translation vector, with respect to the appropriate co-ordinate frame.
- Also represented as $P = \begin{bmatrix} t_x & t_y & t_z & q_x & q_y & q_z & q_w \end{bmatrix}$ where the first three elements represent T'_{3*1} and the next 4 elements represent the rotation in quaternion format (axis of rotation and amount of rotation about that axis.

APRILTAGS

- Visual fiducial system
- Fast detection and recognition from a distance - robust to lighting and angle
- Accurate pose estimation
- Stuck on objects for quick testing and usage, as a benchmark and just for fun!





E. Olson, "Apriltag: A robust and flexible visual fiducial system," in Proc. IEEE ICRA, Shanghai, China, 2011.

PROBLEMS WITH APRILTAGS DETECTIONS

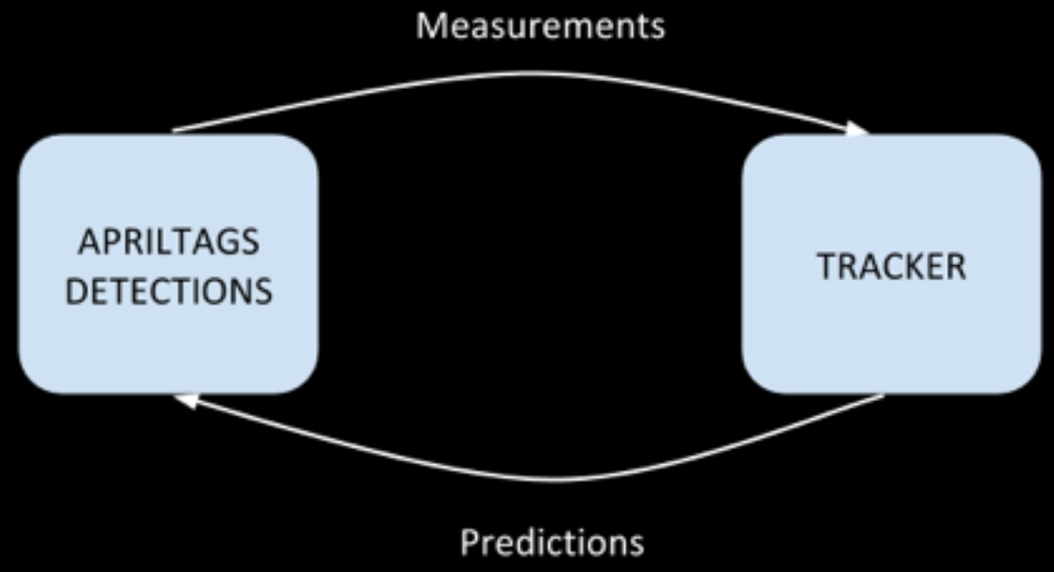
- Loss of frames - AprilTags flicker a lot, particularly at distances. This may be a problem for an active grasping application.

- There is considerable uncertainty in the position of a tag.

- Two ambiguous poses returned when not askew enough, due to multiple possible solutions for the PnP solver.
  (A HUGE problem at times)

CLOSING THE LOOP - TRACKING APRILTAGS

- Maintain the state of the AprilTags to survive frame loss
- Tolerate errors in measurement and ambiguous poses
- Method used - Extended Kalman Filter for pose, with gating



KALMAN FILTER FOR DUMMIES

- A method that continuously processes a noisy input stream and returns the best estimate of the next state of a system.

- Primarily two logical cycles - prediction (before next measurement) and correction (after next measurement).

- The essence is captured in the equation

$$\hat{X_k} = K_k.Z_k + (1 - K_k).\hat{X_{k-1}}$$

- Here, at time step k, $\hat{X}_k$ is the estimated state after that iteration, $\hat{Z}_k$ is the measurement at that time, and K_k a quantity known as the Kalman Gain, which assigns importance to the measurement or prediction accordingly.

R.E. Kalman (1960). "A New Approach to Linear Filtering and Prediction Problems". Journal of Basic Engineering: 35–45.

EXTENDED KALMAN FILTER FOR TRACKING POSE

- EKF is a non-linear version of the normal KF which linearizes about the current estimate of mean and covariance.
- State of the tag is a vector comprising the pose in translation-quaternion format, the linear and angular velocities. The next estimate is non-linearly related to the current state, hence the EKF.
- Measurement is the current pose observation for the tag.
- Benefits - Resistance to loss of a few frames; can tolerate uncertainty and noise in measurements, and also ambiguous poses (next slide)

Goddard, J. 1997. Pose and Motion Estimation From Vision Using Dual Quaternion-Based Extended Kalman Filtering. PhD thesis,Knoxville

RESOLVING AMBIGUOUS POSES - GATING

- A measurement validation gate was used to disallow the degradation of the state estimate by an erroneous measurement.

- This requires the computation of a value called the normalized innovation squared,

$$\epsilon_v(k) = v(k)^T S(k)^{-1} v(k)$$

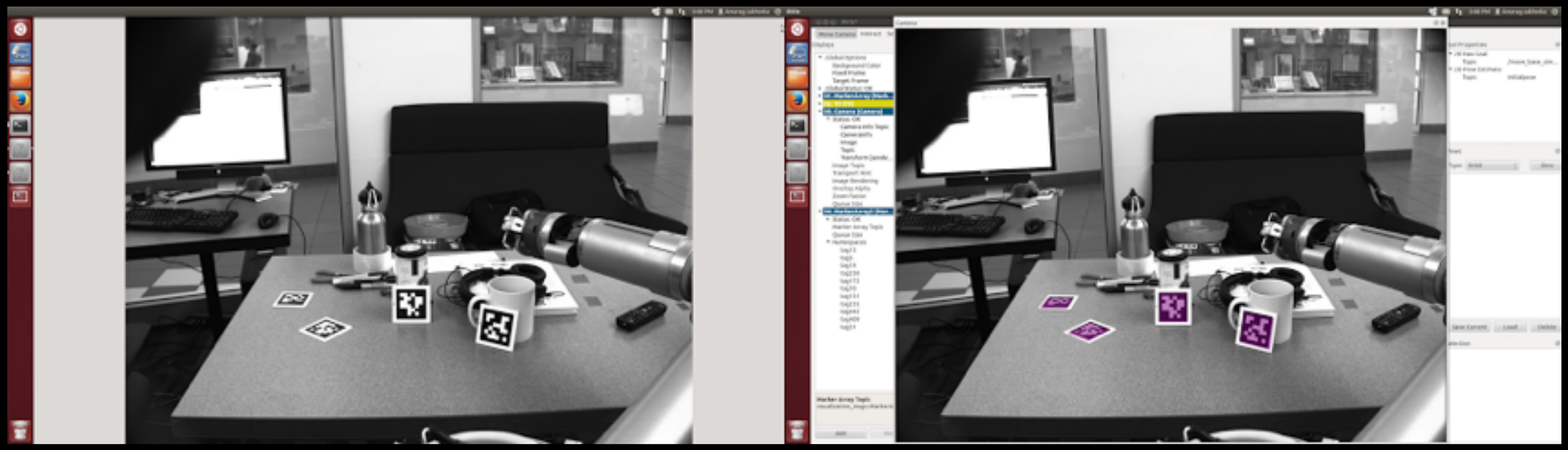
where $v(k)$ is the difference between the estimated state and the observation and $S(k)$ is a procedural matrix.

- If this value is out of (user-defined) bounds, then the measurement is rejected. We made further use of the idea that ambiguous poses would re-project to similar areas in the camera frame.

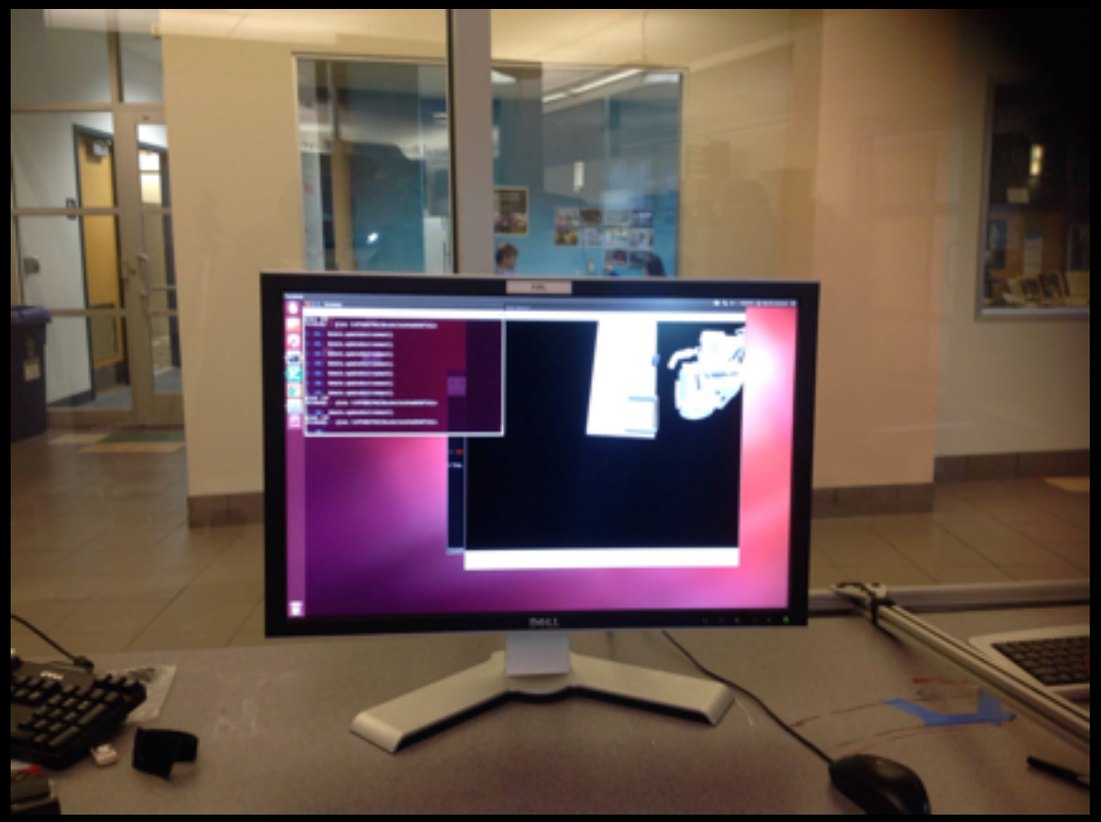
PRELIMINARY OUTCOME

- Module to track tags based on EK, remember them for some frames in case of data los and forget them if they are lost for too long. This reduced the average error in detections and improved the usage of the tags

- The module was used as part of a 3D movie about robots that National Geographic sh in our lab. HERB's task was to clear up a dinner table. We stuck tags on objects to help him localize them in world space.



SCIENCE AT WORK!



MISCELLANEOUS TASKS
AUXILIARY

- Extrinsic Calibration of the vision sensors (transformations)
- Fixing some bugs in the intrinsic parameters of the vision sensors
- Organizing the namespaces of the perception module
- Object pose detection using a bundle of AprilTags, and prior knowledge about tag placement on the objects (for the demonstration)
- Organizing the vision system - ROS nodes will exist but remain dormant until their topics are subscribed to.

SOMETHING MORE...

- As a domestic assistant, HERB will be required to recognize and estimate the pose of a variety of household objects.

- Clearly, AprilTags, which only give the local pose about the tag, will not be enough.

- Prior knowledge about domestic objects and their appearance needs to be leveraged, given the high demand for accuracy in close range.

ROCK
Robust Object Category and Kinematic pose

Presentation due to Aaron Walsman, Carnegie Mellon University

PROBLEM FORMULATION

Visual What and Where

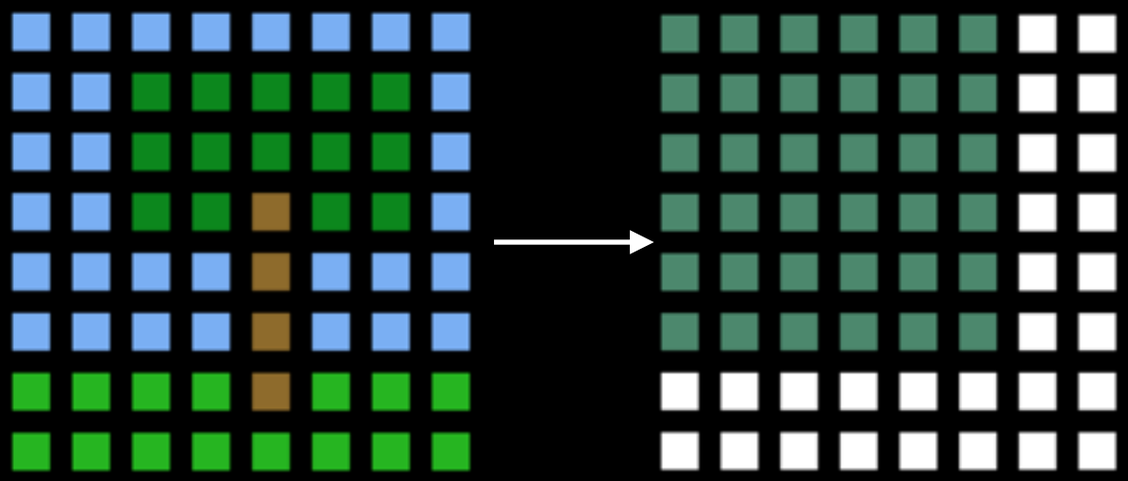
PROPERTIES TO INCORPORATE

- Lightweight model
- High Discriminability
- Invariance to viewpoint and lighting
- Usage of strong pose priors
- Scale to several categories easily
- Adjustable computation time/accuracy ratio

RELATED CONCEPTS

INTEGRAL IMAGE

- A method for faster, more efficient image processing
- Allows quick and effective summation of pixel values for a subset of the image

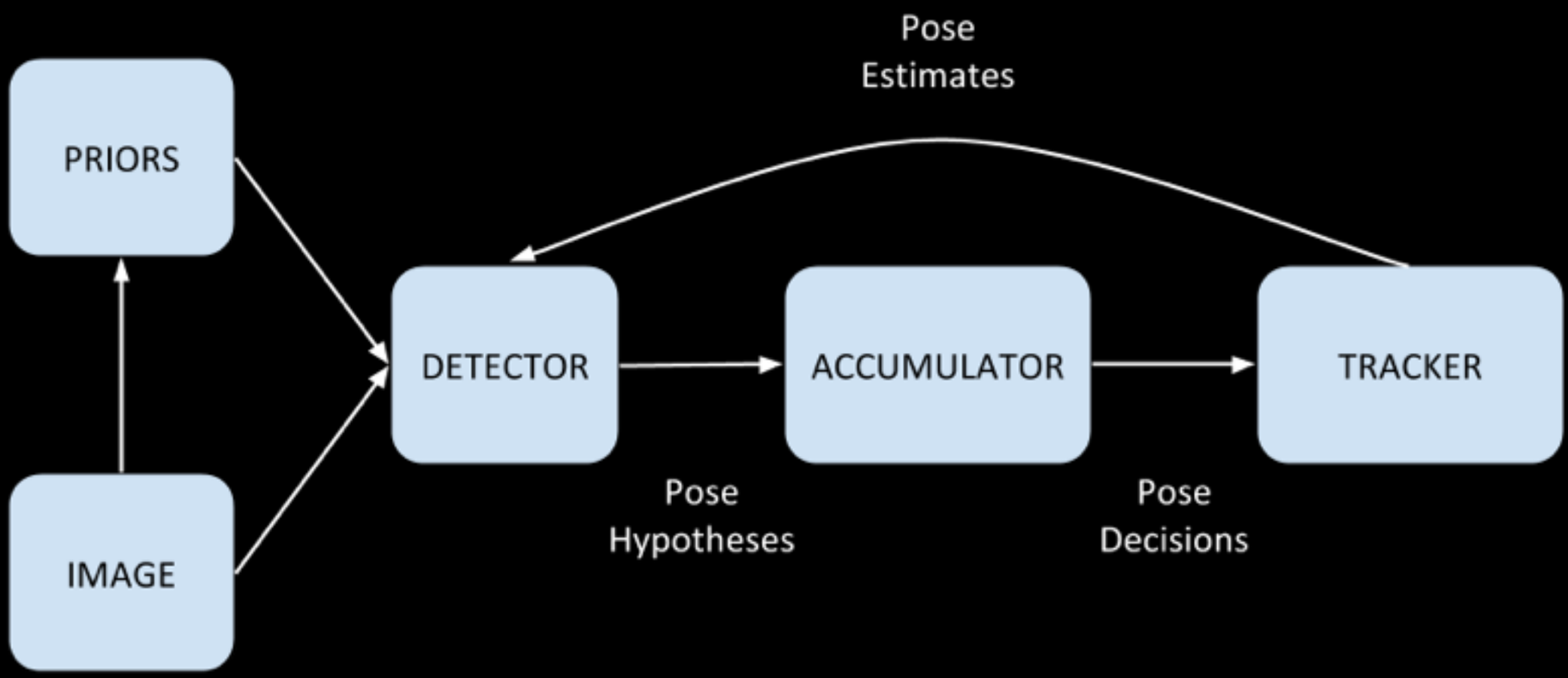


$$ii(x, y) = \sum_{x'<=x, y'<=y} i(x', y')$$

TEMPLATE MATCHING

- A way to search for a patch in a larger image
- A sliding window technique adapted to search over the image space for possible occurrences

P Viola, M Jones. Rapid object detection using a boosted cascade of simple features - Computer Vision and Pattern Recognition, 2001



ROCK WORKFLOW

A REPRESENTATIVE DIAGRAM

OBJECT MODELS

- A database of models of the common objects is used by ROCK, both for recognition and pose estimation.
- Several dozens of models can be easily stored, leading to scalability.
- Construction of the models is a simple but important process.

OBJECT MODELS (contd.)

AUTODESK 123DCATCH
USED TO GENERATE
3D MODELS

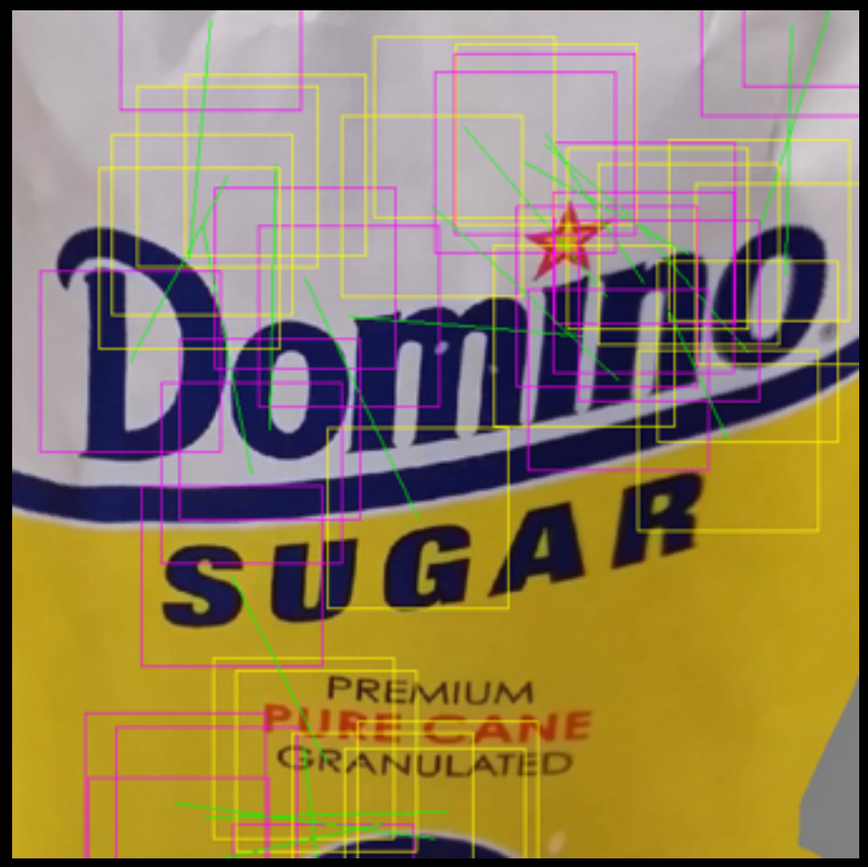
OBJECT MODELS (contd.)

A mesh of vertices, edges and faces generated

OBJECT MODELS (contd.)

A view of the texture mesh for an object

OBJECT MODELS (contd.)

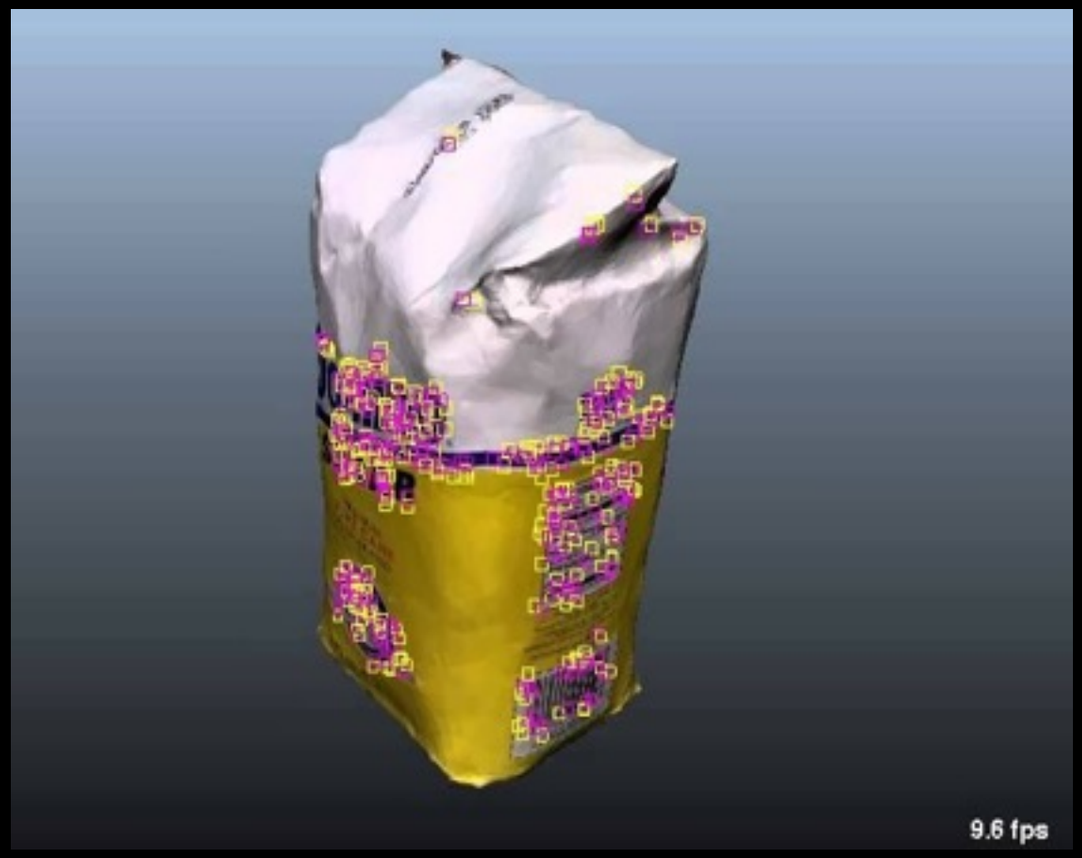


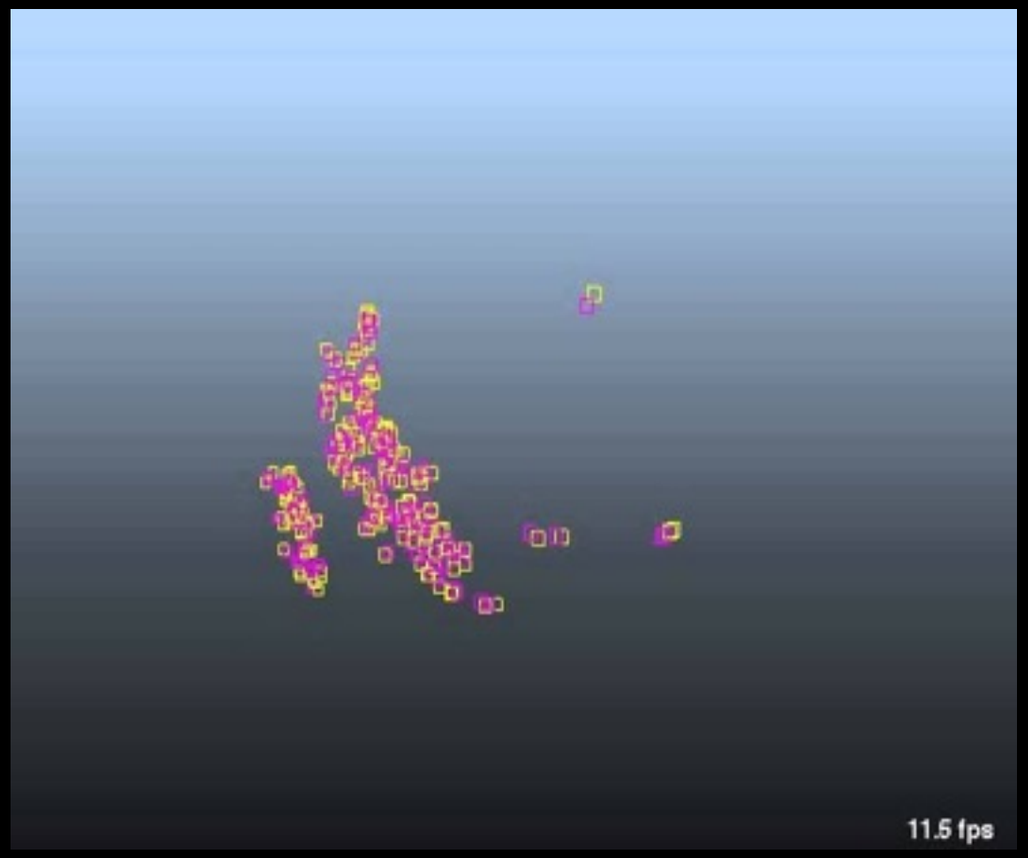
Information of the patches of the object is stored
They will be used later for comparison

OBJECT MODELS (contd.)

There are patches of various sizes
This is used to represent coarse to fine layers

OBJECT MODELS (contd.)





The resulting model with patches
Point features (ORB) of the model are also stored
but not shown here

PRIORS

- Certain aspects of the test image, or any extra knowledge we have, helps generate pose hypotheses to test out at the detection stage.

- Given point features obtained from the test image, and with the nature of the corresponding feature in the model, it is possible to obtain a hypothesis for the pose of the object (with some wacky math we won't get into here)

- Other extra prior information in less general cases may be used (knowledge that it is somewhere on a fixed tabletop, for instance)

DETECTIONS

Given a set of pose hypotheses, we need to test each of them to determine how accurate that pose is. We do this by projecting the sample pattern for the pose onto the image and testing the expectation.

correct: 13.27% ✖

correct 64.29% ✔

DETECTIONS
(PATCH INFORMATION)

- Big idea - The difference between corresponding patches (in LAB colour space) is fairly invariant to illumination and rotation.

- Earlier, during model building, patch processing had been referenced. For a patch P defined by a bounding box, it's value is obtained by averaging the L,A,B values of the patch thus:

$$P(l) = avg\left(\sum_{i=y1}^{y2}\sum_{j=x1}^{x2} L(i,j)\right) \quad P(a) = avg\left(\sum_{i=y1}^{y2}\sum_{j=x1}^{x2} A(i,j)\right) \quad P(b) = avg\left(\sum_{i=y1}^{y2}\sum_{j=x1}^{x2} B(i,j)\right)$$

$$P \equiv \{P(l), P(a), P(b)\}$$

- The difference between pairs of nearby patches on the model is stored as a vector

DETECTIONS
(PATCH DIFFERENCES)

$$d(P_1, P_2) = \begin{bmatrix} P_1(l) > P_2(l) \\ P_1(l) - P_2(l) > \delta_l \\ P_1(a) > P_2(a) \\ P_1(a) - P_2(a) > \delta_a \\ P_1(b) > P_2(b) \\ P_1(b) - P_2(b) > \delta_b \end{bmatrix}$$

The above is the nature of the vector that represents the relationship between a given pair of patches. The difference vectors for every considered pair, is stored apriori for the model. The thresholds are set by experimentation

DETECTIONS
(CALCULATING THE SCORE)

- For each pose hypothesis, the model is overlaid on the test image and the differences between those patches are computed, and then compared with the corresponding difference vector of the model.

- The response (strength) of a pose hypothesis is therefore

$$R_H = \sum_i sim(M_{d_i}, H_{d_i})$$

where M_{d_i} refers to a difference vector for a certain pair of patches in the model and where H_{d_i} is the corresponding difference in the image for the pose hypothesis to be tested.

- A variety of similarity metrics are now being explored for the response, so as to account for noise, randomness and occlusion.
For instance, a stricter check mechanism reduces the chance of false positives but also increases the effect of noise.

DETECTIONS

Start with coarse samples and refine.

DETECTIONS

Start with coarse samples and refine.

DETECTIONS

Start with coarse samples and refine.

ACCUMULATION

- The detection stage gives us a set of pose hypotheses for a particular object, along with those responses.

- We need to accumulate this information and obtain the best estimate of the number of instances, and the pose of each.

- Current idea - Use coarser layer (less discriminatory; computed faster) to reject obvious true negatives; sample better poses at finer layers to get clearer results, and then cluster those results to make a decision.

MEAN-SHIFT CLUSTERING

- Clustering a set of *n* data points, with an unknown prior number of clusters, based on some similarity metric (Generalization of K-means)

- A number of initial centres which are gradually refined as

$$x^{t+1} = \frac{\sum_{i=1}^{n} g(\|x - x_i\|^2/h^2)x_i}{\sum_{i=1}^{n} g(\|x - x_i\|^2/h^2)} x^t$$

where $g(\|x\|)$ represents a Gaussian kernel (response falls off exponentially with distance), and h is the bandwidth that determines the allowable size of a cluster. x_i of course refers to the data points.

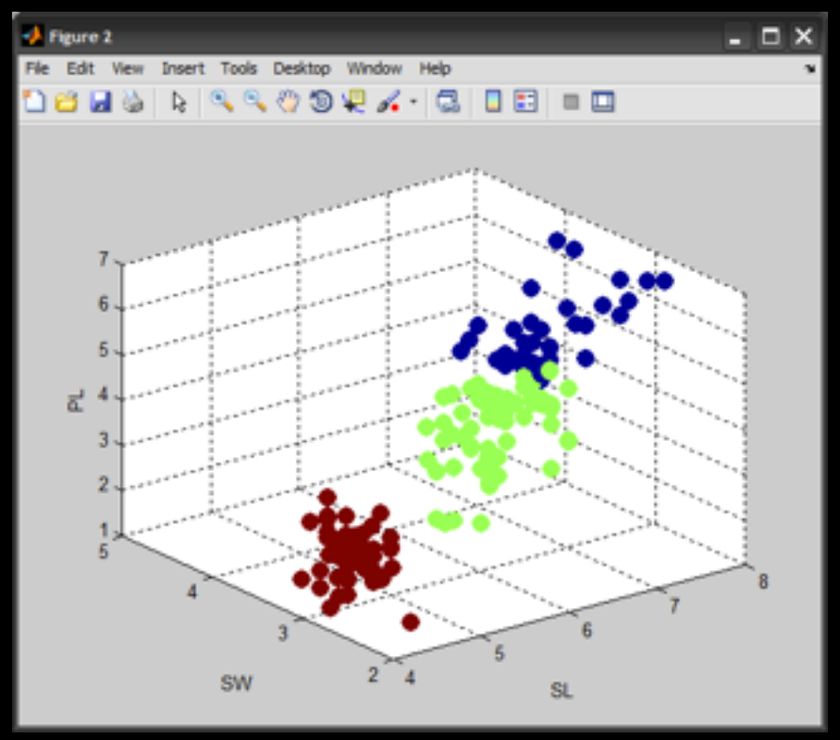
D Comaniciu, P Meer. Mean shift: A robust approach toward feature space analysis Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (5), 603-619

ACCUMULATION
CLUSTERING - 3D

- The first level is to cluster a set of pose hypotheses based only on their 3D position in world space.

- This will disambiguate between obviously different instances.

- The bandwidth for this clustering will depend on the real-world dimensions of the object, which are known.



ACCUMULATION
CLUSTERING - 6D

- For each cluster, there is either a single instance, or multiple instances close to each other.

- The space of poses is not a vector space but rather SE(3), a Special Euclidean space. Mean-shift cannot be used directly with it.

- SE(3) has a closely associated vector space called se(3). (Courtesy Lie Algebra). There are well-defined operators to go from one space to the other. Conversion to this vector space allows mean-shift to be applied.

- The clustering for each centre gives us the best average response, and also if there are two instances close to each other.

R. Subbarao, Y. Genc, and P. Meer, "Nonlinear mean shift for robust pose estimation," in IEEE Workshop on Applications of Computer Vision, February 2007, pp. 6–6.

FUTURE WORK

- Formalizing the mathematics of the various phases
- Converging upon a metric for patch comparisons, on a method for sampling poses around a promising one, and for using pose estimates in one iteration for feedback in the next frame.
- Connecting the components together and parallelizing where possible
- AprilTags - Better pose correction and tracking via fusion of textured image with depth information

THANK YOU!

QUESTIONS?