

Metrics and Evaluation

January 28, 2003

Class Meeting 5

William Thompson, Lord Kelvin (1821-1907):

“When you can measure what you are speaking about, and express it in numbers, you know something about it;


But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind;

It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.”

Announcements

- Paper presentation schedule finalized. Available on course website.
- First presentation: This Thursday (1/30).
- Questions regarding Motor Schemas or material from cancelled class?

Objectives

- “Show and Tell” – Roomba 
- Understand **metrics** for multi-robot systems:
 - General issues
 - Hierarchic Social Entropy
- Understand **evaluation techniques** for multi-robot systems
 - Theoretical proofs
 - Empirical studies
- Assignment #2 overview

Metrics

- **What** are metrics?
 - Standards of measurement
 - Performance indicators
- **Why** do we need metrics?
 - To evaluate, compare, and contrast various approaches to multi-robot control
- **Two primary types** of multi-robot system metrics:
 - Task dependent
 - Task independent

Measuring Performance vs. Measuring Intelligence

- One definition of measurement of intelligence:
 - “Intelligence is measured in terms of the diversity of purposes that can be achieved under a range of environments.”
- Tests for measuring intelligent systems:
 - IQ tests
 - Turing test
- In general, very hard to measure intelligence
 - We'll instead measure performance

Task-Dependent Metrics

- **Task-Dependent Metrics:**
 - Specific to the mission being solved
- **Relatively easy to define task-dependent metrics:**
 - Time of task completion
 - Total energy used
 - Team formation error
 - Average number of targets under observation
 - Cumulative distance traveled
 - Etc.

Typically, Task-Dependent Metrics Are Directly Related to Goal of Multi-Robot Mission

<u>Example Mission</u>	<u>Useful metric</u>
Formation-keeping	Cumulative formation error
Foraging	Rate of puck collection; Percentage of pucks collected
Search	Rate of area covered; Percentage of area covered
Target-tracking	Percentage of targets tracked

Metrics Reflecting Robot Team Scalability Issues

- Collect mission-specific data in terms of scalability issues:
 - Function of robot team size
 - Degree of inter-robot interference
 - Number of communications; required communication bandwidth

<u>Example Mission</u>	<u>So, rather than this metric:</u>	<u>Use this metric:</u>
Formation-keeping	<ul style="list-style-type: none"> • Cumulative formation error 	<ul style="list-style-type: none"> • Cumulative formation error as a function of robot team size • Required communications bandwidth as a function of robot team size
Foraging	<ul style="list-style-type: none"> • Rate of puck collection; • Percentage of pucks collected 	<ul style="list-style-type: none"> • Rate of puck collection as team size increases; • Percentage of pucks collected as team size increases • Number of robot collisions (i.e., interference) as team size increases
Etc.	<ul style="list-style-type: none"> • Etc. 	<ul style="list-style-type: none"> • Etc.

Thought Exercise – Task-Dependent Metric Specifications

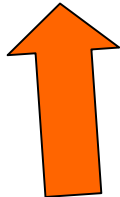
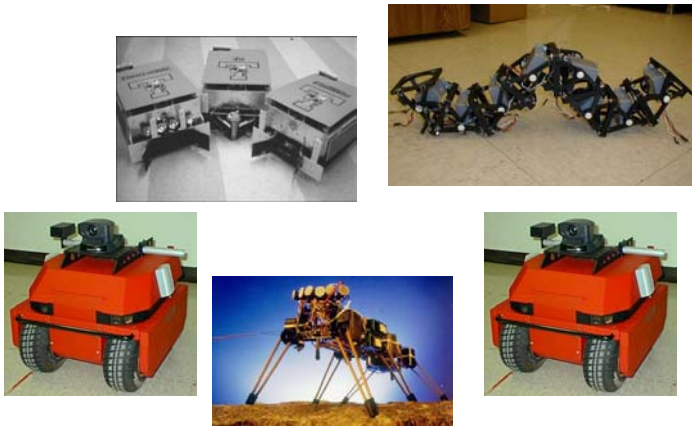
- What metric(s) could be used to evaluate the following multi-robot team behaviors?
 - Flocking
 - Herding
 - Path planning
 - Soccer
 - Box pushing
 - Traffic control
 - Reconfigurable robots

Task-Independent Metrics

- Much harder to define **task-independent** metrics
- Example **task-independent** metrics:
 - Team diversity
 - Team coherence
 - Robustness
 - Fault tolerance
 - Adaptivity
 - Reliability
 - Flexibility
- **Challenge:** How to quantify these values?

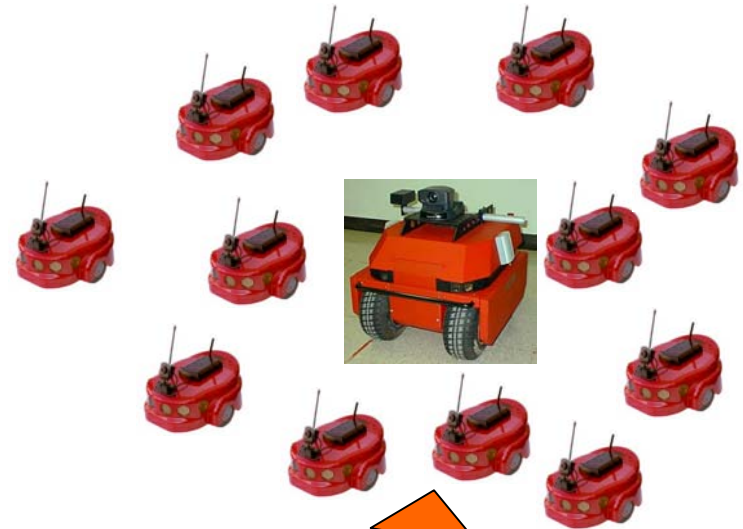
An Example Task-Independent Metric: Team Diversity (Balch)

- Prior to Balch's work, robot teams were either:
 - a. Homogeneous
 - b. Heterogeneous
- But, we want a more specific measurement of heterogeneity



This robot team

seems very different



from this team.

Why Measure Diversity?

- Can study impact of diversity on performance
- Can study impact of task factors on diversity
- Focus in Balch's metric:
 - Behavior diversity in mechanically similar agents
 - Distribution of agents between homogeneous subsets is at core of the meaning of diversity
 - Idea: Societies with members equally distributed among subsets are the most diverse

Simple Social Entropy

- Based on work of Shannon (1949):
 - Sought to quantify the randomness of an information source
 - Developed information entropy measure
 - Diversity and information are closely related concepts
- Formulation for multi-robot teams:
 - \mathbf{R} is a society of N agents, with $\mathbf{R} = \{r_1, r_2, r_3, \dots, r_N\}$.
 - \mathbf{C} is a classification of \mathbf{R} into M possibly overlapping subsets
 - c_i is an individual subset of \mathbf{C} with $\mathbf{C} = \{c_1, c_2, c_3, \dots, c_M\}$.
 - $p_i = \frac{|c_i|}{\sum_{j=1}^M |c_j|}$ is the proportion of agents in the i^{th} subset; $\sum p_i = 1$
 - Simple social entropy of agent society given by: $H(R_a) = \sum_{i=1}^M p_i \log_2(p_i)$

(Examples, page 100)

Limitations to Simple Social Entropy

- Loss of information when diversity is summarized in a single number
- There are potentially an infinite number of societies matching any particular value of diversity
- Single number doesn't tell:
 - How many classes of agents there are
 - How many agents in each class
- Lack of sensitivity to degree of difference between agents

Classification and Clustering

- Differences between agents likely various along a continuous spectrum, rather than in a binary fashion
- Numerical taxonomy tools:
 - Goal: build a taxonomic tree (dendrogram) that reflects spatial distribution of elements in system
 - Strong similarities between elements are grouped near bottom
 - Clustering algorithm is parameterized by h , the maximum distance allowed between elements in the same subset
 - $D(a,b)$ = difference between elements a and b
- Hierarchic classifications:
 - Any member of lower-ranking taxon is also a member of a higher-ranked taxon

(Example, page 104)

Clustering Algorithm for Agent Society

- Initialize N clusters with $c_i = \{r_i\}$
- For each cluster c_i : For each r_j (except r_i) in \mathbf{R} :
 - if $(D(r_j, r_k) \leq h)$ for every r_k already in c_i ,
 - add element r_j to cluster c_i
- Discard redundant clusters

(Example, page 107)

Hierarchical Social Entropy

- Augmented notation:

$H(\mathbf{R},h) = H(\mathbf{R})$ for the clustering of \mathbf{R} at taxonomic level h

- Comprehensive evaluation of diversity should account for clustering at all taxonomic levels
- Hierarchic social entropy:

$$S(R) = \int_0^{\infty} H(R, h) dh$$

(Examples, pages 108-109)

Why Hierarchic Social Entropy is Useful Metric

- Can account for continuous differences in agent society
- Value = 0 when all elements are identical
- Enables total ordering of societies on basis of diversity
- Can distinguish differences between societies regardless of scale

Behavioral Difference

- How do you compare behaviors across agents?
- Fundamental idea: Look for differences in agents' behavioral coding
- Given: perceptual features and set of behaviors:
 - Two agents are absolutely behaviorally equivalent if and only if they select the same behavior in every perceptual state
- Behavioral difference between two robots r_a and r_b is:

$$D(r_a, r_b) = \int \frac{(p_a^i + p_b^i)}{2} |\pi_a(i) - \pi_b(i)| di$$

Examples

- Multi-robot foraging (pages 116 and following)
- Multi-robot soccer (pages 122 and following)

Key Aspects of Balch's Diversity Metric

- Simple social entropy
- Behavioral difference
- Hierarchic social entropy