Australia's National Science Agency

Inferring Temporal Compositions of Actions Using Probabilistic Automata

Rodrigo Santa Cruz[1,2], Anoop Cherian[3], Basura Fernando[4], Dylan Campbell[2], and Stephen Gould[2]

[1]The Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[2]Australian Centre for Robotic Vision (ACRV), Australian National University, Canberra, Australia
[3]Mitsubishi Electric Research Labs (MERL), Cambridge, MA
[4]A*AI, A*STAR Singapore

Compositionality in Computer Vision - CVPR20 - June 15th 2020
rodrigo.santacruz@csiro.au
www.rfsantacruz.com

Compositional Action Recognition

The task of recognizing complex activities expressed as **temporally-ordered compositions** of simple and atomic actions in videos.

Corner kick → **Ball traveling** → **Goal**

Problem Formulation

$a^+ \triangleq a \succ a^\star$
One-or-more repetition

Action Patterns

Describe complex activities by **regular expressions of subset of primitive actions**:

Primitives
$\mathcal{A} = \{a_i\}_{i=1}^{M}$

Alphabet
$\Sigma = \{w \in \mathcal{P}(\mathcal{A})\}$

Operators
$\mathcal{O} = \{\succ, |, \star\}$

Sequential
Alternative
Recursive

Ex: “driving (a_d) and talking on the phone (a_{tc}) or to someone (a_{ts}) repeatedly just after he got in the car (a_{gc})”

$$a_{gc} \succ (\{a_d, a_{tc}\}|\{a_d, a_{ts}\})^+$$

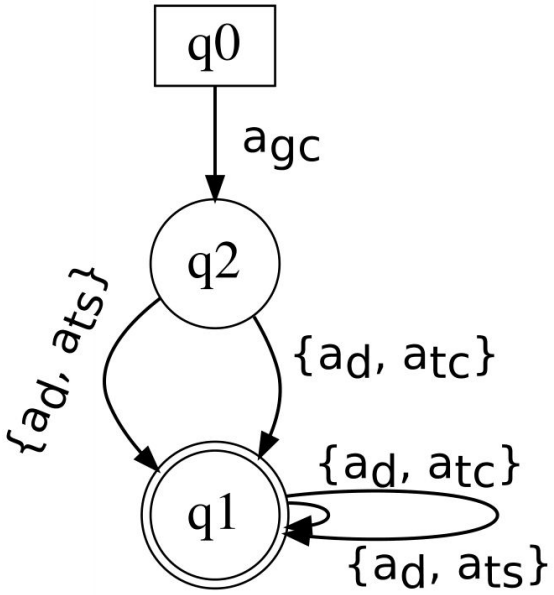
Then, our goal is to model a function **f** that assigns high values to a video **v** if it depicts the action pattern described by the regular expression **r** and low values otherwise.

$$f_r : \mathcal{V} \rightarrow [0, 1]$$

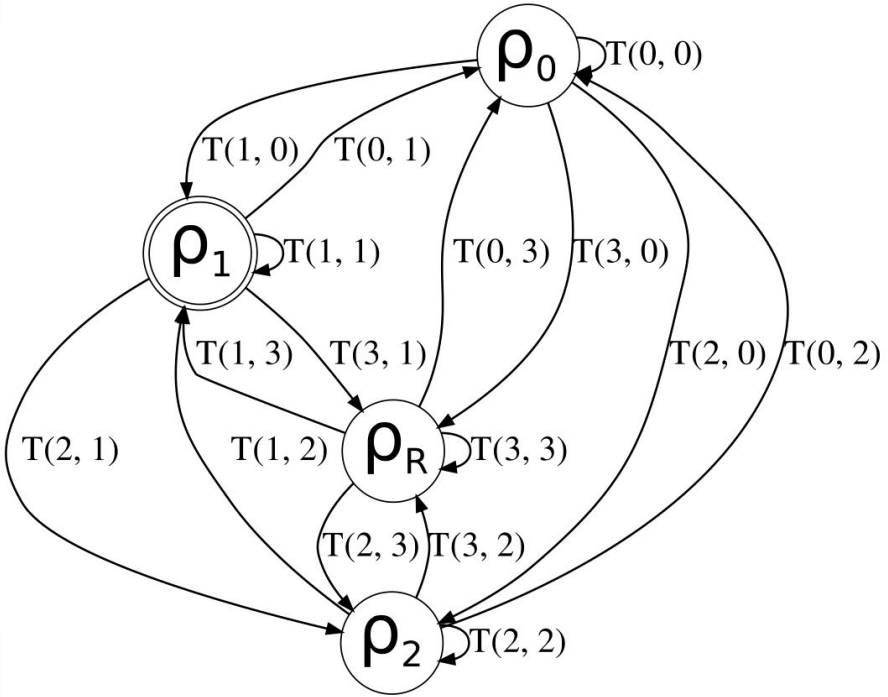
CSIRO

Proposed Models

➔ **Deterministic Inference (DFA based)**

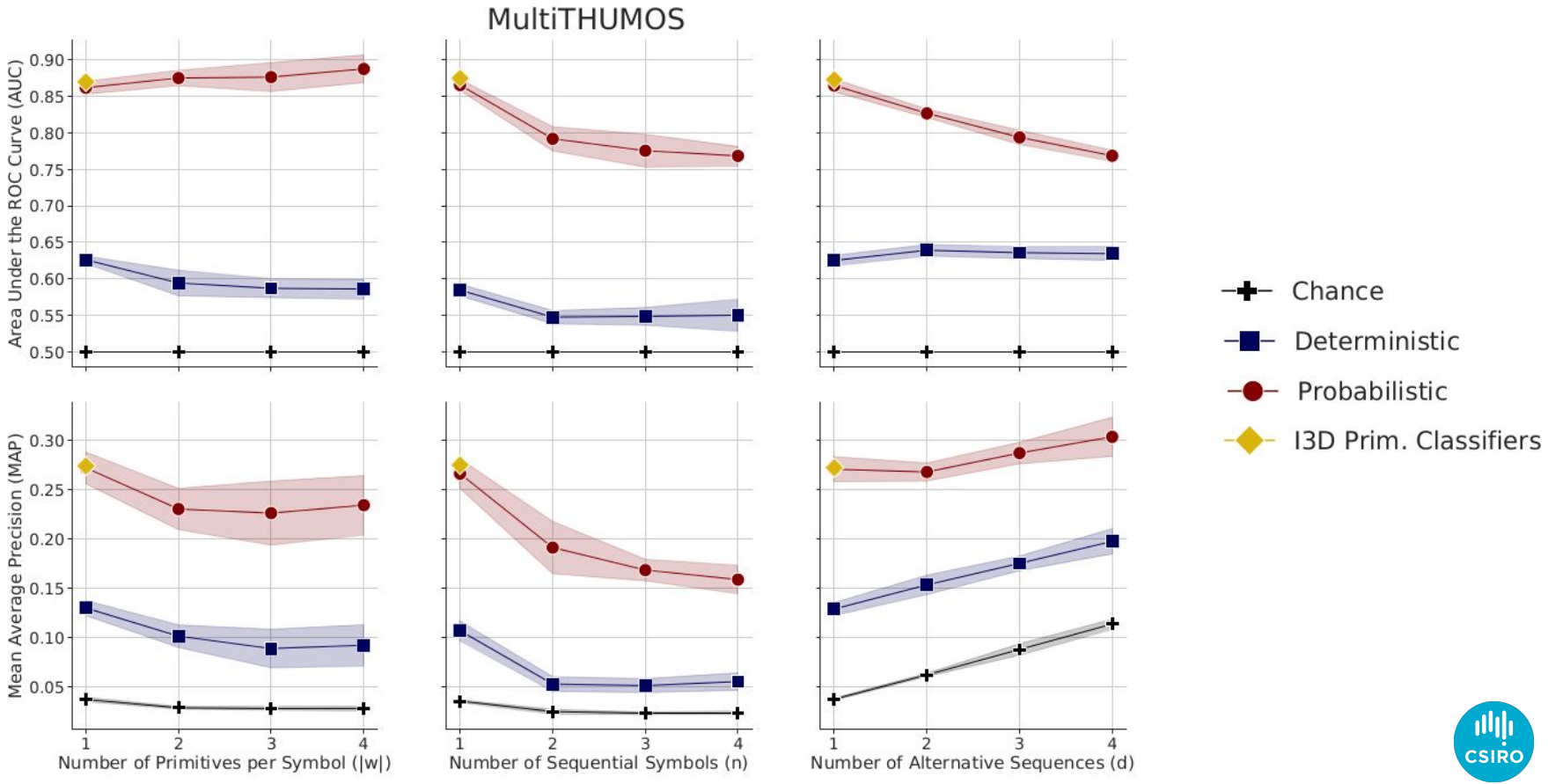


➔ **Probabilistic Inference (PA based)**

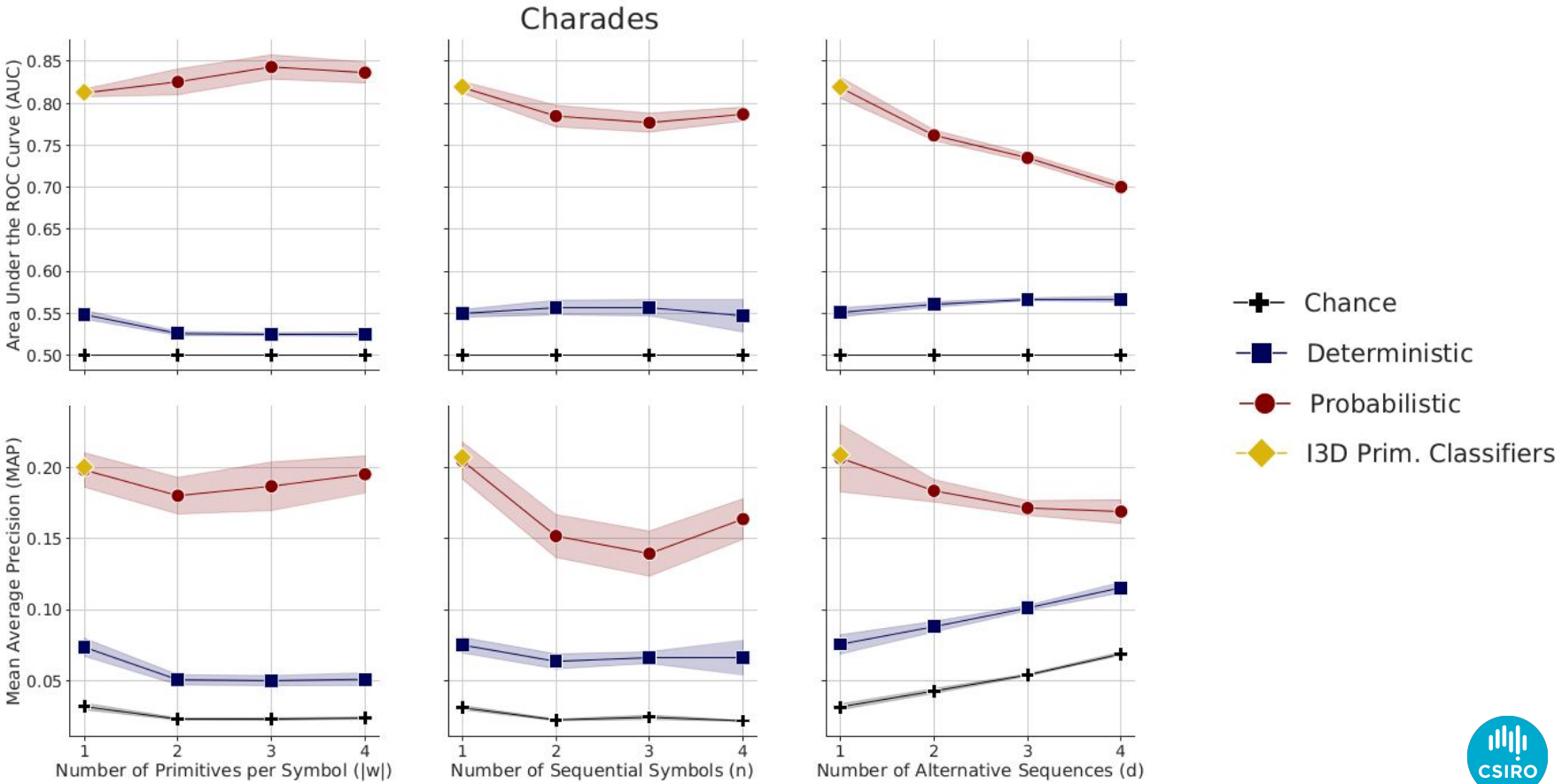


where $T(i,j) = \sum_{w \in \Sigma} T_{ij}(w) p(w|x)$

Experiments - Activity Recognition - MultiTHUMOS



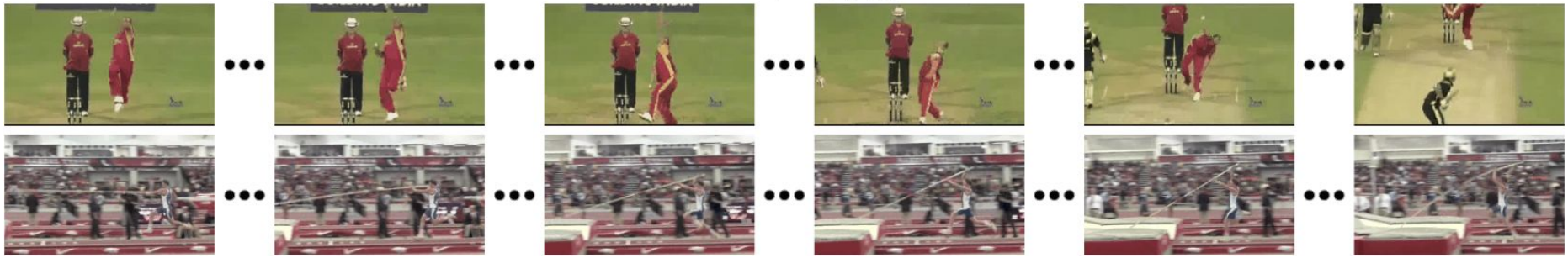
Experiments - Activity Recognition - Charades



Experiments - Qualitative Results

$$\{hg,\ pg\}^+ \succ \{hg,\ dg\}^+$$

$$\{r\}^+ \succ \{cb \mid pp\}^+$$



Primitives: holding a glass (**hg**), pouring water into the glass (**pg**), drinking from the glass (**dg**), running (**r**), cricket bowling (**cb**), and pole vault planting (**pp**).

Australia's National Science Agency

Inferring Temporal Compositions of Actions Using Probabilistic Automata

Rodrigo Santa Cruz[1,2], Anoop Cherian[3], Basura Fernando[4], Dylan Campbell[2], and Stephen Gould[2]

[1]The Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[2]Australian Centre for Robotic Vision (ACRV), Australian National University, Canberra, Australia
[3]Mitsubishi Electric Research Labs (MERL), Cambridge, MA
[4]A*AI, A*STAR Singapore

Compositionality in Computer Vision - CVPR20 - June 15th 2020
rodrigo.santacruz@csiro.au
www.rfsantacruz.com