

Cognitive developmental approach towards the realization of human-like visual scene understanding: Framework and core technologies

Akisato Kimura⁽¹⁾, Kunio Kashino⁽¹⁾, Ken Fukuchi⁽²⁾, Kazuma Akamine⁽²⁾, Kouji Miyazato⁽²⁾, Shigeru Takagi⁽²⁾

(1) NTT Communication Science Laboratories, NTT Corporation, Japan, (2) Okinawa National College of Technology, Japan

Contact: Akisato Kimura <akisato@ieee.org>

Ultimate goal Realize human-like visual scene understanding

- Can detect, recognize and retrieve registered objects from cluttered visual scenes
- Can detect unregistered “objects”, find them “unregistered”, and request their “information”

How we humans understand visual scenes ?

Babies naturally acquire the ability to do it.

Focusing salient objects

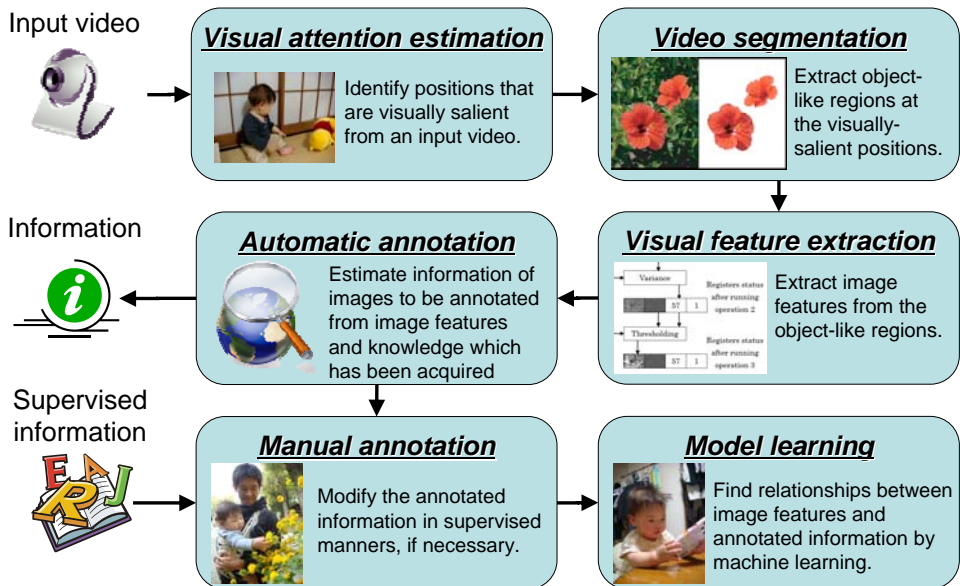


Modeling objects from appearance

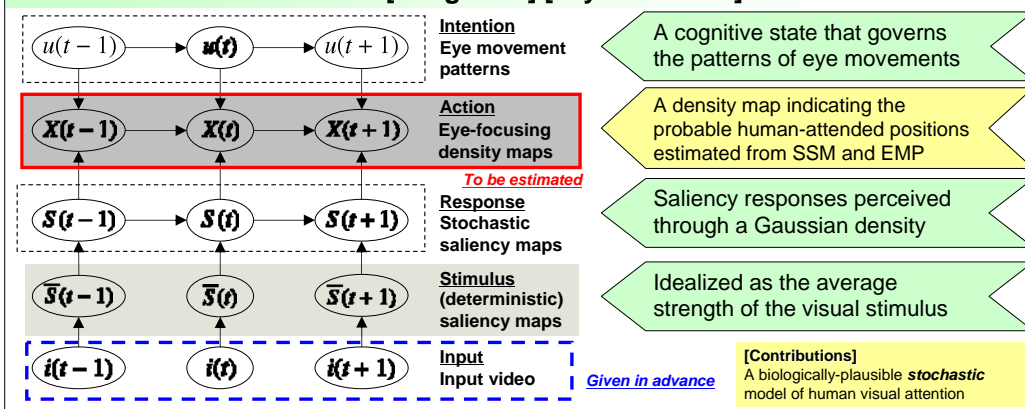


Obtaining verbal information from “parents”

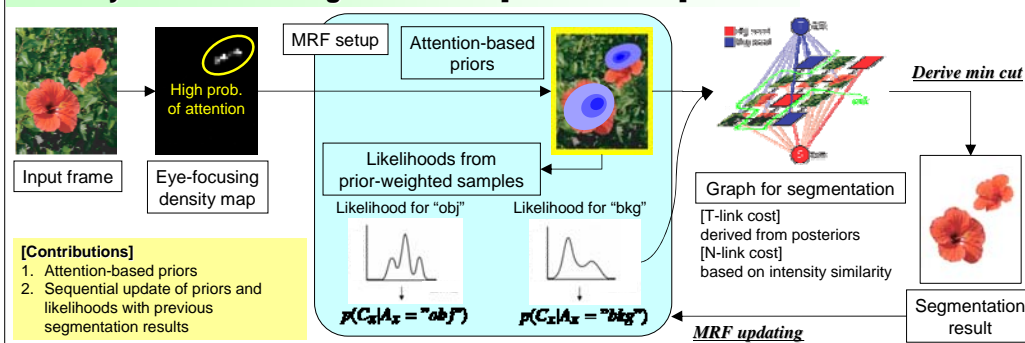
Our framework of visual scene understanding



Visual attention estimation [Pang 2008] [Miyazato 2009]



Saliency-based video segmentation [Fukuchi 2009]



Perspectives: The next step towards the “ultimate goal”

- **Smooth transition of learning strategy:** (1) Supervised → semi-supervised + reinforcement (2) Bottom-up attention-based → Top-down knowledge-based
- **Scalability** for a large amount of acquired “knowledge”
- **Interaction strategy** to educe “intrinsic information” from partners (humans, possibly other systems)
- **Multi modality:** Audiovisual attention, tactile perception, “shape from tactile”

[Pang 2008] Pang, Kimura et al. “A stochastic model of selective visual attention with a dynamic Bayesian network,” Proc. ICME2008 (Lecture)
 [Miyazato 2009] Miyazato, Kimura, et al. “Real-time estimation of human visual attention with dynamic Bayesian network and MCMC-based particle filter,” Proc. ICME2009 (Long paper, Lecture)
 [Fukuchi 2009] Fukuchi, Miyazato et al. “Saliency-based video segmentation with graph cuts and sequentially-updated priors,” Proc. ICME2009