Local Descriptor Groupings in Reinforcement Learning of Sensory-Motor Attention

Christin Seifert, Gerald Fritz, and Lucas Paletta

JOANNEUM RESEARCH Forschungsgesellschaft - Institute of Digital Image Processing

Computational Perception Group (CAPE)

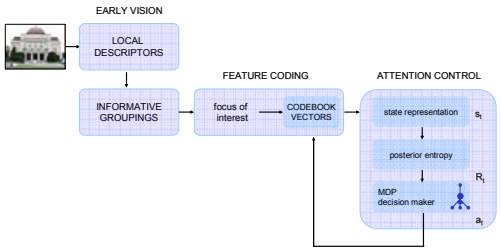
Wastiangasse 6, A - 8010 Graz, Austria

E-mail: {lucas.paletta}@joanneum.at

CONCEPT

- Iterative *object recognition* from *local descriptor groupings and geometry*
- *Transsaccadic* encoding of objects is *learned*
- Object recognition in a *perception-action* framework
- *Perceptual meta-states* in sensory-motor attention
- Local descriptors are *grouped* according to *information content*
- *Purposeful visual grammar* from learned feature-action encodings

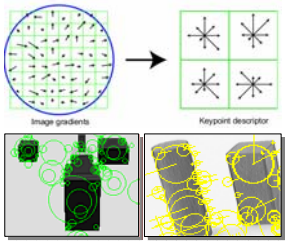
ATTENTIVE RECOGNITION PROCESS



Human Previous research on behavioural modelling of saccadic image interpretation (Henderson, 1982 Psychological Science : 51 - 55) has emphasised the sampling of informative parts under visual attention to guide visual perception. We propose a system of sequential attention for object recognition that (i) groups n-tuples of local gradient based image descriptors (Lowe, 2004 International Journal of Computer Vision 60 91 - 110) being scale, rotation, and to high degree illumination tolerant, defining a vocabulary of prototypical code descriptors, (ii) selects only informative groupings for further processing, (iii) learns a predictive mapping from a current perceptual state in a Markov decision process to a next saccadic action, and (iv) present a model of object recognition being capable of integrating sequential information by minimization of entropy in the Bayesian modeling of object hypotheses. The innovative abstraction level of informative groupings provides perceptual meta-states in sensory-motor attention, enabling the learning of a purposeful grammar integrating atomic feature-saccade mappings into a meaningful recognition behaviour. We demonstrate highly accurate recognition of outdoor facades in a mobile vision application, using the sensory-motor context of trans-saccadic object recognition.

INFORMATIVE GROUPINGS

Informative features are selected using an information theoretic saliency measure on SIFT descriptors (for the informative approach see Fritz et al., ICPR 2004). These features support focusing attention on most salient image regions for further investigation (Fritz et al., AAAI 2004).



SIFT features: *Keypoints are local maxima of the DOG function. The SIFT descriptor is computed from the orientation histogram in a local environment of the keypoint. The contribution of neighbouring features is weighted by a Gaussion function.*

Determining the local information content from the appearance imagettes:

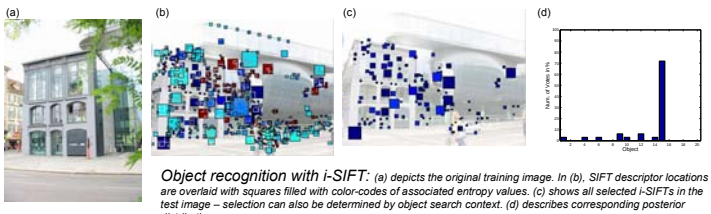
- The SIFT features $\boldsymbol{x}_i$ in the image, i.e. vectors containing 128 floating points are mapped in a low-dimensional subspace using PCA: $\boldsymbol{g}_i = \mathbf{E}\,\boldsymbol{x}_i$.
- In order to get the information content of a sample $\boldsymbol{g}_i$, we estimate the Shannon entropy:

$$H(O \mid \mathbf{g}_i) \equiv -\sum_k P(o_k \mid \mathbf{g}_i) \log P(o_k \mid \mathbf{g}_i).$$

- Approximate the posterior distribution $P(o_k|\boldsymbol{g}_i)$, inside a Parzen window within a local neighborhood of size

$$\varepsilon \geq \| \boldsymbol{g}_i - \boldsymbol{g}_j \|.$$

- The estimate about the Shannon entropy $\hat{H}(O \mid \boldsymbol{g}_i)$ provides a measure of ambiguity with respect to object identification within a single local observation $\boldsymbol{g}_i$.



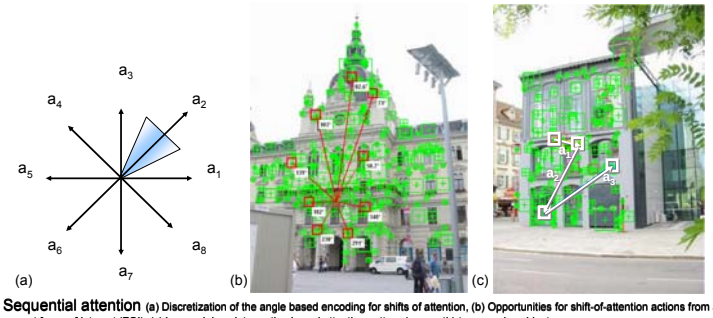
Object recognition with i-SIFT: *(a) depicts the original training image. In (b), SIFT descriptor locations are overlaid with squares filled with color-codes of associated entropy values. (c) shows all selected i-SIFTs in the test image – selection can also be determined by object search context. (d) describes corresponding posterior distribution.*

ATTENTIVE SACCADE SHIFTS

Attention on informative local image patterns is shifted between largest local maxima derived by the information theoretic saliency measure. Saccadic actions originate from a randomly selected maximum and target towards one of the n-best ranked maxima (here: n=10). Saccadic actions are categorized into 8 principal directions. At each local maximum, the extracted local feature is associated to a codebook vector of nearest distance in feature space.

Foci of Interest (FOIs) and codebook vectors

- Compute posterior entropy of local descriptors.
- Apply thresholding on descriptor entropy and generate entropy-sorted list.
- Start attention sequence with lowest entropy keypoint.
- Assign codebook vector to keypoint in the FOI.
- Codebook vectors (k=20) extracted from unsupervised learning on the keypoint distribution.
- Actions are categorized in 8 principle directions.
- The next action is derived from Q-Learning classifier.



Sequential attention **(a) Discretization of the angle based encoding for shifts of attention, (b) Opportunities for shift-of-attention actions from a current focus of interest (FOI), (c) Learned descriptor-action based attention pattern (scanpath) to recognize object o_{14}**

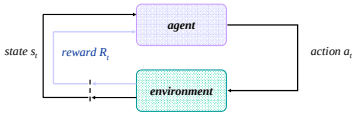
TSG-20 Database (Tourist Sights of Graz, http://dib.joanneum.at/cape/TSG-20)

Q-LEARNING OF WORKING MEMORY

We formalize the sequence of action selections in sequential attention as a Markovian Decision Process MDP (with state space S, action space A, transition function P, reward function R) and are searching for optimal solutions with respect to the object recognition task. Since the probabilistic transition function P cannot be known beforehand, the probabilistic model of the task is estimated via reinforcement learning, e.g., by ***Q-learning*** (Watkins & Dayan, Machine Learning 1992), which guarantees convergence to an optimal policy applying iterative updates of the Q-function

$$Q(s,a) = Q(s,a) + \alpha[R + \gamma(\max_{a'} Q(s',a') - Q(s,a))].$$

where α is the learning rate, γ controls the impact of a current action on future policy return values.



Definitions:

Markovian recognition state: s_i := $[f_1, a_1, f_2, a_2, f_3, \ldots]$. f_i represents *ST working memory*.

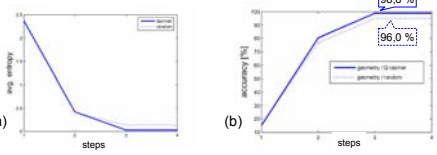
Reward function: entropy decrease R := -dH = -(H_2-H_1), H... conditional entropy.

Entropy at each s_i : frequency of object specific traces visiting s_i.

Actions: saccade shifts in 2D, angle discretized, a_1,..., a_8, a = (N, NE, E, SE, ...).

RESULTS

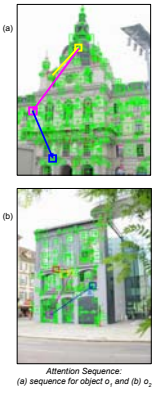
- *20 objects*
- *4 views*
- *20 train images*
- *20 test images*
- *320 x 240 pixels*
- *350 SIFT keys per image on average*



(c)

method	*accuracy*	*computation time*
Random	*96.0 %*	*2800 ms*
q-SIFT	*98.8 %*	*1500 ms*

Performance:
(a) entropy decreases faster with q-SIFT, (b) accuracy increases faster with q-SIFT, (c) performance comparison



Attention Sequence: (a) sequence for object o_1 and (b) o_2

ACKNOWLEDGMENT *This work is supported by the European Commission funded projects MACS under grant number FP6-004381 and MOBVIS under grant number FP6-511051, and by the the FWF Austrian Joint Research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04.*

FSP COGNITIVE VISION S9104-N04