Reinforcement Learning of Informative Attention Patterns for Object Recognition

Lucas Paletta, Gerald Fritz, and Christin Seifert
Computational Perception Group, Institute of Digital Image Processing
JOANNEUM RESEARCH Forschungsgesellschaft mbH
A-8010 Graz, Austria
Email: lucas.paletta@joanneum.at

***Abstract*— Attention is a highly important phenomenon emerging in infant development [1]. In human perception, sequential visual sampling about the environment is mandatory for object recognition purposes. Sequential attention is viewed in the framework of a saccadic decision process that aims at minimizing the uncertainty about the semantic interpretation for object or scene recognition. Methodologically, this work provides a framework for learning sequential attention in real-world visual object recognition, using an architecture of three processing stages. The first stage rejects irrelevant local descriptors providing candidates for foci of interest (FOI). The second stage investigates the information in the FOI using a codebook matcher. The third stage integrates local information via shifts of attention to characterize object discrimination. A Q-learner adapts then from explorative search on the FOI sequences. The methodology is successfully evaluated on representative indoors and outdoors imagery, demonstrating the significant impact of the learning procedures on recognition accuracy and processing time.**

I. Introduction

To function well, infants need to be selectively attending to only a small portion of the information available. At the same time, they must be responsive to improtant events as they occur [1]. Interdependencies between learning, attention, and decision making have been frequently emphasized [1], [2] but did not yet lead to working solutions in real world environments, particularly in computer vision. Recent research in neuroscience [3], [4] and experimental psychology [5] has confirmed evidence that decision behavior plays a dominant role in human selective attention in object and scene recognition. E.g., there is psychophysical evidence that human observers represent visual scenes not by re-constructing but merely by purposive encodings via meaningful attention patterns [6], [7] probing only few relevant features from a scene. This leads on the one hand to the assumption of transsaccadic object memories [4], and supports theories about the effects of sparse information sampling due to change blindness when humans cannot compare dynamically built sparse representations of a scene under impact of attentional blinks [8]. Current biologically motivated computational models on sequential attention reflect the encoding of scenes and relevant objects from sequential attention in the framework of neural network modeling [7] and probabilistic decision processes [10], [11].

The original contribution of this work is to demonstrate the learning of attention patterns in a computational approach that is inspired by human perception. By providing a scalable approach for the learning of sequential visual attention for the purpose of object recognition in real-world environments, it demonstrates the importance of cascaded processing stages for the selection, memorization and re-cognition of discriminative information.

Firstly, the methodology proposes to integrate local information only at locations that are relevant with respect to the task, in terms of an information theoretic saliency measure. Secondly, it enables to apply efficient strategies to group informative local descriptors using a decision maker. The decision making agent uses Q-learning to associate *shift of attention*-actions to cumulative reward with respect to a task goal, i.e., object recognition. Reward is determined for the reduction of entropy for recognition. Objects are represented in a framework of perception-action, providing a transsaccadic *working* memory that stores useful grouping strategies of a kind of *hypothesize and test* behavior.

In computer vision, recent research has been focusing on the integration of information received from local descriptors into a more global analysis with respect to object recognition [12], [13]). The solutions are assuming statistical independence of the local responses, exclude segmentation problems by assuming single object hypotheses in the image, or assume regions with uniformly labelled operator responses.

In object recognition terms, this method enables to match not only between local feature responses, but also taking the geometrical relations between the specific features into account, thereby defining their more global visual configuration. The proposed method is outlined in a perception-action framework, providing a sensorimotor decision maker that selects appropriate saccadic actions to focus on target descriptor locations. The advantage of this framework is to become *able to start interpretation from a single local descriptor* and, to continuously and iteratively integrate local descriptors 'on the fly' while evaluating the current geometric configuration for efficient discrimination.

Fig. 1 illustrates the closed loop object recognition process. Visual information is attended for recognition exclusively at *salient* image locations, using a cascaded attention framework to keep complexity low. In a first processing stage (*early vision*), salient image locations are selected using an information theoretic measure with respect to object discrimination [14].

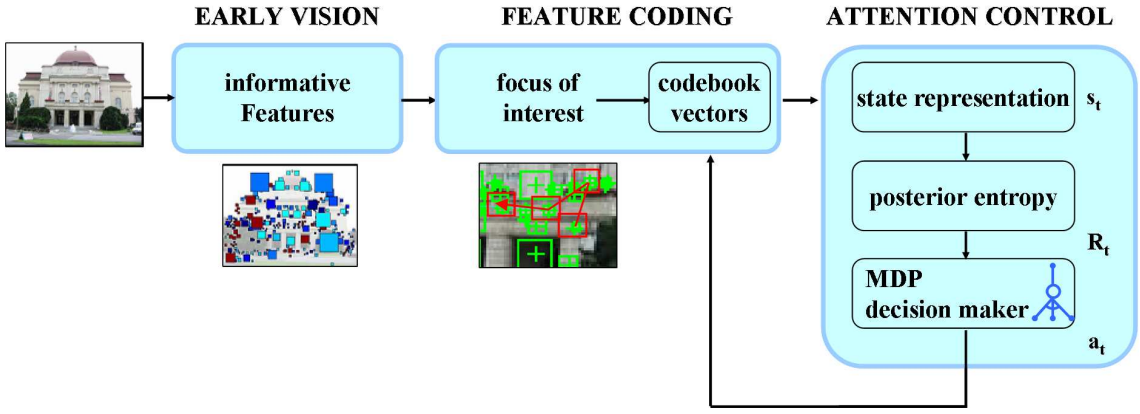


Fig. 1. Concept of cascaded sequential attention for object recognition. In early vision, the system extracts informative local descriptors and focus of interest, where descriptors are encoded with respect to codebook vectors. Descriptor-action sequences defi ne the state, posterior and entropy decrease to drive useful actions – closing the loop.

The information in the focus of interest is then matched to codebook vectors to receive weak object hypotheses (*feature coding*). Descriptor-action sequences determine recognition states that are then associated with object posteriors that define the decreasse in posterior entropy (*reward*) and drive selection of *shift-of-attention actions*.

In the training stage, the reinforcement learner performs trial and error search on useful actions, receiving reward from entropy decreases. In the test stage, the decision maker demonstrates feature grouping and matching between the encountered and the trained attentive sensorimotor patterns. The method is evaluated in experiments on object recognition using the reference COIL-20 (indoor imagery) and the TSG-20 object (outdoor imagery) database, proving the method being computationally feasible and providing rapid convergence in the discrimination of objects.

II. Informative Foci of Interest

In the first two processing stages, we determine informative local descriptors (Sec. II-A) and investigate the focus of interest for weak object hypotheses (Sec. II-B). Relating the information theoretic cost measure with respect to all individual pixels, we extract a saliency map, i.e., a biologically motivated intermediate representation used in visual attention [15] (i) to relate image content directly to cost measure, and (ii) to easily determine regions of interest from maxima in the saliency map values.

In this work, descriptors are either represented by normalized brightness (appearance) patterns [14], or by the Scale Invariant feature Transform (SIFT) [13]. While appearance patterns provide fundamental analysis for each pixel, SIFT descriptors are more sparsely distributed, but they are known to be rotation-, scale- and, to a high degree, illumination invariant.

A. Saliency from Local Information Content

We determine the saliency from an information theoretic measure to evaluate an early vision feature (descriptor, i.e., a pattern of visual information) with respect to its utility for a given task, i.e., object recognition. The resulting local entropy value is then associated to the corresponding pixel in the saliency map.

The object recognition task is formally related to the sampling of local descriptors $\mathbf{f}_i$ in feature space $\mathcal{F}$, $\mathbf{f}_i \in \mathcal{R}^{|\mathcal{F}|}$, where o_i denotes an object hypothesis from a given object set Ω. We need to estimate the entropy $H(O|\mathbf{f}_i)$ of the posteriors $P(o_k|\mathbf{f}_i)$, $k = 1 \ldots \Omega$, Ω is the number of instantiations of the object class variable O. Shannon conditional entropy denotes

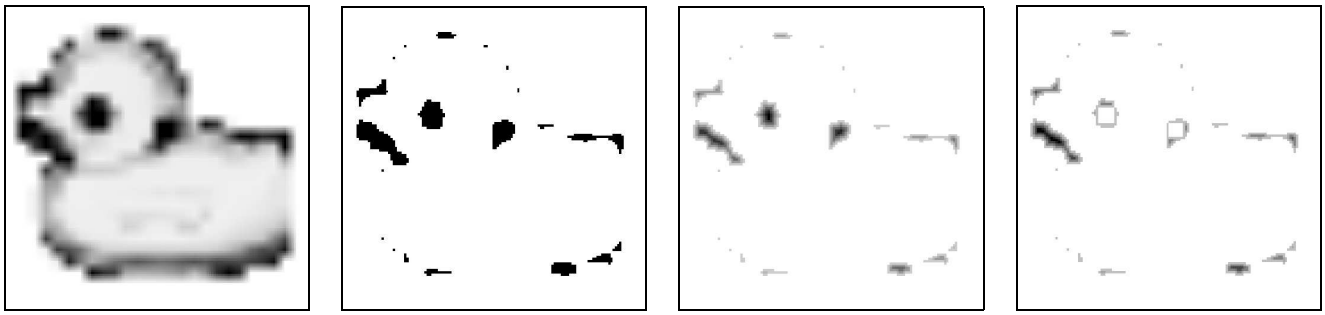
$$H(O|\mathbf{f}_i) \equiv -\sum_k P(o_k|\mathbf{f}_i) \log P(o_k|\mathbf{f}_i). \quad (1)$$

Instead of a global estimate on the posterior, we approximate the posteriors at $\mathbf{f}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window of a local neighborhood ϵ, $||\mathbf{f}_i - \mathbf{f}_j|| \leq \epsilon$, $j = 1 \ldots J$. The estimate about the conditional entropy $\hat{H}(O|\mathbf{f}_i)$ provides then a measure of ambiguity in terms of characterizing the information content with respect to object identification within a single local observation $\mathbf{f}_i$ [14].

B. Sequential Focus from Saliency Maps

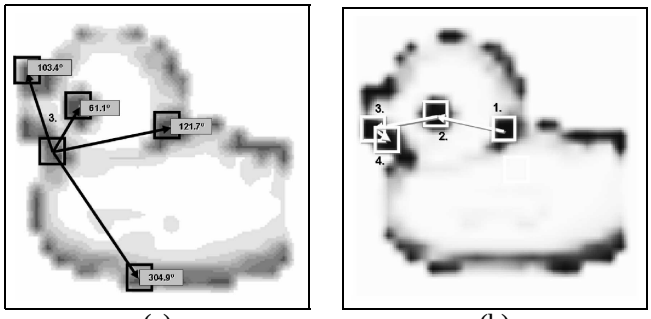
Attention on local descriptors is shifted between the largest local maxima of the information theoretic saliency measure (Sec. II-A).

Fig. 2 depicts the principal stages in selecting the FOIs. From the saliency map (a), one computes a binary mask (b) that represents the most informative regions with respect to the conditional entropy in Eq. 1, by selecting each pixel's contribution to the mask from whether its entropy value H is smaller than a predefined entropy threshold H_Θ, i.e., $H < H_\Theta$. (c) applying a distance transform on the binary regions of interest results mostly in the accurate localization of the entropy minimum. The maximum of the local distance transform value is selected as FOI. Minimum entropy values and maximum transform values are combined to give a location of interest for the first FOI, applying a 'Winner-takes-it-all' (WTA) principle [16]. (d) Masking out the selected maximum of the first FOI,



(a) (b) (c) (d)

Fig. 2. Extraction of foci of interest from an information theoretic saliency map (Sec. II). (a) Saliency from the entropy in local brightness patterns (dark=low entropy). (b) Binarization from thresholding for most informative regions. (c) Distance transform on informative regions. (d) Masking of already processed regions (inhibition of return).



(a) (b)

Fig. 3. Generation of attention patterns. (a) The Shift-of-attention action originates in a randomly selected saliency maximum and is directed towards four next ranked target foci. (b) Learned attention pattern (scanpath) to characterize and recognize the object.

one can apply the same WTA rule, selecting the maximum saliency. This masking is known as 'inhibition of return' [17] in human visual attention.

III. Sensorimotor Sequential Attention

Sequential attention shifts the focus of attention between the most informative patterns in the order of associated saliency values, in this sense representing a step-wise generation of a scanpath [6]. There is two kind of information that characterizes objects for discrimination from scanpaths, (i) the visual information within the focus of attention, and (ii) the geometry between the sequentially accessed FOIs, i.e., the shift-of-attention action translating between FOIs. In this work we claim that the pattern in the FOI must not necessarily be represented in finest detail but an approximate characterization will suffice to give a weak object hypothesis. This renders the algorithm tolerant to noise and failures in the local interpretation, but on the other hand gives rise to analyse the spatial context, i.e., the geometry between the descriptors, in more detail.

Descriptor encodings The visual information in the FOI is associated to a prototypical reference vector to give a weak object hypothesis: At each local maximum, the extracted local pattern $\mathbf{g}_i$ is associated to a codebook vector Γ_j of nearest distance $d = \arg min_j ||\mathbf{g}_i - \Gamma_j||$ in feature space. The codebook vectors can be estimated from k-means clustering of a training sample set $G = \mathbf{g}_1, \cdots, \mathbf{g}_N$ of size N ($k = 20$ in the experiments). The focused local information pattern is therefore associated to the label of the *k*-th prototype vector, gaining discrimination merely from the geometric relations between focus encodings in order to discriminate attention patterns.

Action The shift-of-attention actions target in the proposed method towards one out of next n best-ranked maxima (e.g., n=4 in Fig. 3a) within the information theoretic saliency map. Saccadic actions originate from a randomly selected local maximum of saliency and target towards one of the remaining (n-1) best-ranked maxima via a saccadic action $a \in A$ (Fig. 3a). The individual action and its corresponding angle $\alpha(x, y, a)$ is then categorized into one out of $|A| = 8$ principal directions ($\Delta a = 45°$).

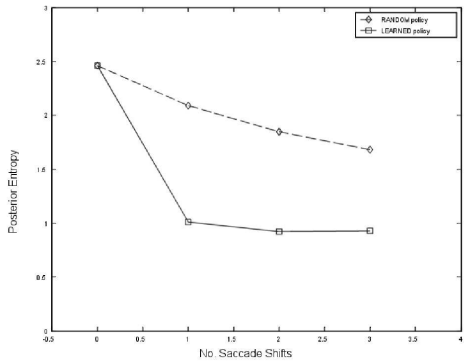
Scanpath An individual state s_i is finally represented by a complete (or part of) a sequential attention pattern, i.e., the scanpath. The attention pattern of length n is encoded by a sequence of descriptor encodings Γ_j and actions $a \in A$, i.e.,

$$s_i = (\Gamma_1, a_2, \cdots, \Gamma_{n-1}, a_n, \Gamma_n)\,. \tag{2}$$

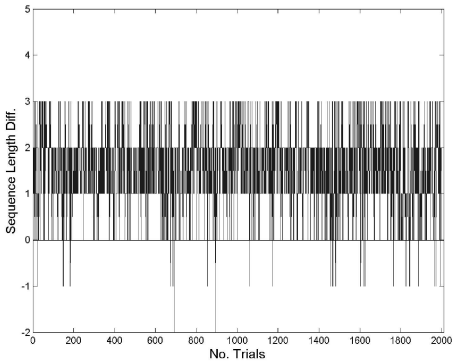
Posteriors In order to characterize the discriminative value of a scanpath, we determine an estimate on the posterior on object hypotheses, given a particular descriptior-action sequence. The posterior is estimated from frequency histogramming: Within the object learning stage, random actions will lead to arbitrary descriptor-action sequences, i.e., attention patterns. For each attention pattern, we protocol the number of times it was experienced in the context of the corresponding object in the database. From this we are able to estimate a mapping from states s_i to posteriors, i.e., $s_i \mapsto P(o_k|s_i)$, by monitoring how frequent states are visited under observation of particular objects. From the posterior we compute the conditional entropy $H_i = H(O|s_i)$ and the *information gain* with respect to actions leading from state $s_{i,t}$ to $s_{j,t+1}$ by

$$\Delta H_{t+1} = H_t - H_{t+1}. \tag{3}$$

An efficient strategy aims then at selecting in each state $s_{i,t}$ the action a^* that would maximize the information gain



(a)



(b)

Fig. 4. Performance evaluation on appearance patterns (Sec. II). (a) Learned strategies lead to lower posterior entropy levels within shorter attention sequences. (b) Random strategies require more actions to attain an entropy threshold (task goal) (threshold $H_{goal} = 1.2$).

$\Delta H_{t+1}(s_{i,t}, a_{k,t+1})$ received from attaining state $s_{j,t+1}$, i.e.,

$$a^* = \arg max_a \Delta H_{t+1}(s_{i,t}, a_{k,t+1}). \quad (4)$$

IV. Q-Learning of Attentive Saccades

In each state of the sequential attention process (Sec. III), a decision making agent is asked to perform a strategy to select an action to arrive at a most reliable recognition decision. Learning to recognize objects means then to explore different descriptor-action sequences, to quantify consequences in terms of a utility measure, and to adjust the control strategy thereafter. In the following we motivate to define sequential attention as a decision process, and address to use reinforcement learning to extract the optimal policy from explorative search since we lack a precise model of the underlying statistics.

Markov decision processes (MDPs) have already been introduced for object recognition by [18] in the sense of optimal selection of visual procedures. Here, the MDP will provide the general framework to outline sequential attention for object recognition in a multistep decision task with respect to the discrimination dynamics. An MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, \delta, \mathcal{R})$ with state recognition set $\mathcal{S}$, action set $\mathcal{A}$, probabilistic transition function δ and reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \Pi(\mathcal{S})$ describes a probability distribution over subsequent states, given the attention shift action $a \in \mathcal{A}$ executable in state $s \in \mathcal{S}$. In each transition, the agent receives reward according to $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto R$, $\mathcal{R}_t \in R$. The agent must act to maximize the utility $Q(s, a)$, i.e., the expected discounted reward

$$Q(s,a) \equiv U(s,a) = E\left[\sum_{n=0}^{\infty} \gamma^n \mathcal{R}_{t+n}(s_{t+n}, a_{t+n}))\right], \quad (5)$$

where $\gamma \in [0, 1]$ is a constant controlling contributions of delayed reward.

We formalize a sequence of action selections $a_1, a_2, \cdots, a_n$ in sequential attention as an MDP and are searching for optimal solutions with respect to finding action selections so as to maximizing future reward with respect to the object recognition task. With each shift-of-attention, the entropy reduction gives feedback about the reduction of uncertainty and therefore the quality of a related recognition decision. With each action, the reward in terms of information gain (Eq. 3) in the posterior distribution on object hypotheses, is received from attention shift a by

$$\mathcal{R}(s, a) := \Delta H. \quad (6)$$

Since the probabilistic transition function $\Pi(\cdot)$ cannot be known beforehand, the probabilistic model of the task is estimated via reinforcement learning, e.g., by Q-learning [19] which guarantees convergence to an optimal policy applying sufficient updates of the Q-function $Q(s, a)$, mapping recognition states s and actions a to utility values. The Q-function update rule is

$$Q(s,a) = Q(s,a) + \alpha\left[R + \gamma(max_{a'} Q(s', a') - Q(s,a))\right], \quad (7)$$

where α is the learning rate, γ controls the impact of a current shift of attention action on future policy returns.

The decision process in sequential attention is determined by the sequence of choices on shift actions at a specific focus of interest (FOI). The agent selects then the action $a \in \mathcal{A}$ with largest $Q(s, a)$, i.e.,

$$a_T = \arg max_{a'} Q(s_T, a'). \quad (8)$$

V. Experimental Results

The sequential attention methodology was applied to experiments with (i) indoor imagery (COIL-20 database), and with (ii) outdoor imagery (TSG-20 database) on the task of object recognition. The indoor images do not contain any illumination or noise artefacts, therefore we expect and finally prove high accuracy in the recognition results, similar to existing methodologies but still proving superiority of learned in contrast to random decision policies. Outdoor images are much more challenging with respect to variance in the viewpoints, the illumination, and also the distance to the objects (scale). There, we proved that the geometry in the sequential attention provided good discrimination, but above all, that the learned policy can significantly outperform standard recognition methodology, both with respect to recognition accuracy and computing times.

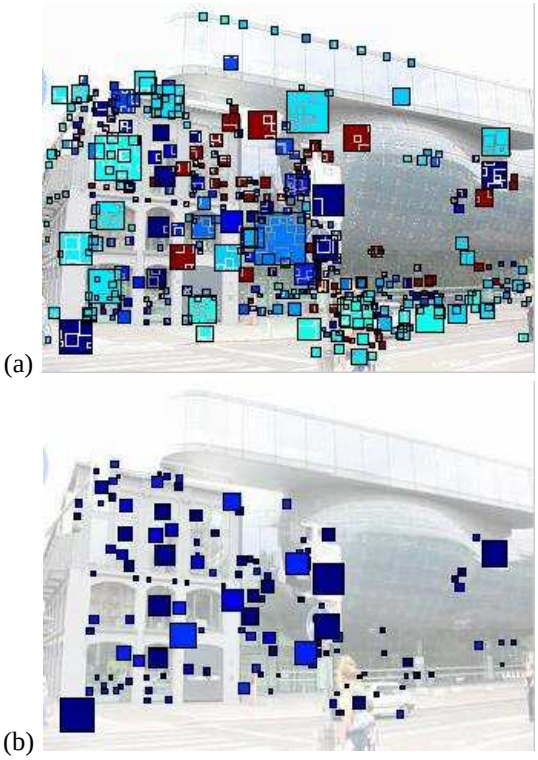


Fig. 5. Informative descriptors for early vision. (a) Position and scale of local descriptors (SIFT [13], (b) Selection of most informative (dark coded) descriptors for further processing (Sec. II).

A. Local Appearance Descriptors (indoors)

The indoor experiments were performed on 1440 images of the COIL-20 database (20 objects and 72 views by rotating each object by 5° around its vertical rotation axis), investigating up to 5 FOIs in each observation sequence, associating to $k = 20$ codebook vectors from informative appearance patterns, in order to determine the recognition state, and deciding on the next saccade action to integrate the information from successive image locations. Fig. 4a represents the learning process, illustrating more rapid entropy decreases from the learned in contrast to random action selection policy. Fig. 4b visualizes the advantages from learning by requiring less actions to attain more informative recognition states. The recognition rate after the second action was 92% (learned) in contrast to 75% (random). A characteristic learned attention scanpath is depicted in Fig. 3b.

B. SIFT Descriptors (outdoors)

In the outdoor experiments, we decided to use a local descriptor, i.e., the SIFT descriptor (Sec. II) due to its superior robustness to viewpoint, illumination and scale changes. The experimental results were obtained from the images of the TSG-20 database (Fig. 6a, 20 objects and 2 views by approx. 30° viewpoint change), investigating up to 5 FOIs in each observation sequence, associating to $k = 20$ codebook vectors to determine the recognition state, and deciding on the next saccade action to integrate the information from successive image locations. Fig. 7a visualizes the progress gained from

TABLE I

PERFORMANCE COMPARISON BETWEEN LEARNED AND RANDOM SEQUENTIAL ATTENTION (SEQA) POLICIES ON TSG-20 (WITH SIFT), AND STATE-OF-THE-ART INFORMATIVE SIFT RECOGNITION [20], COMPARING RECOGNITION *accuracy* AND COMPUTING *times*.

METHOD	ACCURACY (%)	TIME (MS)
Q-LEARN SEQA	98.8± 0.4	1500
RANDOM SEQA	96.0± 1.2	1200
I-SIFT	97.5± 0.9	2800

the learning process in requiring less actions to attain more informative recognition states. Fig. 7b reflects the corresponding learning process, illustrating more rapid entropy decreases from the learned in contrast to random action selection policy. The recognition rate after the second action was $\approx 98.8\%$ (learned) in contrast to $\approx 96.0\%$ (random, see Table I). A characteristic learned attention scanpath is depicted in Fig. 3b.

Fig. 5 depicts the principal stages in the selection of FOIs. (a) depicts the test image overlaid with squares brightness-coded with respect to associated entropy values (dark=low). (b) depicts the selection of the most informative descriptors from (a). Fig. 6 illustrates (b) various opportunities for action from a given FOI, and (c) a learned sequential attention sequence using the SIFT descriptor.

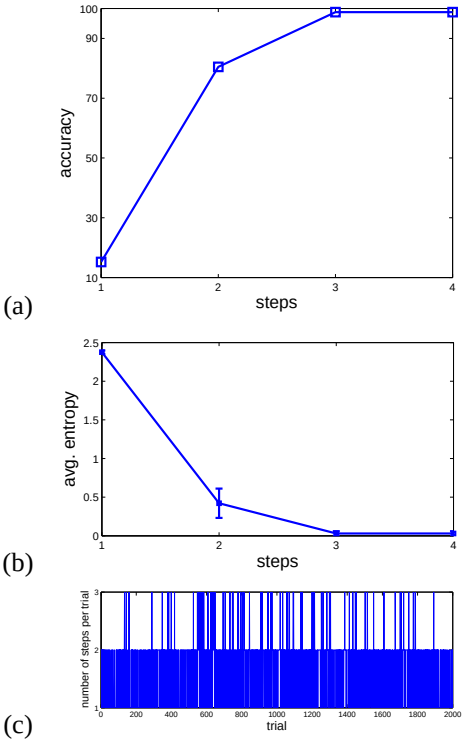
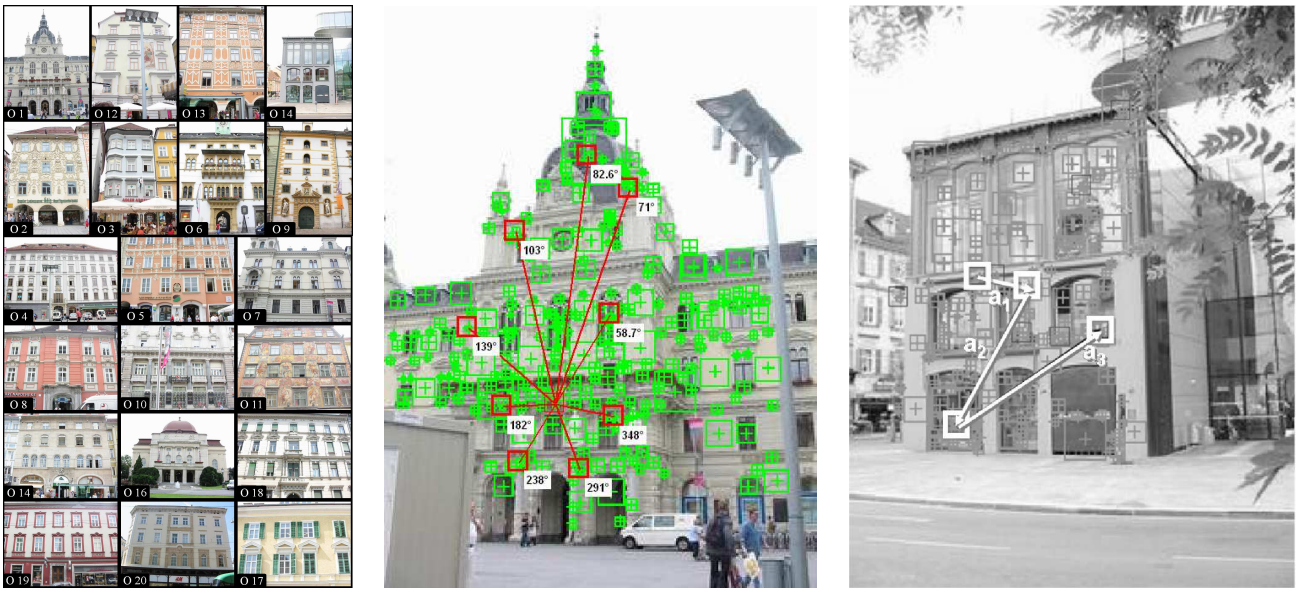


Fig. 7. Performance evaluation of learned policy. (a) Step-wise improvement in recognition accuracy. (b) Step-wise entropy reduction. (c) Number of actions required to attain task goal (entropy threshold).



(a) (b) (c)

Fig. 6. (a) Objects of the TSG-20 database (Sec. V). (b) Opportunities for shift-of-attention actions from a current FOI. (c) Learned descriptor-action based attention pattern (scanpath) to recognize object o_{14}.

VI. Conclusions

The proposed methodology significantly extends previous work on sequential attention and decision making by providing a scalable framework for learning attention in real world object recognition. The three-stage process, (i) determining information theoretic saliency, (ii) characterizing the visual information in the FOI, and (iii) integrating local descriptive information in a perception-action recognition process is robust with respect to viewpoint, scale, and illumination changes using the standrad descriptor SIFT [13], and finally provides rapid attentive matching by requiring only very few local descriptor samples to be evaluated for object discrimination. Future work will be directed towards hierarchical reinforcement learning in order to provide local grouping schemes that will be globally integrated.

Acknowledgment

This work is supported by the European Commission projects MACS (FP6-004381) and MOBVIS (FP6-511051), and by the FWF Austrian Joint Research Project Cognitive Vision (sub-projects S9103-N04, S9104-N04).

References

[1] H. Ruff and M. Rothbart, *Attention in early development*. New York, NY: Oxford University Press, 1996.

[2] P. Dayan, S. Kakade, and P. Montague, "Learning and selective attention," *Nature Neuroscience*, no. 3, pp. 1218–1223, 2000.

[3] G. Deco, "The computational neuroscience of visual cognition: Attention, memory and reward," in *Proc. International Workshop on Attention and Performance in Computational Vision*, 2004, pp. 49–58.

[4] H. Deubel, "Localization of targets across saccades: Role of landmark objects," *Visual Cognition*, no. 11, pp. 173–202, 2004.

[5] J. Henderson, "Human gaze control in real-world scene perception," *Trends in Cognitive Sciences*, vol. 7, pp. 498 –504, 2003.

[6] L. W. Stark and Y. S. Choi, "Experimental metaphysics: The scanpath as an epistemological mechanism," in *Visual attention and cognition*, W. H. Zangemeister, H. S. Stiehl, and C. Freska, Eds. Amsterdam, Netherlands: Elsevier Science, 1996, pp. 3–69.

[7] I. Rybak, I. G. V., A. Golovan, L. Podladchikova, and N. Shevtsova, "A model of attention-guided visual perception and recognition," *Vision Research*, vol. 38, pp. 2387–2400, 1998.

[8] R. Rensink, J. O'Regan, and J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, vol. 8, pp. 368–373, 1997.

[9] C. Bandera, F. Vico, J. Bravo, M. Harmon, and L. B. III, "Residual Q-learning applied to visual attention," in *International Conference on Machine Learning*, 1996, pp. 20–27.

[10] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," in *Proc. International Conference on Autonomous Agents*, 2001, pp. 457–464.

[11] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. European Conference on Computer Vision*, 2000, pp. 18–32.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] G. Fritz, C. Seifert, L. Paletta, and H. Bischof, "Rapid object recognition from discriminative regions of interest," in *Proc. National Conference on Artificial Intelligence, AAAI 2004*. San Jose, CA, 2004, pp. 444–449.

[14] L. Paletta, E. Rome, and H. Buxton, "Attention architectures for machine vision and mobile robots," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos, Eds. Amsterdam, Netherlands: Elsevier Science, 2005, pp. 642–648.

[15] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[16] S. Tipper, S. Grisson, and K. Kessler, "Long-term inhibition of return of attention," *Psychological Science*, vol. 14, pp. 19–25–105, 2003.

[17] B. Draper, J. Bins, and K. Baek, "ADORE: adaptive object recognition," in *Proc. International Conference on Vision Systems, ICVS 1999*. Las Palmas de Gran Canaria, Spain, 1999, pp. 522–537.

[18] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3,4, pp. 279–292, 1992.

[19] G. Fritz, C. Seifert, L. Paletta, and H. Bischof, "Building recognition using informative local descriptors from mobile imagery," in *Proc. Scandinavian Conference on Image Analysis, SCIA 2005*. Joensuu, Finland, in print, 2005.