Cascaded Sequential Attention for Object Recognition with Informative Local Descriptors and Q-learning of Grouping Strategies[*]

Lucas Paletta, Gerald Fritz, and Christin Seifert
Institute of Digital Image Processing
JOANNEUM RESEARCH Forschungsgesellschaft mbH
Graz, A-8010, Austria
{lucas.paletta,gerald.fritz,christin.seifert}@joanneum.at

Abstract

The contribution of this work is to provide a three-stage architecture for sequential attention to provide a system being capable of sensorimotor object detection in real world environments. The first processing stage provides selected foci of interest in the image based on the extraction of information theoretic saliency of local image descriptors (i-SIFT). The second stage investigates the information in the local attention window using a codebook matcher, providing local weak hypotheses about the identity of the object under investigation. The third stage then proposes a shift of attention to a next attention window. The working hypothesis is to expect a better discrimination from the integration of both the individual local FOA patterns and the geometric relation between them, providing a model of more global information representation, and feeding into a recognition state in the Markov Decision Process (MDP). A reinforcement learner (Q-learner) performs then explorative search on useful actions, i.e., shifts of attention, towards locations of salient information, developing a strategy of useful action sequences being directed in state space towards the optimization of discrimination by information maximization. The method is evaluated in experiments using the COIL-20 database (indoor imagery) and the TSG-20 database (outdoor imagery) to demonstrate efficient performance in object detection tasks, proving the method being more accurate and computationally much less expensive than standard SIFT based recognition.

1. Introduction

Recent research in neuroscience [1, 2] and experimental psychology [3, 4, 5] has provided evidence that decision behavior plays a dominant role in human selective attention for object and scene recognition . E.g., there is psychophysical evidence that human observers represent visual scenes not by extensive reconstructions 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 [5], 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 identify shift invariant descriptions of sampling sequences [9], and 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].

In computer vision, recent research has been focusing on the integration of information received from single local descriptor responses into a more global analysis with respect to object recognition [13, 14]). State-of-the-art solutions, such as, (i) identifying the MAP hypothesis from probabilistic histograms [15], (ii) integrating responses in a statistical dependency matrix [13], and (iii) collecting evidence for object and view hypotheses in parametric Hough space [14], provide convincing performance under assumptions, such as, statistical independence of the local responses, excluding segmentation problems by assuming single object hypotheses in the image, or assuming regions with uniformly labelled operator responses. An integration strategy closing methodological gaps when above assumptions are violated should therefore (i) cope with statistical dependency between local features of an object, (ii) enable to segment multiple targets in the image and (iii) provide convincing evidence for the existence of object regions merely on the geometry than on the relative frequency of labelled local responses.

The original contribution of this work is to provide a scalable framework for cascaded sequential attention in real-world environments. Firstly, it proposes to integrate

[*] 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 FWF Austrian Joint research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04.

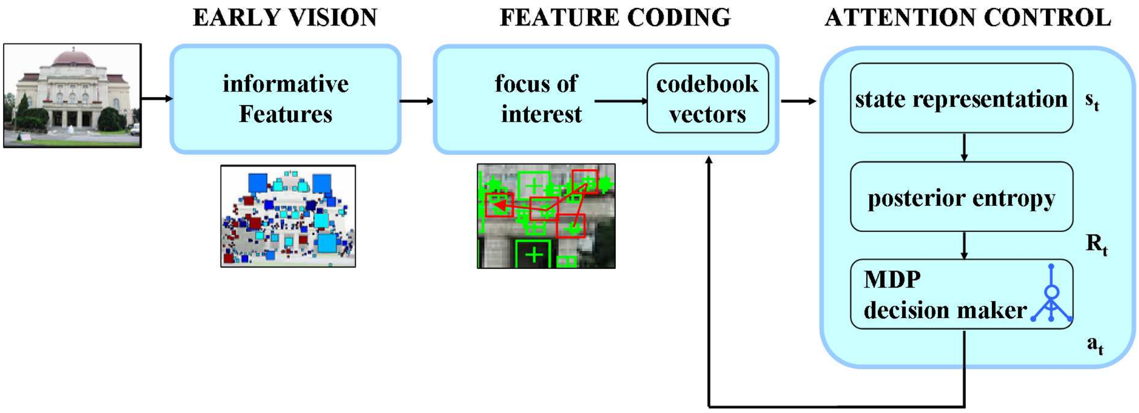


Figure 1: Concept of the proposed perception-action system for object recognition. A module for early vision extracts informative SIFT descriptors [12] from the input image and associates codebook vectors. Sequential attention operates on the geometry between these vectors and statistically reinforces promising feature-action configurations.

local information only at locations that are relevant with respect to 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 used Q-learning to associate *shift of attention*-actions to cumulative reward with respect to a task goal, i.e., object 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 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, by continuously and iteratively integrating local descriptor responses 'on the fly', being capable to evaluate the complete geometric configuration from a set of few features.

The saccadic decision procedure is embedded in a cascaded recognition process (Fig. 1) where visual evidence is probed exclusively at salient image locations. In a first processing stage, salient image locations are determined from an entropy based cost function on object discrimination. Local information in terms of code book vector responses determine the recognition state in the Markov Decision Process (MDP). In the training stage, the reinforcement learner performs trial and error search on useful actions towards salient locations within a neighborhood, receiving reward from entropy decreases. In the test stage, the decision maker demonstrates feature grouping by matching between the encountered and the trained saccadic 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.

2 Informative Foci of Interest for Object Detection

In the propopsed method, attention on informative local image patterns is shifted between the largest local maxima derived from a local feature saliency map (Fig. 4). Informative features are selected using an information theoretic saliency measure on local descriptor patterns as described in detail. The following sections describe the informative feature method from [15] and relate the resulting saliency map to the sequential attention approach.

2.1 Saliency Maps from Local Information Content

We determine the information content from a posterior distribution with respect to given task specific hypotheses. In contrast to costly *global* optimization, we expect that it is sufficiently accurate to estimate a *local* information content, by computing it from the posterior distribution within a sample test point's local neighborhood in feature space [15].

The object recognition task is applied to sample 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 $P(o_k|\mathbf{f}_i)$,

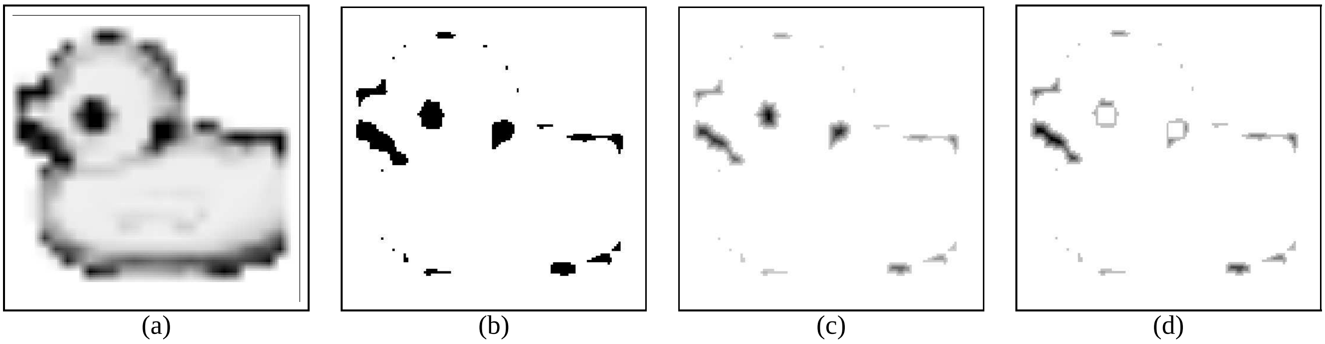


Figure 2: Extraction of FOI (focus of interest) from an information theoretic saliency measure map. (a) Saliency map from the entropy in the local appearances (9×9 pixel window). (b) Binary mask from a thresholded entropy map representing most informative regions ($H_\Theta = 0.2, H < H_\Theta$ white pixels). (c) Distance transform on most informative regions. (d) Inhibition of return for the first 2 FOIs (black regions in informative areas) for maximum saliency extraction from WTA (winner-takes-all) computation [16].

$k = 1 \dots \Omega$, Ω is the number of instantiations of the object class variable O. Shannon conditional entropy denotes

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

We approximate the posteriors at $\mathbf{f}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window of a local neighborhood ϵ,

$$||\mathbf{f}_i - \mathbf{f}_j|| \leq \epsilon, \qquad (2)$$

$j = 1 \dots J$. We weight the contributions of specific samples $\mathbf{f}_{j,k}$ - labeled by object o_k - that should increase the posterior estimate $P(o_k|\mathbf{f}_i)$ by a Gaussian kernel function value $\mathcal{N}(\mu, \sigma)$ in order to favor samples with smaller distance to observation $\mathbf{f}_i$, with $\mu = \mathbf{f}_i$ and $\sigma = \epsilon/2$. 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 $\mathbf{f}_i$.

We receive sparse instead of extensive object representations, in case we store only *selected* descriptor information that is *relevant for classification* purposes, i.e., *discriminative* $\mathbf{f}_i$ with $\hat{H}(O|\mathbf{f}_i) \leq \Theta$. A specific choice on the threshold Θ consequently determines both storage requirements and recognition accuracy. For efficient memory indexing of nearest neighbor candidates we use the adaptive *K-d* tree method.

The local patterns are projected into eigenspace, a Parzen window approach is used to estimate the local posterior distribution $P(o_k|\mathbf{g}_i)$, given eigencoefficient vector $\mathbf{g}_i$ and object hypothesis o_k. The information content in the pattern is computed from the Shannon entropy in the posterior. These features support focus of attention on most salient, i.e., informative image regions for further investigation [17].

2.2 Foci of Interest from Informative Saliency Maps

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 n-best ranked maxima – represented by a focus of interest (FOI) – in the saliency map. At each local maximum, the extracted local pattern is associated to a codebook vector of nearest distance in feature space.

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 pixels 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, one can apply the same WTA rule, selecting the maximum saliency. This masking is known as 'inhibition of return' in the psychology of visual attention [18].

3 Sensory-motor Patterns of Sequential Attention

Sequential attention shifts the focus of attention in the ranked order of maximum saliency, providing an integration

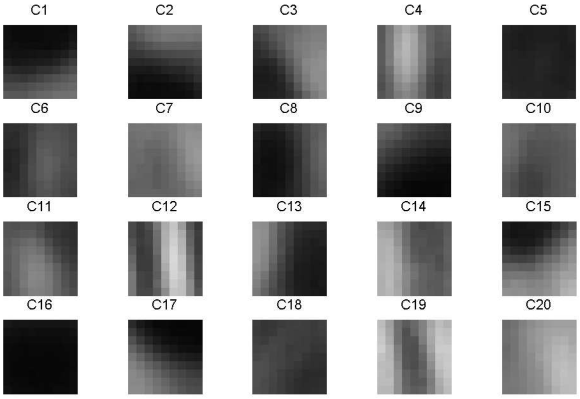
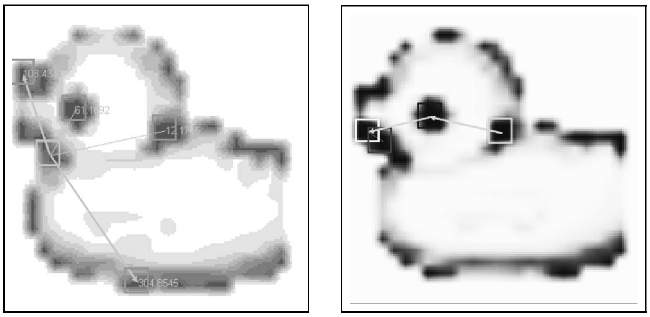


Figure 3: Set of codebook patterns that represent the space of all informative patterns. The patterns have been found by k-means clustering.



(a) (b)

Figure 4: Saccadic attention pattern. (a) Saccadic actions originating in a FOE, directed towards 4 possible target FOIs. (b) Learned attention pattern (scanpath) to recognize the object.

of the visual information in the sampled focused attention windows. In the proposed method, saccadic actions operate on n best-ranked maxima (e.g., n=5 in Fig. 4a) of the information theoretic saliency map. At each local maximum, the extracted local pattern $\mathbf{g}_i$ is associated to a codebook vector Γ_j of nearest distance

$$d = \arg min_j ||\mathbf{g}_i - \Gamma_j|| \qquad (3)$$

in feature space. The codebook vectors were 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, see Fig. 3). The focused local information patterns (in Fig. 4b: the appearance patterns) are therefore associated and thereby represented by prototype vectors, gaining discrimination mainly from the geometric relations between descriptor encodings (i.e, the label of the associated codebook vector) to discriminate saccadic attention patterns. 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. 4a). The individual action and its corresponding angle $\alpha(x, y, a)$ is then categorized into one out of $|A| = 8$ principal directions ($\Delta a = 45°$) (Fig. 5).

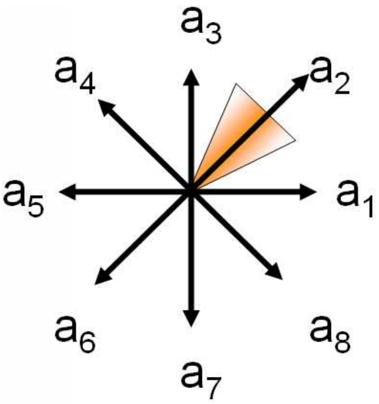


Figure 5: Discretization of the angular encoding for shifts of attention.

An individual state s_i of a saccadic pattern of length N is finally represented by the sequence of descriptor encodings Γ_j and actions $a \in A$, i.e.,

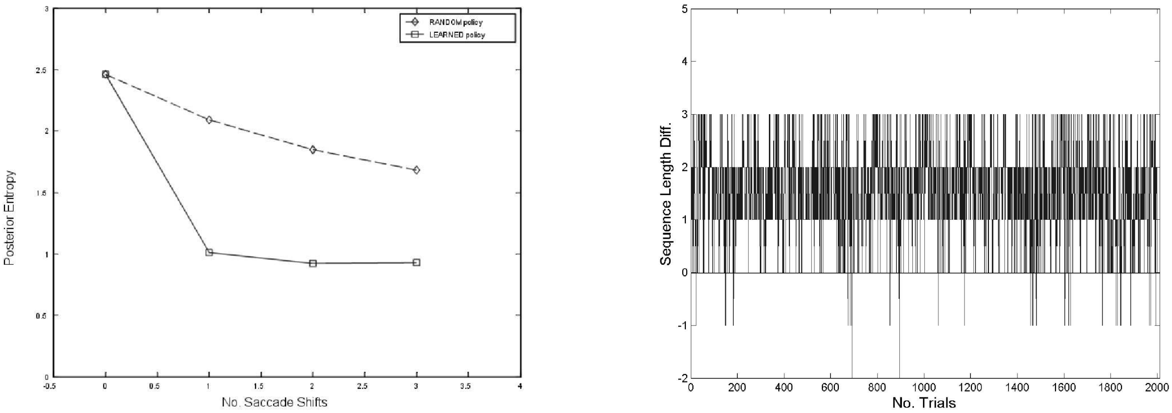
$$s_i = (\Gamma_{n-N}, a_{n-N-1}, \cdots, \Gamma_{n-1}, a_n, \Gamma_n) \, . \qquad (4)$$

Within the object learning stage, random actions will lead to arbitrary descriptor-action sequences. For each sequence pattern, we protocol the number of times it was experienced per 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 $\Delta H_{t+1} = H_t - H_{t+1}$. An efficient strategy aims then at selecting in each state $s_{i,t}$ exactly the action a^* that would maximize the information gain $\Delta H_{t+1}(s_{i,t}, a_{k,t+1})$ received from attaining state $s_{j,t+1}$, i.e.,

$$a^* = \arg max_a \Delta H_{t+1}(s_{i,t}, a_{k,t+1}). \qquad (5)$$

4 Q-Learning of Attentive Saccades

In each state of the sequential attention process, a decision making agent is asked to select an actin to drive its classifier towards a reliable 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.



(a) (b)

Figure 6: Performance evaluation. (a) Rapid information gain from learned attention shift policy in contrast to random action selections. (b) The learned strategy requires shorter shift sequences to pass a given threshold on conditional entropy (threshold $H_{goal} = 1.2$).

The Markov decision process (MDP [19]) provides the general framework to outline sequential attention for object recognition in a multistep decision task with respect to the discrimination dynamics. A 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 shai8ft 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

$$Q(s,a) \equiv U(s,a) = E\left[\sum_{n=0}^{\infty} \gamma^n \mathcal{R}_{t+n}(s_{t+n}, a_{t+n}))\right], \tag{6}$$

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

We formalize a sequence of action selections $a_1, a_2, \cdots, a_n$ in sequential attention as a MDP and are searching for optimal solutions with respect to the object recognition task. In the posterior distribution on object hypotheses, the information gain received from attention shift a

$$\mathcal{R}(s,a) := \Delta H. \tag{7}$$

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 [20] 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

$$Q(s,a) = Q(s,a) + \alpha\left[R + \gamma(max_{a'}Q(s',a') - Q(s,a))\right], \tag{8}$$

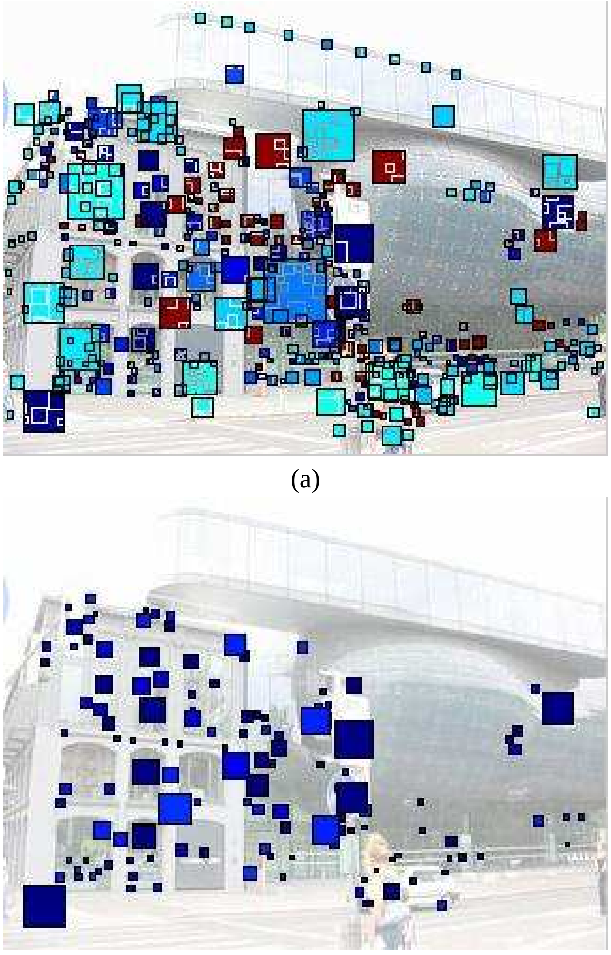
where α is the learning rate, γ controls the impact of a current shift of attention action on future policy return values.

The decision process in sequential attention is determined by the sequence of choices on shift actions at specific focus of interest (FOI). In response to the current visual observation represented by the local descriptor and the corresponding history, i.e., represented by the recognition state, the current posterior is fused to a an integrated posterior. The agent selects then the action $a \in \mathcal{A}$ with largest $Q(s,a)$, i.e.,

$$a_T = \arg max_{a'} Q(s_T, a'). \tag{9}$$

5 Experimental Results

The proposed methodology for cascaded sequential attention was applied to (i) an experiment with indoor imagery (i.e., the COIL-20 database), and to (ii) an experiment with outdoor imagery (i.e.m, the TSG-20 database) on the task of object recognition. The experimental results demonstrate that the informative descriptor method is robustly leading to similar saliency results under various environment conditions, and that the recursive integration of visual information from the informative foci of interest can find good matches to the stored perception-action object representation.



(a)

(b)

Figure 7: Informative descriptors for saliency.

5.1 Indoor Experiments Using Informative Local Appearances

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. 6a represents the learning process, illustrating more rapid entropy decreases from the learned in contrast to random action selection policy. Fig. 6b visualizes the corresponding progress in 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. 4b.

Method	Accuracy $[\%]$	Processing time $[ms]$
i-SIFT	97.5	2800
Sequ. Attention	98.8	1500

Table 1: Sequential attention provides improved performancer with respect to a tuned SIFT recognition method.

5.2 Outdoor Experiments Using Informative SIFT descriptors

In the outdoor experiments, we decided to use a local descriptor, i.e., the SIFT descriptor ([14] Fig. 7) that can be robustly matched to the recordings in the database, despite viewpoint, illumination and scale changes in the object image captures.

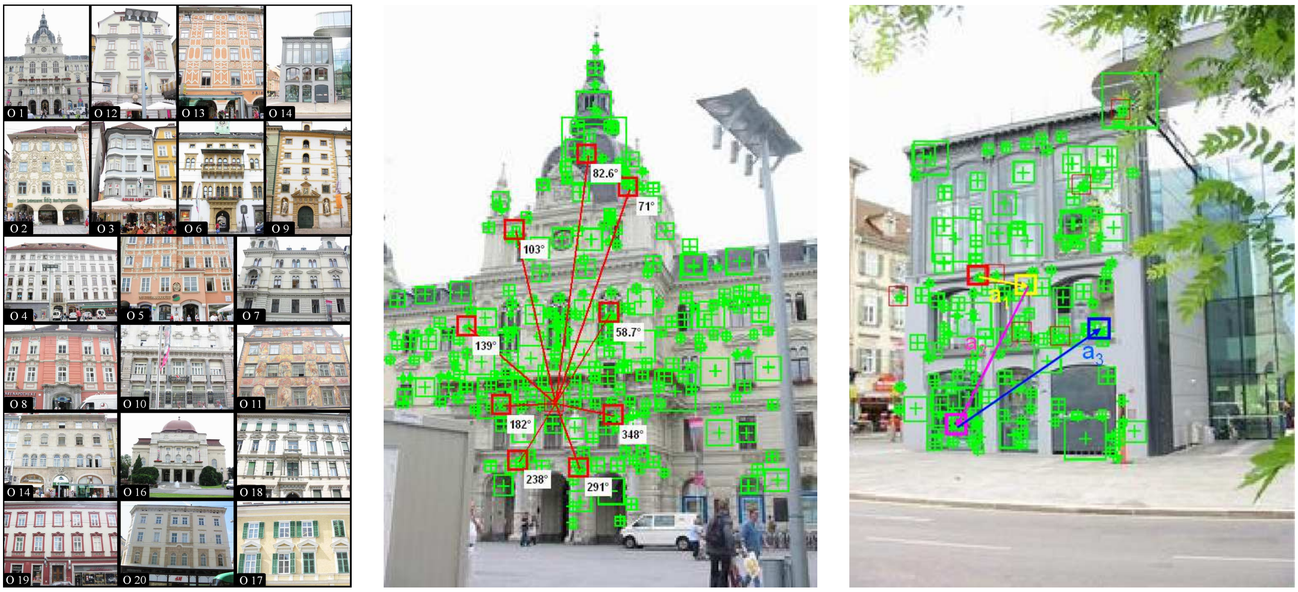
Fig. 7 depicts the principal stages in selecting the FOIs. (a) depicts the original training image. and SIFT descriptor locations are overlaid with squares filled with color-codes of associated entropy values, from corresponding low (red) to high (blue) information values. c) describes a corresponding posterior distribution over all object hypotheses from the MAP hypotheses in the informative SIFT descriptors (i-SIFTs). (d) depicts all selected i-SIFTs in the test image. Fig. 8 illustrates (b) descriptor selection by action and (c) a sample learned sequential attention sequence using the SIFt descriptor.

The experimental results were obtained from the images of the TSG-20 database[1] (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. 9a visualizes the progress gained from the learning process in requiring less actions to attain more informative recognition states. Fig. 9b 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 $\approx 96\%$ (random). A characteristic learned attention scanpath is depicted in Fig. 4b.

6 Conclusions and Future Work

The proposed methodology significantly extends previous work on sequential attention and decision making by providing a proof of concept of a scalable framework for real world object recognition. The three-stage process of determining information theoretic saliency and integrating local

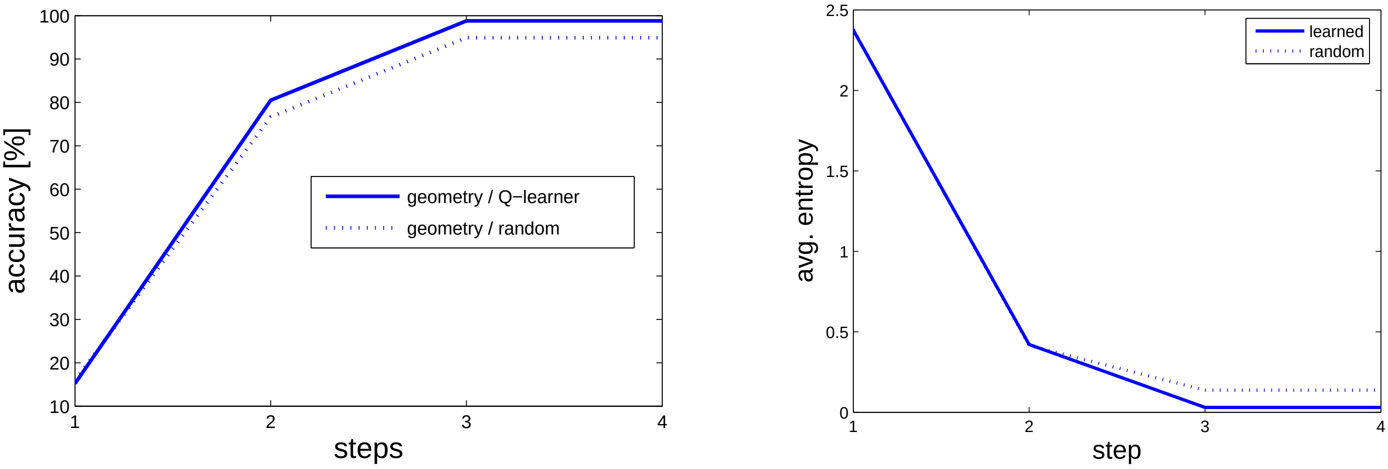
[1] The TSG-20 (Tourist Sights Graz, Fig. 8a) database can be downloaded at the URL http://dib.joanneum.at/cape/TSG-20.



(a) (b) (c)

Figure 8: (a) The TSG-20 database, consisting of images from 20 buildings in the city of Graz, displayed images were used for training (Sec. 5). (b) Saccadic actions originating in a FOI, directed towards 9 potential target FOIs, depicting angle values of corresponding shifts of attention starting in the center SIFT descriptor. (c) Learned descriptor-action based attention pattern (scanpath) to recognize an object.

.



(a) (b)

Figure 9: Performance evaluation. (a) Accuracy improvement from learned attention shift policy in contrast to random action selections. (b) Information gain achieved by learned strategy with each additional perception-action cycle.

descriptive information in a perception-action recognition dynamics is robust with respect to viewpoint, scale, and illumination changes, and provides rapid attentive matching by requiring only very few local samples to be integrated for object discrimination.

Future work will be directed towards hierarchical reinforcement learning in order to provide local grouping schemes that will be integrated by means of a global information integration process from sequential shifts of attention.

References

[1] J.D. Schall and K.G. Thompson, "Neural selection and control of visually guided eye movements," *Annual Review of Neuroscience 22:*, , no. 22, pp. 241–259, 1999.

[2] 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.

[3] A. Gorea and D. Sagi, "Selective attention as the substrate of optimal decision behaviour in environments with multiple stimuli," in *Proc. European Conference on Visual Perception*, 2003.

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

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

[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., pp. 3–69. Elsevier Science, Amsterdam, Netherlands, 1996.

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

[8] R.A. Rensink, J.K. O'Regan, and J.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] M. Li and J.J. Clark, "Learning of position and attention-shift invariant recognition across attention shifts," in *Proc. International Workshop on Attention and Performance in Computational Vision*, 2004, pp. 41–48.

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

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

[12] G. Fritz, C. Seifert, and L. Paletta, "Learning informative SIFT descriptors for object detection," in *Proc. International Conference on Image Processing*, submitted to ICIP 2005.

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

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

[15] G. Fritz, L. Paletta, and H. Bischof, "Object recognition using local information content," in *Proc. International Conference on Pattern Recognition, ICPR 2004*. Cambridge, UK, 2004, vol. II, pp. 15–18.

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

[17] 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.

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

[19] M.L. Puterman, *Markov Decision Processes*, John Wiley & Sons, New York, NY, 1994.

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